How Many Political Parties Should Brazil Have? 
A Data-driven Method to Assess and Reduce Fragmentation in Multi-Party Political Systems

Pedro O.S. Vaz de Melo, Universidade Federal de Minas Gerais

In June 2013, Brazil faced the largest and most significant mass protests in a generation, comparable in size to the protests that triggered the collapse of the military dictatorship in 1984 [1]. The 2013 protests had been exacerbated by the broader disenchantment of the population towards the party system in Brazil [2]. Banners with sentences such as “no party represents me” or “we don’t have a party, we are Brazil!” were commonly seen among the protesters. In response to these protests, the government proposed a program of political reform [1]. However, more than one year has passed and very little has been done.

Why are people so unhappy with the Brazilian party system? To illustrate its incapability, consider the following examples. 19 months before Rio de Janeiro stages South America’s first Olympic games, an Evangelical pastor without any link to sports was nominated as Brazil’s new sports minister. He replaced communist Aldo Rebelo, who oversaw preparations for the World Cup and was highly criticized for subsidizing the construction of white-elephant football stadiums [3]. Aldo Rebelo, who in 1994 proposed a bill that prohibits the adoption of any technological innovation in local, state and federal agencies, is, ironically, Brazil’s new minister of science and technology. Moreover, in 2010 elections, Tiririca, a well known entertainer whose career began as a circus clown in Brazil, was first elected to represent São Paulo in Congress, winning the most votes of any candidate in the country with the slogans “It can’t get any worse” and “What does a Congressman do? In fact, I do not know, but vote for me and I will tell you”.

One of the main causes of Brazil’s political inefficiency is its highly fragmented party system [4]. This is a system with many political parties and with no one party being able to obtain an absolute majority in the representative assembly. The more fragmented the party system is, the less likely it is that the president’s party will control a majority of seats in the legislature. Therefore, presidents are usually forming informal coalition governments, needing to build cross-party coalitions to implement most major policies [5]. Under these circumstances, many (if not most) deputies spend the bulk of their time arranging jobs and pork-barrel projects for their constituents in exchange for legislative support [6]. Also, parties rarely organize around national-level questions, which means that Congress rarely deals with serious social and economic issues [6]. As a consequence, individualism, clientelism, and personalism, rather than programmatic appeals, dominate electoral campaigns [6]. More generally, party system fragmentation impacts the electoral dynamic, the process of coalition formation, governing, and ultimately, the survival of political systems in presidential democracies [7]. In Brazil, party system fragmentation has reached one of the highest levels ever found in the world [8][9]. After 2014 elections, the number of parties represented in Congress grew from 23 to 28.
Besides Brazil, many countries have party systems with high levels of fragmentation, such as Bolivia, Bulgaria, Denmark, Ecuador, Finland, France, Guatemala, India, Israel, Italy, Netherlands and Thailand [9]. In theory, the number of parties of a country can be explained by its electoral and social structures [10; 11; 12]. With regard to the electoral structure, while plurality elections favor two-party competition, proportional representation (PR) electoral systems create fragmented party systems [13; 4]. Concerning the social structures, the more socially heterogeneous a country is, the more electoral parties it will have [14; 15]. Social heterogeneity is measured either by the number of linearly independent ideological dimensions (e.g. religious and socio-economic) being discussed in the society or the number of social cleavages (e.g. centre-periphery and state-church) a country has [10; 15]. Thus, is it possible to measure if a fragmented party system is a reflex of a socially heterogeneous society? In this direction, how can we determine whether an electoral system is optimally fragmented? And if it is not optimally fragmented, how can we optimize it?

To answer these questions, I propose **ARRANGE**, a dAta dRiven method for Assessing and reducing party fragmentation in a country. Inspired by the broad spectrum and advances of data analysis methods [16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30], **ARRANGE** uses as input the roll call data, i.e., the votes given by congressmen on bills and amendments, as a proxy for political preferences and ideology. The idea, regularly employed by political scientists [24; 23], is that congressmen who give the same vote regularly share the same political ideology and, therefore, should belong to the same party. Using this insight, **ARRANGE** reorganizes the party system of a given country by trying to find the minimum amount of social cleavages that divides its congressmen into coherent voting blocks, i.e., sets of congressmen whose members votes similarly. These coherent voting blocks would be the new political parties of the analyzed country and would allow us to assess its actual level of fragmentation. If **ARRANGE** divides the congressmen into a much lower number of political parties than the actual number, then it is possible to conclude that party fragmentation in this particular country is much higher than it should be. If this is the case, **ARRANGE** immediately provides a new congressmen–party configuration that both reduces the country’s level of party fragmentation and increases intra-party similarity, which could potentially increase the efficiency of the party system [8; 4; 5; 6].

To the best of my knowledge, this is the first work that proposes a data-driven method to assess and potentially reduce the number of parties in political systems. In summary, the main contributions of this paper are:

1. **A new problem is addressed**: what is the minimum number of political parties a country should have given the roll votes of its congressmen? Again, although party fragmentation has been extensively studied in the literature, to the best of my knowledge, this is the first time roll call data has been used to assess it.

2. **ARRANGE**, a fast and parsimonious method that receives as input roll call data and outputs a new party system configuration that potentially reduces its actual level of fragmentation.

3. **It is shown that Brazil has and had many ideologically redundant parties**, i.e., parties that are similar in the ideological space. Thus, if today Brazil has one of the highest levels of party system fragmentation in the world (more than 20 parties), this work proves it can be much lower (down to 5 parties).

2. FUNDAMENTALS AND RELATED WORK

The constant advancement of information systems allows, at a growing rate, more data to be stored and generated from the most diverse situations. It is fascinating that, behind all these data, we see the reflection of the environment itself. In order to find knowledge in this invaluable evolving database, a growing number of data-driven methods are being proposed along various research areas. For instance, there are data-driven methods to predict hospital mortality from instance-based patient data [16] and flu epidemics [31]. In the social sciences, Silva et. al. proposed two data-driven methods to quantitatively characterize cultural behaviors of geographical regions [17; 18] and Park et. al. [19] designed and evaluated a measure that captures diversity of musical tastes from social
media data. In economics, [22] proposed a data-driven approach to understand online consumer behavior and engagement with brands. For the benefit of the industry sector, there are data-driven methods to monitor industrial processes [21] and to assist the development and deployment of intelligent transportation systems [20]. The constant advancement of information systems allows, at a growing rate, more data to be stored and generated from the most diverse situations. It is fascinating that, behind all these data, we see the reflection of the environment itself. In order to find knowledge in this invaluable evolving database, a growing number of data-driven methods are being proposed along various research areas. For instance, there are data-driven methods to predict hospital mortality from instance-based patient data [16] and flu epidemics [31]. In the social sciences, Silva et. al. proposed two data-driven methods to quantitatively characterize cultural behaviors of geographical regions [17, 18] and Park et. al. [19] designed and evaluated a measure that captures diversity of musical tastes from social media data. In economics, [22] proposed a data-driven approach to understand online consumer behavior and engagement with brands. For the benefit of the industry sector, there are data-driven methods to monitor industrial processes [21] and to assist the development and deployment of intelligent transportation systems [20].

In the political sciences, data analysis methods from roll votes primarily focuses on the estimation of cleavages and ideologies across congressmen and parties [32, 23, 24] to characterize and predict legislative behavior [33, 34, 25]. Recently, with the advancement of political weblogs and online social networks, researchers are also extracting political knowledge from user generated data on the Web. There are studies that focus on mining political opinions [26] and profiles [27] from the texts users post on social media applications. Others, such as [35], extract political opinions from general texts, such as statement records of U.S. senators and online news. More recently, Leman Akoglu [28] classified the political polarity of individuals using roll call votes of U.S. congressmen and texts posted on political forums. The idea behind all these studies is that political preferences tend to be stable over time and can be predicted accurately.

Still in the political sciences, the study of party systems is one of its largest sub-fields [11]. Within this sub-field, since Duverger’s seminal paper [13], many studies focused on predicting and understanding the factors that determine the number of parties that compete in a given polity [36, 12, 37, 7, 9]. In summary, there are two lines of thought: one that emphasizes the role of electoral laws in structuring coalition incentives, and another that emphasizes the importance of preexisting social cleavages. Another fundamental problem in this sub-field is to count the number of parties by taking into account their relative size [38]. If, for instance, a party has a very small percentage of seats in Congress (e.g. one seat of one thousand), then it should be counted accordingly. The metric that considers this is called the effective number of parties. Conceptually, the effective number of parties is simply the number of “viable” or “important” political parties in a party system that includes parties of unequal sizes. Since Laakso and Taagepera’s seminal work [38], several ways of computing the effective number of parties were proposed [39, 40, 41]. The number of effective parties is a frequent metric for assessing party system fragmentation in a country [41].

The high interest in these problems comes from the fact that the actual number of parties usually determines the number of effective parties, or how fragmented a party system is [38]. Highly fragmented party systems can affect governance drastically [42]. The more fragmented the party system is, the less likely it is that the president’s party will control a majority of seats in the legislature. Simone Bohn [7] reviewed the literature and concluded that party system fragmentation impacts the electoral dynamic, the process of coalition formation, governing, and ultimately, the survival of political systems in presidential democracies. Thus, in this paper, we measure party fragmentation by counting both the actual and the effective number of parties.

Another crucial factor for governance is party discipline, i.e., the ability of a political party to get its members to support the policies of their party leadership. Mainwaring and Shugart [5] assessed the effects of this on the costs of governing. If parties are not disciplined, presidents will be forced to rely on ad-hoc coalitions based on the distribution of patronage to individual legislators, which raises the costs of governing and reduces policy coherence. Limongi and Figueiredo [8] argued that “institutional engineering” should focus on electoral formulas that reduce party fragmentation and
increase party discipline. Brazil is a special case in politics for its high level of party fragmentation, being consistently analyzed in the literature [15, 37, 8, 21, 7]. Thus, using Brazil as the use case makes this work specially challenging but rewarding.

Finally, it is important to emphasize that this work differs significantly from those that focus on algorithms to find communities in networks [30, 44]. Although such algorithms could be applied here to detect communities of congressmen that are ideologically similar and, therefore, could compose a political party, this is very different from our problem in two major aspects. First, here our goal is to find the minimum number of communities (in our case, political parties), which is an optimization problem not addressed by community detection algorithms, which usually aim to maximize modularity [45] or any other cohesion metric [46]. Second, while traditional community detection algorithms do not allow two disjoint subgraphs to be part of the same community, our major constraint here is party discipline. Thus, here we allow two disjoint subgraphs (in our case, two ideologically dissimilar groups of congressmen) to be part of the same community if the party discipline constraint is satisfied. The comparison between the communities (political parties) generated by the proposed method and by the state of the art community detection algorithms is left for future work.

3. DATA DESCRIPTION

All the data used in this work was collected from the Open Data (Dados Abertos) project of the House of Representatives (or Chamber of Deputies) of Brazil. In total, I collected 744,195 thousand roll votes on 774 bills of 1,582 thousand congressman that worked in the House of Representatives of Brazil from November, 4th, 1998 to December, 3rd, 2014. The reason for this particular time interval is related to the purpose of this work. Since party discipline is a fundamental metric of evaluation, I only collected the bills in which the party leaders declared the desired vote for their fellow partisans. More than 95% of the votes given during this period had a declared party leader vote. Moreover, note that congressmen vote for bills and their amendments. An amendment is a proposition presented as ancillary to the bill, to amend its form or content. Thus, bills and amendments compose a total of 2,162 thousand propositions to be voted by the congressmen. Each congressman may or may not agree with the vote of his/her party leader. There are, in total, 35,216 thousand declared votes of the leaders of the 36 parties that had congressmen elected for the House of Representatives during the analyzed period.

4. FORMAL DEFINITIONS

As discussed previously, reducing party fragmentation only makes sense if party discipline does not decrease significantly. One way to measure the discipline of a congressman is to compute the fraction of votes given by him/her through his political life that agreed with his/her party leader. However, in Brazil it is common for congressmen to switch parties. Also, it is well known that some parties demand (or inspire) higher levels of discipline than others [8]. Thus, instead of analyzing the discipline of the whole political life of a congressman, I will analyze his/her discipline as a member of a single party, i.e., a congressman may have different levels of discipline if he/she was member of different parties during his/her political life. For simplicity, from now on I will assume that the vote given by the leader of the party was given by the party itself, e.g., I will call party vote the vote given by the leader of the party.

Thus, given the set of \( m \) congressmen \( U = \{u_1, u_2, ..., u_m\} \) and the \( N \) political parties that compose the set \( P = \{p_1, p_2, ..., p_N\} \), I define a partisan \( a := (u, p) \) as the tuple formed by a congressman \( u \) and a political party \( p \). The set containing all \( M \) partisans is defined as \( A = \{a_1, a_2, ..., a_M\} \). These partisans’ job is to vote for the set \( B = \{b_1, b_2, ..., b_n\} \) of \( n \) bills and amendments that were put to vote in the House of Representatives during the analyzed period. From now on, I will use the term propositions to refer to both bills and amendments. Since a partisan \( a \) had not necessarily voted for all propositions, I define the set \( B_a \subseteq B \) as the set of propositions that were voted by partisan \( a \). I also define \( A_p \) as the set of partisans which are members of party \( p \), i.e., \( A_p = \{a_i | a_i := (u, p') \land p' = p\} \).
Before the partisans give their votes for a given proposition \( b \), their parties have to announce their votes for \( b \). Again, since a party \( p \) had not necessarily voted for all propositions, I define the set \( B_p \subseteq B \) as the set of propositions that were voted by party \( p \). For each proposition \( b \in B_p \) of a given party \( p \), there is a vote \( v_{pb}^b \) associated with it. In the same way, for each proposition \( b' \in B_a \) of a given partisan \( a \), there is a vote \( v_{ab}^{b'} \) associated with it. Thus, a party \( p \) and a partisan \( a \) have, respectively, a set of votes \( V_p \) and \( V_a \), where \(|V_p| = |B_p|\) and \(|V_a| = |B_a|\). The set of all votes given by partisans and parties is simply \( Y \).

A vote \( v_{pb}^b \) given by party \( p \) on a proposition \( b \) may be of four types: \( Y \) (yes), \( N \) (no), \( O \) (obstruction) and \( F \) (free), i.e., \( v_{pb}^b \in \{Y, N, O, F\} \). If the vote is \( Y \) (\( N \)), the party approves (disapproves) the proposition. If the vote is \( O \), the party is trying to avoid the vote on the proposition, i.e., its partisans are called to withdraw from the plenary. Finally, if the vote is \( F \), its partisans are free to vote at will. Similarly, a partisan \( a \) vote \( v_{ab}^{b'} \) on a proposition \( b' \) may be of three types: \( Y, N, O \), i.e., \( v_{ab}^{b'} \in \{Y, N, O\} \). If the vote is \( Y \) (\( N \)), the partisan approves (disapproves) the proposition. If the vote is \( O \), the partisan withdrew from plenary.

In this work, the vote is the fundamental feature that determines a preference or ideology. Since for all propositions in our dataset we have both the vote of the partisan and the party, here I define a general function \( \text{agrees}(v_1, v_2) \) that receives two votes as input and outputs 1 if the votes are in accordance or 0 otherwise. This function is defined as:

\[
\text{agrees}(v_1, v_2) = \begin{cases} 
1 & \text{if } v_1 = v_2 \text{ or } v_1 = F \text{ or } v_2 = F \\
0 & \text{otherwise.} 
\end{cases} 
\]  

(1)

Note that both \( v_1 \) and \( v_2 \) can be a vote of a partisan or a party. Also note that a \( F \) vote implies accordance, since the party that gave that vote does not particularly care about its members votes. Once we know how to compare votes, we can propose a way to compare the similarity \( \text{sim}(i, j) \) between two vote sets \( V_i \) and \( V_j \). It is given as:

\[
\text{sim}(i, j) = \sum_{b \in B_i \cap B_j} \text{agrees}(v_i^b, v_j^b) \times |B_i \cap B_j|^{-1}. 
\]  

(2)

In summary, \( \text{sim}(i, j) \) sums all the votes in agreement between the vote sets \( V_i \) and \( V_j \) considering only the propositions that are in both sets. From Equation(2) we can define the three levels of discipline that we will use throughout this paper. First, I define partisan discipline as the discipline \( d_{a \rightarrow p} \) of a partisan \( a := (u, p) \) towards his/her party \( p \), calculated as:

\[
d_{a \rightarrow p} = \text{sim}(a, p) | a := (u, p). 
\]  

(3)

Second, I define party discipline as the discipline \( d_p \) throughout all the votes that were given by partisan members of the party \( p \), calculated as:

\[
d_p = \frac{\sum_{a \in A_p} |B_a| \times d_{a \rightarrow p}}{\sum_{a \in A_p} |B_a|} 
\]  

(4)

Finally, I define overall discipline as the discipline \( d_a \) throughout all the votes given in the House of Representatives during the analyzed period, calculated as:

\[
d_a = \frac{\sum_{a := (u, p)} |B_a| \times d_{a \rightarrow p}}{\sum_{a := (u, p)} |B_a|} 
\]  

(5)

5. POLITICS IN BRAZIL

In this section I will show a summarized view of the fundamental characteristics of Brazilian party system that are relevant to the purpose of this work. Because all parties in Brazil are often referred by their acronyms, I will also use their acronyms instead of their names. Thus, please refer to Table[1]
Table I. Current and historical parties of Brazil.

| Name                                      | Acronym | \( B_p \) |
|-------------------------------------------|---------|------------|
| Partido do Movimento Democrático Brasileiro | PMDB    | 2,163      |
| Partido da Social Democracia Brasileira   | PSDB    | 2,163      |
| Partido dos Trabalhadores                 | PT      | 2,163      |
| Partido Democrático Trabalhista           | PDT     | 2,152      |
| Partido Socialista Brasileiro             | PSB     | 2,144      |
| Partido Trabalhista Brasileiro            | PTB     | 2,144      |
| Partido Popular Socialista                | PPS     | 2,137      |
| Partido Comunista do Brasil               | PCDOB   | 2,127      |
| Partido Verde                             | PV      | 1,798      |
| Partido Progressista                      | PP      | 1,500      |
| Partido Social Cristo                      | PSC     | 1,333      |
| Partido Socialismo e Liberdade            | PSOL    | 1,230      |
| Partido da República                      | PR      | 1,134      |
| Partido da Frente Liberal                 | PFL     | 1,081      |
| Democratas                                | DEM     | 1,077      |
| Partido Republicano Brasileiro            | PRB     | 1,061      |
| Partido Liberal                           | PL      | 1,001      |
| Partido da Mobilização Nacional           | PMN     | 981        |
| Partido Social Liberal                    | PSL     | 865        |
| Partido Humanista da Solidariedade        | PHS     | 662        |
| Partido Trabalhista Cristo                | PTC     | 606        |
| Partido Progressista Brasileiro           | PBB     | 567        |
| Partido Trabalhista do Brasil             | PCDOB   | 471        |
| Partido Social Democrático                 | PSOL    | 466        |
| Partido Social Trabalhista                | PST     | 464        |
| Partido de Reedificacao da Ordem Nacional | PRONA   | 442        |
| Partido Republicano Progressista          | PRP     | 356        |
| Partido Renovador Trabalhista Brasileiro  | PRTB    | 179        |
| Partido Republicano da Ordem Social       | PROS    | 147        |
| Solidariedade                             | SDD     | 146        |
| Partido Ecológico Nacional                | PEN     | 100        |
| Partido Trabalhista Nacional              | PTN     | 92         |
| Partido dos Aposentados da Nação           | PAN     | 75         |
| Partido Social Democrata Cristão          | PSDC    | 38         |
| Partido Republicano Brasileiro            | PMR     | 33         |
| Partido Socialista dos Trabalhadores Unificados | PSTU   | 8          |

for a list of all parties’ names and their respective acronyms and sizes, in this case given by the size of their \( B_p \).

First, in Fig. 1 I show the historical participation of all parties in the House of Representatives during the analyzed period. Party participation, which I will interchangeably call party size, is represented by the total number of propositions that were voted by the members of the party (horizontal axis) and the total number of partisans that are and were members of the party (vertical axis). Observe the heterogeneity of this universe. From the three biggest parties (PMDB, PSDB and PT), with hundreds of partisans who voted for thousands of propositions, to the two smallest ones (PSTU and PMR), which together have only three partisans and forty one propositions, there are another 31 parties with very distinct levels of representation.

In order to verify if there is any correlation between participation and party discipline, I show in Fig. 2 for each party \( p \), the total number of votes given by partisans members of \( p \) (horizontal axis) and the party discipline \( d_p \) (vertical axis). Note that there is no apparent relationship between party discipline and participation. In fact, the Pearson’s correlation coefficient between party discipline and the total number of votes is 0.24, but since the \( p-value \) for testing the hypothesis of no correlation is 0.28, we cannot affirm the correlation is significant. Nevertheless, as already observed by [8] using a smaller dataset, party discipline in Brazil is consistently high: no party has a historical party discipline below 0.75 and only two parties have a figure below 0.85.
In Fig. 3 we show the behavior of partisans discipline in Brazil during the analyzed period. First, observe in Fig. 3a the Cumulative Distribution Function (CDF) of all partisans’ disciplines. Note that the curve representing partisan discipline in Brazil is not very far away from the ideal curve, where all partisans have discipline of 1.0. Thus, together with party discipline, partisan discipline is also usually high in Brazil: only 6.3% of partisans have discipline lower or equal to 0.8. In Fig. 3b, we plot the heatmap of partisans using their disciplines and total number of votes. The color bar at the right indicates the number of partisans in a given area of the map. Observe that the vast majority of partisans are located in the upper-right of the heatmap, i.e., they have given many votes and have high partisan discipline. The Pearson’s correlation coefficient of 0.03 and $p-value = 0.23$ indicate that there is no correlation between discipline and participation.
Although party fragmentation in Brazil has reached one of the highest levels ever found in the world [8; 9], we have seen that party and partisan discipline is consistently high. Does this mean that the level of party fragmentation in Brazil is necessary? According to the seminal doctrine of responsible party government [47], parties must differ sufficiently between themselves, providing the electorate with a proper range of choice between alternative actions. Given that, I reformulate the previous question: is the actual level of party fragmentation in Brazil a consequence of a high number of sufficiently different parties?

Instead of performing a deep clustering analysis to answer this question, I will apply the Principal Component Analysis (PCA) [48] technique to the matrix $M_{VBR}$ composed by the voting vectors $V_{a_i}$ of each partisan $a_i$. PCA is a widely used statistical technique for unsupervised dimension reduction. It transforms the data into a new coordinate system such that the greatest variance is achieved by projecting the data into the first coordinate, namely principal component, the second greatest variance is achieved by projecting into the second coordinate, the second component, and so on. In order to draw more interesting conclusions from this analysis, I will also apply PCA to the matrix $M_{VUS}$ composed by the voting vectors of the U.S. congressmen. For this, I will use the same dataset used in [28], which consists of votes on 1655 bills in The House of Representatives in years 2009-2010 by 451 representatives. In the matrices the YES votes are represented as 1, the NO votes as $-1$, and the $F$, $O$ and non-attendance as 0. To make the comparison more precise, I will only use the votes in years 1999-2000 for constructing $M_{VBR}$. This period comprises votes on the 349 bills by 767 contemporary partisans of 18 parties in the first two years of president Cardoso in power. Also, it is the two year period that gives the higher explained variance by the first two components of the PCA.

In Fig. 4 I show the first two components of the PCA for both $M_{VBR}$ and $M_{VUS}$, where each point represents a congressman and each symbol represents a party. Observe that, for both USA and Brazil the first two components explain a significant part of the variance: 67% and 53%, respectively. However, only the USA PCA can visually divide the members of different parties. For Brazil, many members of different (same) parties are located together (apart). This suggests that parties in Brazil are not sufficiently different to justify one of the highest levels of party fragmentation ever found in the world [8; 9].

6. THE ARRANGE METHOD

In this section I describe the method ARRANGE, which has, basically, two steps. First, based on the votes given by party leaders, it tries to find pairs of parties that can be merged into one. The idea is that parties that always give the same vote could be merged into a single party. Then, ARRANGE attempts to assign partisans to new parties with the objective of minimizing the total number of
parties receiving partisans and preserving party discipline. Finally, I describe the quality outputs of \textit{ARRANGE}. Formally, the problem tackled by \textit{ARRANGE} is:

**Problem 1.** Given a set of parties $\mathcal{P}$, a set of partisans $\mathcal{A}$, a set of bills and amendments $\mathcal{B}$ and the set of all votes given by parties and partisans $\mathcal{V}$, find the minimum set of parties $\mathcal{P}^*$ to which the partisans in $\mathcal{A}$ can be assigned in a way that overall, party and partisan disciplines are maximized.

6.1. Merging Political Parties

In order to assign partisans to other parties, it is necessary to formally define the ways it can be done. Thus, here I define an \textbf{option} as the descriptor of the party that can receive an external partisan as member. More formally, given a partisan $a := (u, p_i)$, his/her current party $p_i$, and his/her set of propositions $B_a$, an option $o_a := (a, p_j, \text{sim}(a, p_j))$ is a tuple composed by the partisan $a$, a party $p_j \neq p_i$, and the similarity value $\text{sim}(a, p_j)$ between $a$ and $p_j$. More importantly, the option $o_a = (a, p_j, \text{sim}(a, p_j))$ exists if and only if $B_a \subseteq B_{p_j}$, i.e., if the party $p_j$ has voted for all propositions in $B_a$. The set $O_a = \{o_1^a, o_2^a, ...\}$ is composed by all the options of partisan $a$ or, more formally

$$O_a = \{o_a := (a := (u, p_i), p_j, \text{sim}(a, p_j)) : p_i, p_j \in \mathcal{P} \land p_j \neq p_i \land B_a \subseteq B_{p_j}\}.$$  

Moreover, an option $o_a := (a, p_j, \text{sim}(a, p_j))$ of partisan $a := (u, p_i)$ is characterized as a \textbf{good option} if $\text{sim}(a, p_j) \geq \text{sim}(a, p_i)$. The set of good options $O_a^*$ for partisan $a := (u, p_i)$ is defined as

$$O_a^* = \{o_a = (a := (u, p_i), p_j, \text{sim}(a, p_j)) \in O_a : \text{sim}(a, p_j) \geq \text{sim}(a, p_i)\}.$$  

In Fig. 5a, I show the histogram of the number of good options $|O_a^*|$ for all partisans of our dataset. Consistent with the high levels of party and partisan discipline, observe that the majority of partisans do not have a single good option, i.e., for 69% of the partisans there is no other party in Brazil that offers more similar voting vectors. In fact, only $\approx 11\%$ of the partisans have more than three good options. This result suggests that it is very difficult to reduce the number of parties in Brazil by moving partisans from one party to another.

The main reason for the low number of good options is related to the short lifetime of many parties in Brazil, as it can be observed in Fig. 1. Since an option exists if and only if the propositions voted on by the partisan is a subset of the propositions voted on by the parties, many partisans with a long history of votes cannot find options for them among the small parties. Thus, here I propose
a method for creating new parties by merging existing ones. The method is based on a simple idea: if two parties are not contemporary or are contemporary, but all votes given by them are equal, then these parties can be merged into a new one.

More formally, two parties \( p_i \) and \( p_j \) can be merged into a new party \( p_{i,j} \) if one of the two conditions below is met:

— **C1.** \( B_{p_i} \cap B_{p_j} = \emptyset \), i.e., parties \( p_i \) and \( p_j \) have not voted for a common proposition.

— **C2.** \( \forall b \in B_{p_i} \cap B_{p_j} : v_{b_{p_i}} = v_{b_{p_j}} \), i.e., parties \( p_i \) and \( p_j \) have given the same vote for all common propositions.

Given these two conditions, the first thing we have to do is find all pairs of parties to which at least one of the two conditions is valid. Once this is done, for every pair of parties \((p, q)\) that can be merged, we create a merged party \( p_{i,j} \), for which the sets of propositions and votes are, respectively,

\[
B_{p_{i,j}} = B_{p_i} \cup B_{p_j} \quad \text{and} \quad V_{p_{i,j}} = V_{p_i} \cup V_{p_j}.
\]

All parties created in this step are put in set \( P' \). After this, we repeat this process by verifying, for each party \( p \in P' \), all parties \( q \in P \) that can be merged to \( p \). We merge \( p \) to \( q \) into \( p_{i,j} \) as previously, but this time adding the merged parties to \( P'' \). Once this process is done, we copy \( P'' \) to \( P' \), empty \( P'' \), and restart the process of finding parties \( q \in P \) eligible to be merged to parties \( p \in P' \). The process ends when set \( P'' \) does not receive a new merged party. All parties \( p \in P' \) that were not merged are put in the final set of merged parties \( P_M \). This whole process is described in Algorithm 1.

Basically, what Algorithm 1 does is to find all possible \( k \)-combinations \( \binom{N}{k} \) of set \( P \) for all \( 1 \leq k \leq N \). It is well known that \( \sum_{k=1}^{N} \binom{N}{k} = 2^N - 1 \), making the worst-case complexity for this problem to be \( O(2^N) \). Nevertheless, in practice, finding parties in \( P \) that can be merged with parties in \( P' \) gets significantly harder as \( k \) increases, which, in practice, makes Algorithm 1 computationally feasible for the problem in question. For the case of the 36 Brazilian parties, \( k \) went up to 6 and the algorithm stopped, generating a set \( P^M \) containing 95 parties, in which only 5 were not a product of a merge, namely PDT, PMDB, PPS, PSDB and PT. Besides these, Algorithm 1 generated 12 parties of size 2, 17 parties of size 3, 20 parties of size 4, 31 parties of size 5 and 10 parties of size 6.

In Fig. 5b, I show the histogram of the number of good options \( |O^*_u| \) for all partisans considering the new set of merged parties \( P^M \). Observe that the number of partisans that do not have a single good option dropped from \( \approx 69\% \) to \( \approx 33\% \), all of them being members of the five parties that were not merged. Moreover, the number of partisans that have more than three good options grew from \( \approx 12\% \) to \( \approx 47\% \). Thus, I conclude that merging parties significantly raises the chances of reducing the number of parties in Brazil by moving partisans from one party to another without decreasing party discipline.
6.2. Finding the Minimum Set

Now that most of the partisans have multiple good options, we can find ways of redistributing them among the parties in $P^M$. The idea in this redistribution is to find a set of parties $P^* \subset P^M$ that has a lower cardinality than $P$, i.e., $P^* \subset P^M$ is a set of parties able to receive all partisans $a \in A$ as members and the number of parties $N^* = |P^*|$ in $P^*$ has to be lower than the actual number of parties $N = |P|$. However, this has to be done cautiously, since there are two partially conflicting goals:

(1) Minimize the number of parties;
(2) Maximize party and partisan discipline.

These goals are conflicting because minimizing the number of eligible parties to receive partisans as members implies reducing the options for moving partisans and, as a consequence, the number of good options. If a partisan does not have a good option, he/she is obliged to stay in his/her party in order to not decrease his/her partisan discipline. On the other hand, maximizing the discipline implies in maximizing the size of the set of good options and, therefore, the number of parties to be considered has to be as high as possible.

In order to solve this conflict, I model this redistribution problem as a set cover problem (SCP) [49]. SCP is a well studied problem for the field of approximation algorithms [50], being also one of Karp’s 21 NP-complete problems shown to be NP-complete. In summary, given a set of elements $\{1, 2, ..., m\}$ (called the universe) and a set $S$ of $n$ sets whose union equals the universe, the SCP is to identify the smallest subset of $S$ whose union equals the universe. More formally, given a universe $\mathcal{X}$ and a family $S$ of subsets of $\mathcal{X}$, a cover is a subfamily $C \subseteq S$ of sets whose union is $\mathcal{X}$.

In our case, the universe $\mathcal{X}$ is the set of partisans $A$ and the family $S$ of subsets of $\mathcal{X}$ is a family of subsets of partisans $A_p^M \in A$ where each subset $A_p^M$ is composed by the partisans that are eligible for moving to party $p \in P^M$. In order to build $A_p^M$, it is necessary to recalculate the set of options $O_a$ and good options $O_a^*$ of each partisan $a$ with respect to the merged parties in $P^M$. Once this is done, I define that each set $A_p^M$ is composed by all the partisans $a$ that have a good option $o_a := (a, p, \text{sim}(a, p))$, i.e., $A_p^M = \{a | a \in A \land p \in P^M \land \exists o_a := (a, p, \text{sim}(a, p)) \in O_a^*\}$.

Since this problem is NP-complete, I recur to a greedy algorithm to solve it. Literature shows that the greedy algorithm is essentially the best possible polynomial time approximation algorithm for set cover under plausible complexity assumptions [51]. The greedy algorithm for set covering chooses sets according to one rule: at each stage, choose the set that contains the largest number of partisans.
of uncovered elements. It can be shown [49] that this algorithm achieves an approximation ratio of $H(s)$, where $s$ is the size of the set to be covered and $H(n)$ is the $n$-th harmonic number: $H(n) = \sum_{k=1}^{n} \frac{1}{k} \leq \ln n + 1$.

For our specific problem, this algorithm works as follows, being described in Algorithm[2] First, we create two empty sets: $\mathcal{P}^*$, which will contain the final set of parties, and $\mathcal{A}^i$, which receives the covered partisans during the process. Then, while $\mathcal{A}^i$ does not contain all partisans, we find the party $p \in \mathcal{P}^M$ to which $\mathcal{A}^M_p$ contains the largest number of uncovered partisans. Then, we add all partisans $a \in \mathcal{A}^M_p$ to $\mathcal{A}^i$, make $\mathcal{A}_p^M$ an empty set (so it is not selected again), and add $p$ to the final set of parties $\mathcal{P}^*$.

Note that so far this process guarantees that all partisans will, at least, have the same partisan discipline as their actual ones, since only good options are used to build the sets $\mathcal{A}^M_p, p \in \mathcal{P}^M$. Nevertheless, it is possible that relaxing this constraint a little might diminish the total number of parties that compose $\mathcal{P}^*$ considerably. For instance, we may allow a partisan to be member of a party if their similarity is at most 0.05 smaller than his/her similarity with his/her actual party. This relaxation increases the number of partisans that are eligible to be member of other parties and, therefore, may reduce the size of $\mathcal{P}^*$.

Thus, here I introduce the parameter $\delta$, which is the maximum allowed difference between the actual partisan discipline and the future one. We accommodate this in $\text{ARRANGE}$ by simply changing the way the set of good options is constructed. With the introduction of $\delta$, the set of good options $\mathcal{O}^*_a$ is defined as:

$$\mathcal{O}^*_a = \{[a := (u, p_i), p_j, sim(a, p_j)] \in \mathcal{O}_a : p_i \in \mathcal{P} \land p_j \in \mathcal{P}^M \land sim(a, p_j) \geq sim(a, p_i) - \delta\}. \quad (8)$$

In summary, after generating the set of merged parties $\mathcal{P}^M$ using Algorithm[1] it is necessary to create the sets $\mathcal{A}^M_p$ for all $p \in \mathcal{P}^M$ considering $\delta$. These sets will contain all partisans that are eligible for being members of party $p$ given $\delta$. This is done by selecting a value for $\delta$ and then running the procedure $\text{FIND}$ of Algorithm[2] If, for instance, $\delta = 0$, then no partisans are allowed to decrease their discipline when moving to a different party. On the other side, if $\delta = 1$, partisans are allowed to be members of any party that voted for all propositions that they voted on, i.e., partisan discipline is not a constraint. After running Algorithm[2] we will have a minimum set of parties $\mathcal{P}^*$ able to accommodate all partisans in $\mathcal{A}$. Then, all we have to do is to assign a party $p \in \mathcal{P}^*$ to each partisan $a \in \mathcal{A}$. This is done by selecting the option $o_a = (a := (u, p_i), p_j, sim(a, p_j)) \in \mathcal{O}^*_a$ that gives the maximum similarity value $sim(a, p_j)$ for all parties $p_j \in \mathcal{P}^*$ and making $p_j$ the new party of partisan $a$.

Algorithm 2 Find the minimum set of parties $\mathcal{P}^*$

```
1: procedure buildOptions($\mathcal{A}^M_p, \delta$)
2:   for all $a \in \mathcal{A}$ do
3:       $\mathcal{O}_a \leftarrow \{[a := (a, q, sim(a, q))] : p \in \mathcal{P} \land q \in \mathcal{P}^M \land B_a \subseteq B_q\}$
4:       $\mathcal{O}^*_a \leftarrow \{[a := (u, p), q, sim(a, q))] \in \mathcal{O}_a : sim(a, q) \geq sim(a, p) - \delta\}

5: procedure findP($\mathcal{A}^M_p, \delta$)
6:   buildOptions($\mathcal{A}^M_p, \delta$)
7:   $\mathcal{A}' \leftarrow \{\}$
8:   $\mathcal{P}' \leftarrow \{\}$
9:   while $\mathcal{A}' \neq \mathcal{A}$ do
10:      $\mathcal{P}'' \leftarrow \{\}$
11:      $p \leftarrow \arg\max_p |\mathcal{A}^M_p \cup \mathcal{A}|$
12:      $\mathcal{A}'' \leftarrow \mathcal{A}' \cup \mathcal{A}^M_p$
13:      $\mathcal{A}^M_p \leftarrow \{\}$
14:      $\mathcal{P}' \leftarrow \mathcal{P}' \cup \{p\}$
15:   return $\mathcal{P}'$
```

 annotations:
- $\mathcal{O}_a$: the set of uncovered partisans
- $\mathcal{O}^*_a$: the final set of parties
- $\mathcal{A}^M_p$: the set of good options
- $\mathcal{P}^*$: new merged parties
6.3. Quality Signals

The main goal of \textit{ARRANGE} is to reduce party fragmentation by reducing the number of parties that are able to accommodate all elected partisans. Thus, the main output of \textit{ARRANGE} is \(N^*\) (or \(P^*\)). Nevertheless, it is necessary to assess the quality of the \textit{cover} \(P^*\), since it is essential to reduce party fragmentation by achieving desirable levels of party and partisan discipline. But what are the desirable levels of party and partisan discipline?

In Fig. [3], I showed the CDF of the actual partisan discipline distribution during the analyzed period. The most desirable, or \textit{ideal}, partisan discipline distribution is shown as a dashed red line, which represents the situation where all partisans have discipline of 1.0. Thus, a new partisan discipline distribution, generated by Algorithm 2 with parameter \(\delta\) and defined by the random variable \(X_\delta\), is considered \textit{desirable} if its CDF \(F_{X_\delta}(x)\) is closer to the ideal than the actual one, defined by the random variable \(X_0\) and CDF \(F_{X_0}(x)\). More formally, considering that the area under the ideal CDF curve is 0 for both partisan and party discipline distributions, I propose the following definition:

\[ \text{Definition 6.1. A new discipline distribution defined by random variable } X_\delta \text{ and its CDF } F_{X_\delta}(x) \text{ is considered a desirable discipline distribution if } \int_0^1 F_{X_\delta}(x) \, dx - \int_0^1 F_{X_0}(x) \, dx > 0, \]

where \(F_{X_0}(x)\) is the CDF of the actual discipline distribution defined by random variable \(X_0\).

In other words, if the area under \(F_{X_\delta}(x)\) is smaller than the area under \(F_{X_0}(x)\), then \(X_\delta\) represents a desirable discipline distribution. Moreover, given that discipline ranges from 0 to 1, I propose the following lemma:

\[ \text{Lemma 6.2. Given a random variable } X_0 \text{ representing the actual discipline distribution and a random variable } X_\delta \text{ representing a new discipline distribution, if the expected value } E_{X_\delta}[x] \text{ of } X_\delta \text{ is higher than the expected value } E_{X_0}[x], \text{ then } X_\delta \text{ represents a desirable discipline distribution.} \]

\[ \text{Proof. From probability theory, we know that } E_X[x] = \int_0^\infty 1 - F_X(x) \, dx \text{ for a given random variable } X \text{ and CDF } F_X(x). \]

Since discipline has values from 0 to 1, we can write \(E_X[x] = 1 - \int_0^1 F_X(x) \, dx\), or \(\int_0^1 F_X(x) \, dx = 1 - E_X[x]\) for discipline distributions. Then, we can replace Definition 6.1 for \(1 - E_{X_\delta}[x] < 1 - E_{X_0}[x] \) or \(E_{X_\delta}[x] > E_{X_0}[x]\).

Thus, now I can formally define three binary quality signals for the new partisan configuration over the new set of parties \(P^*\) generated by Algorithm 2 with parameter \(\delta\). These signals indicate, respectively, if the new partisan configuration has desirable levels of partisan, party and overall discipline, and is defined as:

- \(Q_1\): 1 if the overall discipline \(d^0_\delta\) of the new configuration is greater than the overall discipline \(d^0_\delta\) of the actual configuration; 0 otherwise. Recall that the overall discipline \(d_\delta\) is defined in Equation 5.
- \(Q_2\): 1 if the expected (average) partisan discipline of the new configuration \(E_{X_\delta}^A[x]\) is greater than the expected party discipline of the actual configuration \(E_{X_0}^P[x]\) or, in other words, if the new partisan discipline distribution is a desirable partisan discipline distribution; 0 otherwise.
- \(Q_3\): 1 if the expected (average) party discipline of the new configuration \(E_{X_\delta}^P[x]\) is greater than the expected party discipline of the actual configuration \(E_{X_0}^P[x]\) or, in other words, if the new party discipline distribution is a desirable party discipline distribution; 0 otherwise.

7. RESULTS

In this section I show the results for \textit{ARRANGE} for a hundred equally distributed values of \(\delta\) between 0 and 1. From now on I will call a configuration \(c_\delta\) the distribution of partisans among parties generated by \textit{ARRANGE} for a particular \(\delta\) value. Moreover, in all plots I will indicate whether the quality signals described in previous section were 1 or 0. Namely, I will use yellow stars for results where all quality signals were 1 (\(Q_1 \land Q_2 \land Q_3\)), blue diamonds when only signals \(Q_2\) and \(Q_3\) were 1.
(Q_2 \land Q_3), red circles when only signal Q_3 was 1 and, finally, green squares when all quality signals were 0. In summary, ARRANGE was able to generate 31 distinct configurations that, compared with the status quo, have (i) a significantly smaller number of parties, (ii) higher discipline of partisans towards their parties and (iii) more even distributions of partisans into parties. Besides comparing with the status quo, ARRANGE will be compared with two random models: random-sq and random-\delta. The competitors can be summarized as:

(1) status quo. It is the existing state of affairs, i.e., the actual and historical situation in Brazil during the analyzed period.

(2) random-sq. It randomly redistributes the partisans among the existing parties. For each partisan \(a \in A\), the model randomly pick an option \([o_a := (a, p, sim)] \in O_a\) and assigns \(a\) to party \(p\). By using the options set \(O_a\) I guarantee that the party \(p\) allocated to \(a\) has voted for every proposition voted by \(a\), i.e., \(B_a \subseteq B_p\).

(3) random-\delta. Works in the same way as random-sq, but instead of allocating partisans to the actual set of parties \(P\), it randomly redistributes the partisans among the minimum set of parties \(P^*\) generated by Algorithm 2.

The random models are used to quantify the payoffs obtained by using ARRANGE when it generates the minimum party set \(P^*\) (random-sq) and it efficiently allocates partisans to the parties of this set (random-\delta). Although not always visible, for all results of both models I also show the 90% confidence interval.

In Fig. 6 I show the number of parties \(N^*\) generated by ARRANGE for different values of \(\delta\). I also show the status quo, i.e., the actual number of parties \(N\) of which partisans were members during the analyzed period, and the number of parties generated by random-sq. First, note that the number of parties generated by ARRANGE is significantly lower than the status quo, decreasing as \(\delta\) increases. Even when no partisans are allowed to decrease its discipline (\(\delta = 0\)), \(N^* = 22\), a number \(\approx 39\%\) lower than 36, the status quo. Moreover, ARRANGE was able to generate a configuration \((c_{0.15})\) with only 13 parties (\(\approx 64\%\) reduction) and with all quality signals equal to 1 \((Q_1 \land Q_2 \land Q_3)\), i.e., with overall, party and partisan disciplines greater than the status quo. When only \(Q_2\) and \(Q_3\) are 1 \((Q_2 \land Q_3)\), ARRANGE could generate a configuration \((c_{0.19})\) with \(N^* = 10\) parties, a \(\approx 72\%\) reduction. Finally, when only \(Q_3\) is 1 \((Q_3)\), ARRANGE could provide a \(\approx 86\%\) reduction in the number of parties by generating a configuration (e.g. \(c_{0.41}\)) with only 5 parties. It is worth mentioning that the number of parties in Brazil is so excessive that even random-sq was able to generate a configuration with fewer parties than the status quo.

Concerning discipline, I show in Fig. 7 the overall discipline (Fig. 7a) and the average party (Fig. 7b) and partisan (Fig. 7c) disciplines of the configurations produced by ARRANGE and its competitors. Observe that for all discipline metrics the results produced by ARRANGE are very similar with the status quo, even when all quality signals are 0. The discipline values decrease significantly only for \(\delta\) values close to 1, when the number of parties generated by ARRANGE is 2 or 1. It is also interesting to note that the random models are able to produce configurations with considerably high levels of discipline, which shows that Brazilian political parties are, on average, very similar to each other.

In order to analyze how well distributed are the partisans among parties, we compute the Gini coefficient \([53]\) for each configuration generated by ARRANGE and its competitors. The Gini coefficient was initially proposed to describe the income inequality in a population \([53]\). It assumes values from 0, which expresses perfect equality, where all parties have the same number of partisans, to 1, which expresses maximal inequality among values, where all partisans are allocated to a single party. Observe in Fig. 8 that ARRANGE is able to produce configurations in which the partisans are more evenly distributed than the status quo for all values of \(\delta\). While the Gini coefficient is \(\approx 0.64\) for the status quo, it decreased to \(\approx 0.43\) for \(c_{0.15}\) \((Q_1 \land Q_2 \land Q_3)\), to \(\approx 0.32\) for configuration \(c_{0.41}\) \((Q_3)\), and to \(\approx 0\) for \(c_{0.5}\) (all quality signals equal to 0, but with discipline values similar with the status quo). Model random-sq has a slightly high Gini coefficient than the status quo because a few
big parties are more likely to randomly receive new partisans, since they appear as an option in the option sets $O_a$ for all $a \in A$.

Another relevant characteristic to be measured in political party systems is switching, i.e., party changes among partisans. As stated by [43], switching effectively destroys the meaning of party labels, raises voters’ information costs, and eliminates party accountability, being a threat to the very core of democratic representation. Thus, in Fig. 9 I plot the number of party changes among partisans that would occur if the configurations generated by ARRANGE and the random models were the reality. Observe that in all scenarios generated by ARRANGE the number of changes is lower than the status quo, random-sq and random-δ. This is another evidence of the importance of reducing party fragmentation to have ideologically well defined parties.

Now I will take a closer look at particular configurations generated by ARRANGE, namely $c_{0.15}$, $c_{0.19}$ and $c_{0.41}$. These are the configurations that have the lowest number of parties and achieved, respectively, quality signals $Q_1 \land Q_2 \land Q_3$, $Q_2 \land Q_3$ and only $Q_3$. In Fig. 10, I show the number of active parties per year for the status quo and these three configurations. First, observe that the number of parties changes constantly over the years in Brazil, this being a harmful consequence of its highly fragmented party system. On the other hand, observe that the configurations generated by ARRANGE are (i) much more stable over the years and (ii) have a significantly smaller number of parties than the status quo.
Besides counting the actual number of parties per year, it is also important to measure the effective number of parties. As described previously, it is a concept which provides for an adjusted number of political parties in a country’s party system, weighting the partisan count per party by their relative strength [38]. In our case, the relative strength refers to their seat share in the parliament. This measure is especially useful to detect trends toward fewer or more numerous parties over time [38]. The number of parties equals the effective number of parties only when all parties have equal strength. In any other case, the effective number of parties is lower than the actual number of parties. It is also a frequent metric for the fragmentation of a party system [9]. Moreover, although several indexes for computing the effective number of parties exist [41], in this paper I use the Golosov index $N_p = \sum_{i=1}^{N} \left(1 + \left(\frac{s_i^2}{s_1}\right) - s_1\right)^{-1}$, where $N$ is the actual number of parties, $s_i$ is the proportional share of each party $p_i$, and $s_1$ is the highest share of a party [41]. For the best of my knowledge,
8. DISCUSSION

In practical terms, the main contribution of ARRANGE to a government and its population is the ability to provide a quantitative assessment of fragmentation in its party system. Given that the
the effective number of parties is expected to reflect the number of issue (or ideological) dimensions in a country \[12\]. Highly fragmented party systems overestimate this number of issue dimensions, providing a distorted view to the population that harms democracy \[47\]. Thus, by mapping plenary votes into ideological preferences, ARRANGE quantitatively provides an “ideal” number of parties for a country given the ideological preferences of its congressmen, revealing the presumable true number of issue dimensions that exists in this particular country. I used the expression "presumable true" because, as verified by \[8\], it is not always true that a disciplined and cohesive party represents an ideological cleavage (or group) in society, i.e., its members may simply be a group of congressmen obeying its leader in order to obtain a particular benefit.

It is also worth mentioning that ARRANGE also provides the list of parties that should ideally exist and which partisans should be their members. Although I know a democratic government cannot implement this solution easily, it can be used to support significant reforms in its political system. With this in mind, I could not fail to mention in this paper that one of the possible reasons for Brazil’s high level of party fragmentation is the so called fundo partidário, which are funds distributed by the federal government to Brazilian political parties for them to spend indiscriminately. A share of the amount paid by the federal government through the fundo partidário is the same for every party, but another part is proportional to the number of elected congressmen, senators and governors by each organization. In 2014, PT, the party with the highest share (16.5%), received R$50,314,999.19 million reais from the fund. On the other hand, PROS, the party with the lowest share (0.16%), received R$493,873.68 thousand reais in 2014 \[53\]. Consider that, in 2014, the exchange rate of the real varied from US$2.19 to US$2.72 U.S. dollars. It is out of the scope of this paper to point fundo partidário as the main culprit for Brazil’s high level of party fragmentation, but it poses as a clear incentive for the creation of many parties in Brazil.

In spite of the fact that this work considers Brazil as its use case, the methods and results shown here can be easily replicated to other countries that have highly fragmented party systems. Carsten Anckar studied party system fragmentation in 77 countries \[9\] and reported high levels of fragmentation for many countries besides Brazil, such as Bolivia, Bulgaria, Denmark, Ecuador, Finland, France, Guatemala, India, Israel, Italy, Netherlands and Thailand. Nevertheless, it is important to point out that this work was motivated and eased by the Open Data initiative of the Brazilian government, that provides public data related to politics and also to many other areas, such as demographics, government spending, budget and road accidents.

9. CONCLUSIONS

In this work, I proposed the method ARRANGE to assess and reduce fragmentation in multi-party political systems. From roll votes data of partisans and their respective party leaders, ARRANGE redistributes the partisans into new parties considering two conflicting objectives: to minimize the number of parties and to maximize party discipline. When applied to Brazilian historical roll call data, ARRANGE was able to generate 23 distinct configurations that, compared with the status quo, have (i) a significantly smaller number of parties, (ii) higher discipline of partisans towards their parties and (iii) more even distributions of partisans into parties. These results show that Brazil has and had many redundant parties, i.e., parties that are very similar ideologically. Thus, if today Brazil has one of the highest levels of party system fragmentation in the World \[8\] \[9\], this work proved it could be much lower. Finally, it is important to point out that ARRANGE is a general method and could be directly applied to analyze fragmentation in any of the many highly fragmented party systems that exists in the world \[9\].

REFERENCES

Saad-Filho A. Mass Protests under 'Left Neoliberalism': Brazil, June-July 2013. Critical Sociology. 2013 Aug;39(5):657–669. Available from: \[http://crs.sagepub.com/cgi/doi/10.1177/0896920513501906\].

Mische A. "Come to the streets, but without parties": The challenges of the new Brazilian protests; 2013. Available from: \[https://mobilizingideas.wordpress.com/2013/09/04/come-to-the-streets-but-without-parties-the-challenges-of-the-new-brazilian-protests/\]
Downie A. Brazil crowds up but stadium usage still problematic; 2014. Available from: [http://in.reuters.com/article/2014/12/23/soccer-world-brazil-stadiums-idINKBN0K10R920141223](http://in.reuters.com/article/2014/12/23/soccer-world-brazil-stadiums-idINKBN0K10R920141223)

Mainwaring S. Rethinking Party Systems in the Third Wave of Democratization: The Case of Brazil. Stanford University Press; 1999. Available from: [http://books.google.com.br/books?id=RLU72_YerAcC](http://books.google.com.br/books?id=RLU72_YerAcC)

Mainwaring S, Shugart MS. Presidentialism and Democracy in Latin America. Cambridge Studies in Comparative Politics. Cambridge University Press; 1997. Available from: [http://books.google.com.br/books?id=663EWRmlMoC](http://books.google.com.br/books?id=663EWRmlMoC)

Ames B. The Deadlock of Democracy in Brazil. Interests, Identites, and Institutions in Comparative Politics. University of Michigan Press; 2009. Available from: [http://books.google.com.br/books?id=RLU72_YerAcC](http://books.google.com.br/books?id=RLU72_YerAcC)

Bohn S. Institutional, societal, and economic determinants of party system size: evidence from Brazil. Political Science Review. 2014 Jun; Available from: [http://ips.sagepub.com/cgi/doi/10.1177/0192512114532968](http://ips.sagepub.com/cgi/doi/10.1177/0192512114532968)

Figueiredo AC, Limongi F. Presidential Power, Legislative Organization, and Party Behavior in Brazil. Comparative Politics. 2000 Jan;32(2):151. Available from: [http://www.jstor.org/stable/422395?origin=crossref](http://www.jstor.org/stable/422395?origin=crossref)

Anckar C. Size and Party System Fragmentation; 2000.

Colomer JM, Tworzecki H. cleavages, issues and parties: a critical overview of the literature; 2005.

Neto OA, Cox GW. Electoral Institutions, Cavlage Structures, and the Number of Parties. American Journal of Political Science. 1997;41(1):pp. 149–174. Available from: [http://www.jstor.org/stable/2111712](http://www.jstor.org/stable/2111712)

TAAGEPERA R, GROFMAN B. Rethinking Duverger’s Law: Predicting the Effective Number of Parties in Plurality and PR Systems - Parties Minus Issues Equals One*. European Journal of Political Research. 2006 May;13(4):341–352. Available from: [http://doi.wiley.com/10.1111/j.1475-6765.1985.tb00130.x](http://doi.wiley.com/10.1111/j.1475-6765.1985.tb00130.x)

Duverger M. Political Parties: Their Organization and Activity in the Modern State. Science editions. Wiley; 1963. Available from: [http://books.google.com.br/books?id=KhcOAAAAQAAJ](http://books.google.com.br/books?id=KhcOAAAAQAAJ)

Lipset SM, Rokkan S. Party systems and voter alignments: cross-national perspectives. International yearbook of political behavior research. Free Press; 1967. Available from: [http://books.google.com.br/books?id=EpNCAAAAIAAJ](http://books.google.com.br/books?id=EpNCAAAAIAAJ)

Stoll H. Social Cleavages and the Number of Parties: How the Measures You Choose Affect the Answers You Get; Comparative Political Studies. 2007

Wu C, Rosenfeld R, Clermont G. Using Data-Driven Rules to Predict Mortality in Severe Community Acquired Pneumonia. PLoS ONE. 2014 Apr;9(4):e89053. Available from: [http://dx.plos.org/10.1371/journal.pone.0089053](http://dx.plos.org/10.1371/journal.pone.0089053)

Silva TH, Vaz de Melo POS, Almeida JM, Salles J, Loureiro AAF. Revealing the City That We Cannot See. ACM Transactions on Internet Technology. 2014 Dec;14(4):1–23. Available from: [http://dl.acm.org/citation.cfm?doid=2699996.2677208](http://dl.acm.org/citation.cfm?doid=2699996.2677208)

Silva TH, Vaz de Melo POS, Almeida J, Musolesi M, Loureiro A. You are What you Eat (and Drink): Identifying Cultural Boundaries by Analyzing Food & Drink Habits in Foursquare. In: Proceedings of 8th AAAI Intl. Conf. on Weblogs and Social Media (ICWSM 2014); 2014. Available from: [http://arxiv.org/abs/1404.1009](http://arxiv.org/abs/1404.1009)

Park M, Weber I, Naaman M, Vieweg S. Understanding Musical Diversity via Online Social Media; 2015. In: Proceedings of 9th AAAI Intl. Conf. on Weblogs and Social Media (ICWSM 2015); 2015. Available from: [https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10570](https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10570)

Zhang J, Wang FY, Wang K, Lin WH, Xu X, Chen C. Data-Driven Intelligent Transportation Systems: A Survey. IEEE Transactions on Intelligent Transportation Systems. 2011 Dec;12(4):1624–1639. Available from: [http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5959985](http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5959985)

Yin S, Ding SX, Xie X, Luo H. A Review on Basic Data-Driven Approaches for Industrial Process Monitoring. IEEE Transactions on Industrial Electronics. 2014 Nov;61(11):6418–6428. Available from: [http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6717991](http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6717991)

de Vries NJ, Carlson J, Moscato P. A Data-Driven Approach to Reverse Engineering Customer Engagement Models: Towards Functional Constructs. PLoS ONE. 2014 Jul;9(7):e102768. Available from: [http://dx.plos.org/10.1371/journal.pone.0102768](http://dx.plos.org/10.1371/journal.pone.0102768)

Shor B, McCarty NM. The Ideological Mapping of American Legislatures. American Political Science Review. 2011;105:530–551.

Poole KT, Rosenthal HL. Ideology and Congress. American Studies. Transaction Publishers; 2011. Available from: [https://books.google.com.br/books?id=eOuZqa7UbqUC](https://books.google.com.br/books?id=eOuZqa7UbqUC)

Clinton JD, Jackman S, Rivers D. The Statistical Analysis of Roll Call Data. American Political Science Review. 2004;98:355–370. Available from: [http://journals.cambridge.org/abstract_S0003055404000194](http://journals.cambridge.org/abstract_S0003055404000194)

Sobkowicz P, Kaschesky M, Bouchard G. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. Government Information Quarterly. 2012;29:470–479.

Pennacchiotti M, Popescu AM. Democrats, republicans and starbucks afficionados. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’11; 2011. p. 430. Available from: [http://dl.acm.org/citation.cfm?doid=2020408.2020477](http://dl.acm.org/citation.cfm?doid=2020408.2020477)
Akoglu L. Quantifying Political Polarity Based on Bipartite Opinion Networks; 2014. Available from: https://aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8073/8100

Gao ZK, Yang YX, Fang PC, Jin ND, Xia CY, Hu LD. Multi-frequency complex network from time series for uncovering oil-water flow structure. Scientific Reports. 2015 Feb;5:8222. Available from: http://www.nature.com/doifinder/10.1038/srep08222

Newman MEJ. Communities, modules and large-scale structure in networks. Nature Physics. 2011 Dec;8(1):25–31. Available from: http://www.nature.com/doifinder/10.1038/nphys2162

Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. Science. 2014 Mar;343(6176):1203–1205. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1248506

Poole KT, Rosenthal H. Patterns of Congressional Voting. American Journal of Political Science. 1991;35(1):228–278.

Snyder JM, Groseclose T. Estimating Party Influence on Roll Call Voting: Regression Coefficients versus Classification Success. American Political Science Review. 2001;95:689–698.

Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. Science. 2014 Mar;343(6176):1203–1205. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1248506

Clinton JD, Meirowitz A. Integrating Voting Theory and Roll Call Analysis: A Framework. Political Analysis. 2003;11:381–396.

Fang Y, Si L, Somasundaram N, Yu Z. Mining Contrastive Opinions on Political Texts using Cross-Perspective Topic Model Categories and Subject Descriptors. Proceedings of the fifth ACM international conference on Web search and data mining. 2012;p. 63–72.

Riker WH. The Number of Political Parties: A Reexamination of Duverger’s Law. Comparative Politics. 1976;9(October):93–106.

Aldecoa R, Marín I. Deciphering Network Community Structure by Surprise. PLoS ONE. 2011 Sep;6(9):e24195. Available from: http://dx.plos.org/10.1371/journal.pone.0024195

Fortunato S. Community detection in graphs. Physics Reports. 2010 Feb;486(3-5):75–174. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0370157309002841

Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E. 2004 Feb;69(2):026113. Available from: http://link.aps.org/doi/10.1103/PhysRevE.69.026113

Aldecoa R, Marín I. Deciphering Network Community Structure by Surprise. PLoS ONE. 2011 Sep;6(9):e24195. Available from: http://dx.plos.org/10.1371/journal.pone.0024195

Ranney A. Curing the Mischiefs of Faction: Party Reform in America. Jefferson memorial lectures. University of California Press; 1975. Available from: http://books.google.com.br/books?id=6mjBAVjJaVCc

Jolliffe IT. Principal Component Analysis. 2nd ed. Springer; 2002.

Chvatal V. A Greedy Heuristic for the Set-Covering Problem. Mathematics of Operations Research. 1979;4(3):pp. 233–235. Available from: http://www.jstor.org/stable/3689577

Alon N, Moshkovitz D, Safra S. Algorithmic construction of sets for k -restrictions. ACM Transactions on Algorithms. 2006 Apr;2(2):153–177. Available from: http://portal.acm.org/citation.cfm?doid=1150334.1150336

Feige U. A threshold of In n for approximating set cover. Journal of the ACM. 1998 Jul;45(4):634–652. Available from: http://portal.acm.org/citation.cfm?doid=285053.285059

Billingsley P. Probability and measure. Wiley series in probability and mathematical statistics. Wiley; 1986. Available from: http://books.google.com.br/books?id=Q2IPQAAMAAJ

Ceriani L, Verme P. The origins of the Gini index: extracts from Variabilit` a e Mutabilit` a (1912) by Corrado Gini. Journal of Economic Inequality. 2012;10(3):421–443.

Justiça Eleitoral do Brasil. Distribuição do Fundo Partidário; 2014. Available from: http://www.jus.br/justica-eleitoral/fundo-partidario/distribuicao-2014