
A VERIFIABLY SECURE AND PROPORTIONAL COMMITTEE ELECTION RULE

A PREPRINT

Alfonso Cevallos
 Web 3.0 Technologies Foundation
 Zug, Switzerland
 alfonso@web3.foundation

Alistair Stewart
 Web 3.0 Technologies Foundation
 Zug, Switzerland
 alistair@web3.foundation

July 28, 2020

ABSTRACT

The property of proportional representation in approval-based committee elections has appeared in the social choice literature for over a century, and is typically understood as avoiding the underrepresentation of minorities. However, we argue that the security of some distributed systems is directly linked to the opposite goal of *avoiding the overrepresentation* of any minority, a goal not previously formalized which leads us to an optimization objective known as *maximin support*, closely related to the axiom of *proportional justified representation* (PJR). We provide a new inapproximability result for this objective, and propose a new election rule inspired in Phragmén’s methods that achieves a) a constant-factor approximation guarantee for the objective, and b) the PJR property. Furthermore, a structural property allows one to quickly *verify* that the winning committee satisfies the two aforementioned properties, even if the algorithm was executed by an untrusted party who only communicates the output. Finally, we present an efficient post-computation that, when paired with any approximation algorithm for maximin support, returns a new solution that a) preserves the approximation guarantee, b) satisfies PJR, and c) can be efficiently verified to satisfy PJR.

Our work is motivated by an application on blockchains that implement *nominated proof-of-stake* (NPoS), where the community must elect a committee of validators to participate in its consensus protocol, and where fighting overrepresentation protects the system against attacks by an adversarial minority. Our election rule enables a validator election protocol with formal and verifiable guarantees on security and proportionality. We propose a specific protocol that can be successfully implemented in spite of the stringent time constraints of a blockchain architecture, and that will be the basis for an implementation in the *Polkadot* network, launched in 2020.

Keywords social choice · approval-based committee election · approximation algorithms · proof-of-stake · blockchain

1 Introduction

The property of proportional representation in approval-based committee elections has been discussed in the literature of computational social choice for a long time, with new mathematical formulations recently proposed to formalize it. The property is typically understood as ensuring that no minority in the electorate is underrepresented by the winning committee. In this paper we complement this notion with an analysis of the opposite goal: fighting overrepresentation of any minority. We establish such objective formally as an optimization problem, related to the axiom of proportional justified representation (PJR), and propose efficient approximation algorithms for it.

Our work is motivated by the problem of electing participants to the consensus protocol – whom we refer to as *validators* – in a decentralized blockchain network, where overrepresentation may affect security. More precisely, under a proof-of-stake mechanism it is assumed that a majority of the native token is in the hands of honest actors, and this proportion of honest actors is to be preserved among the elected validators; hence, if the validator committee is selected via an election process where users have a vote strength proportional to their token holdings, an adversarial

minority should not be able to gain overrepresentation. We propose an efficient election rule for the selection of validators which offers strong guarantees on security and proportional representation, and successfully adapts to the distributed and resource-limited nature of blockchain protocols.

Background on Nominated Proof-of-Stake. Many blockchain projects launched in recently years substitute the inefficient proof-of-work (PoW) component of Nakamoto’s consensus protocol [14] with proof-of-stake (PoS), which is much more efficient and offers security in the presence of a dishonest minority of users, where this minority is to be measured relative to token holdings. We focus on *nominated proof-of-stake* (NPoS), a variant of PoS where at regular intervals a committee of validators is elected by the community to participate in the consensus protocol. Users are free to become validator candidates, or become *nominators*, who approve of candidates that they trust and back them with their tokens. Both validators and nominators thus put their tokens at stake and receive economic rewards (coming from transaction fees and minting of new tokens) on a pro-rata basis, and may also be penalized in case a validator shows negligent or adversarial behavior. Polkadot [6] is a blockchain network launched in 2020 which implements NPoS.

Having the nominator role in the system allows for a massive number of users to indirectly participate in the consensus protocol, in contrast to the number of elected validators which due to operational limitations remains bounded, typically in the order of hundreds. This allows for a large amount of stake to back validators, much higher than any single user’s holding. Intuitively, the higher the stake backing the validators, the higher the **security** of the network; we make this intuition formal further below. As such, one of the goals of the validator election rule is to maximize the stake backing all elected validators. The second, equally important goal is **proportionality**, i.e. picking a committee where nominators are represented in proportion to their stake. We highlight that diverse preferences and factions will naturally arise among nominators for reasons that range from economic and technical to political, geographical, etc. Such diversity of points of view is expected and welcome in a distributed system, and having a committee with proportionality guarantees helps keep the system decentralized and its users satisfied.

Proportional representation in multiwinner elections. Let us now formalize this validator election and its two stated goals. For simplicity, we consider that only nominators have stake, not candidates, and that a nominator’s approval of candidates is dichotomous, i.e. she provides an unranked list of approved candidates of any size. This leads to a vote-weighted, approval-based committee election scheme. There are finite sets N and C of voters (nominators) and candidates respectively, and every voter $n \in N$ provides a list $C_n \subseteq C$ of approved candidates and has a vote strength s_n (their stake). There is also a target number $1 \leq k < |C|$ of candidates to elect.

One of our goals corresponds to the classical notion of proportional representation, i.e. the committee should represent each group in the electorate proportional to their aggregate vote strength, with no minority being underrepresented. Electoral system designs that achieve some version of proportional representation have been present in the literature of social choice for a very long time. Of special note is the work of Scandinavian mathematicians Edvard Phragmén and Thorvald Thiele in the late nineteenth century [17, 18, 19, 20, 24, 12]. Several axioms have been recently proposed to define the property mathematically; we mention the most relevant ones. *Justified representation* (JR) [2] states that if a group $N' \subseteq N$ of voters is cohesive enough in terms of candidate preferences and has a large enough aggregate vote strength, then it has a justified claim to be represented by a member of the committee. *Proportional justified representation* (PJR) [21] says that such a group N' deserves not just one but certain number of representatives according its vote strength, where a committee member is said to represent the group as long as it represents any voter in it. Finally, *extended justified representation* (EJR) [2] strengthens this last condition and requires not only that N' have enough representatives as a group, but some voter in it must have enough representatives individually. It is known that EJR implies PJR, and PJR implies JR, but converse implications are not true [21]. For each of these properties, we say that a committee voting rule satisfies said property if its output committee is always guaranteed to satisfy the property for any input instance. While the most common voting rules usually satisfy JR, they fail to satisfy the stronger properties of PJR and EJR, and up to recently there were no known efficient voting rules that satisfy the latter two. For instance, the *proportional approval voting* (PAV) method [24, 12] proposed by Thiele satisfies EJR but is NP-hard to compute, while efficient heuristics based on it, such as reweighted approval voting, fail PJR [4, 23, 2]. Only very recently have efficient algorithms that achieve PJR or EJR finally been proposed [5, 22, 3, 16].

Among these axioms, we set to achieve **PJR**, defined formally in Section 2, for two reasons. First, because it is more *Sybil resistant* [7] than JR, meaning that a strategic voter may be incentivized to assume several nominator identities in the network under JR, but not under PJR. Second, because PJR seems to be most compatible with our security objective. Indeed, as argued in [16] and [13], the PJR and EJR axioms correspond to different notions of proportionality: while EJR is primarily concerned with the general welfare or satisfaction of the voters, PJR considers proportionality of the voters’ decision power, and our security objective aligns with the latter notion as we explain below.

The security goal: fighting overrepresentation. In a PoS-based consensus protocol, we want a guarantee that as long as most of the stake is in the hands of actors that behave honestly or rationally, carrying any attack will be very costly. If we assume that the attack requires control of some minimum number of validators to succeed, the adversary would

need to use its stake to get these validators elected. The security level is hence proportional to how difficult it is for an entity to gain overrepresentation in the validator election protocol. Interestingly, this is in stark contrast to the classical approach taken in proportional representation axioms, that only seek to avoid underrepresentation.

Recall that each voter $n \in N$ has a vote strength s_n and a list of approved candidates $C_n \subseteq C$. Suppose we want to make it as expensive as possible for an adversary to gain a certain number $1 \leq r \leq k$ of representatives into the k -validator committee, and that few to none of the honest nominators trust these candidates. Then, our goal would be to elect a committee $A \subseteq C$ that maximizes $\min_{A' \subseteq A, |A'|=r} \sum_{n \in N: C_n \cap A' \neq \emptyset} s_n$. This gives a different objective for each value of threshold r . For example, for $r = 1$, maximizing this objective is equivalent to the classical multiwinner approval voting: selecting the k candidates $c \in C$ with highest total approval $\sum_{n \in N: c \in C_n} s_n$. If we are only concerned about a particular threshold, e.g. $r = \lceil (k+1)/3 \rceil$ for a 34% attack¹, then we can fix the corresponding objective. However, different types of attacks require different thresholds, and some attacks succeed with higher probability with more attacking validators. Hence, a more pragmatic approach is to incorporate the threshold into the objective and maximize *the least possible per-validator cost over all thresholds*, i.e.

$$\text{Maximize } \min_{A' \subseteq A, |A'| \neq \emptyset} \frac{1}{|A'|} \sum_{n \in N: C_n \cap A' \neq \emptyset} s_n, \quad \text{over all committees } A \subseteq C \text{ with } |A| = k. \quad (1)$$

We establish in Lemma 6 that this objective is equivalent to the **maximin support objective**, recently introduced by Sánchez-Fernández et al. [22], and which we thus set to optimize. We define it formally in Section 2. The authors in [22] remark that in its exact version, maximin support is equivalent to another objective, maxPhragmen, devised by Phragmén and recently analyzed in [5], and in this last paper it is shown that maxPhragmen is NP-hard and incompatible with EJR. Thus, the same hardness and incompatibility with EJR holds true for our security objective. To the best of our knowledge, the approximability of maximin support has not previously been studied.

Our contribution. Our security analysis in the election of validators leads us to pursue the goal of fighting overrepresentation, and more specifically to the maximin support objective. Conversely, we formalize our goal of proportionality to the PJR property, which fights underrepresentation. We show in this paper that these two objectives complement each other well, and prove the existence of efficient election rules that achieve guarantees for both objectives.

Theorem 1. *There is an efficient election rule for approval-based committee elections that simultaneously achieves the PJR property and a 3.15-factor approximation guarantee for the maximin support objective.*

Our proposed election rule is inspired in the seqPhragmen method [5], and to the best of our knowledge corresponds to the first analysis of approximability for a Phragmén objective. In contrast, several approximation algorithms for Thiele objectives have been proposed; see [13] for a survey. Our optimization approach to proportional representation also provides new tools to discern between different election rules. For instance, MMS [22] and seqPhragmen [5] are two efficient rules that achieve the PRJ property, but we show that while the former provides a constant-factor approximation guarantee for maximin support, the latter does not (Section 3). Next comes the question of applicability: the blockchain architecture adds very stringent time constraints to computations, and may render all but the simplest algorithms unimplementable.² This is because, in order to reach consensus on a protocol, validators typically need to execute it individually, and no further blocks can be produced until most validators finish the computation and agree on the output. However, if an output can be *verified* much faster than it can be computed, then it is possible to dump the computation to *off-chain workers*, who execute it privately and separately from block production, leaving only the output verification task to validators. This is the case for our election rule.

Theorem 2. *There is a linear-time test that takes as input an election instance and an arbitrary solution to it, such that if the test passes then the input solution satisfies the PJR property and a 3.15-factor approximation guarantee for the maximin support objective. Moreover, the output of the election rule mentioned in Theorem 1 always passes this test.*

We believe this to be the most important feature of our proposed election rule relative to others in the literature. Such efficient verification helps ensure transparency in a decentralized governance process, and might be of independent interest. In the context of our motivating application, it makes our election rule implementable into a fast blockchain validator election protocol which provides strong and verifiable guarantees on security and proportionality. We provide details on such a protocol in Section 6, which will be the basis for an implementation in the Polkadot network. Finally,

¹This is the minimum threshold required to carry on a successful attack in classical Byzantine fault tolerant protocols [15].

²For instance, the PoS-based project EOS uses multiwinner approval voting to elect a validator committee, which is a highly efficient election rule but is known to perform poorly in terms of proportionality. This has lead to user dissatisfaction and claims of excessive centralization. For an analysis, we refer to the blogpost “EOS voting structure encourages centralization” by Priyeshu Garg at <https://cryptoslate.com/eos-voting-structure-encourages-centralization/>

we derive from the new election rule a post-computation which, when paired with any approximation algorithm for the maximin support problem, makes it also satisfy PJR in a black-box manner.

Theorem 3. *There is an efficient computation that takes as input an election instance and an arbitrary solution to it, and outputs a new solution which a) is no worse than the input solution in terms of the maximin support objective, b) satisfies the PJR property, and in particular c) can be efficiently tested to satisfy the PJR property.*

Organization of the paper. We start with a thorough complexity analysis of the maximin support problem in Section 3, where we exhibit several approximation algorithms for it, and prove as well that it does not admit a PTAS³ unless $P=NP$ so constant-factor approximations are best possible. Interestingly, many of our approximation analyses are based on network flow theory. In Section 4 we propose a new heuristic, Phragmms, inspired in seqPhragmen [5] but allowing for a more robust analysis and better guarantees than the latter both in terms of the PJR property and the maximin support objective. We use it to prove Theorem 1 and obtain the fastest known algorithm that simultaneously guarantees a) the PJR property and b) a constant-factor approximation for maximin support.

In Section 5 we prove Theorems 2 and 3, and explore how the two aforementioned guarantees can be efficiently tested by a *verifier*, even if the algorithm is privately executed by an untrusted *prover* who only communicates the solution. To do so, we define a parametric version of the PJR property, and a notion of local optimality for solutions. Finally, in Section 6 we propose a validator election protocol for an NPoS-based blockchain network, where we exploit our results. This proposal will be the basis for an implementation in Polkadot.

2 Preliminaries

Throughout the paper we consider the following approval-based multiwinner election instance. We are given a bipartite approval graph $G = (N \cup C, E)$ where N is a set of voters and C is a set of candidates. We are additionally given a vector $s \in \mathbb{R}_{\geq 0}^N$ of vote strengths, where s_n is the strength of n 's vote, and a target number k of candidates to elect, where $0 < k \leq |C|$. For each voter $n \in N$, $C_n := \{c \in C : nc \in E\}$ represents her approval ballot, i.e. the subset of candidates that n approves of, and for each candidate $c \in C$ we denote by $N_c := \{n \in N : nc \in E\}$ the set of voters approving c , where nc is shorthand for edge $\{n, c\}$. To avoid trivialities, we assume that graph G in the input has no isolated vertices. For any $c \in C \setminus A$, we write $A + c$ and $A - c$ as shorthands for $A \cup \{c\}$ and $A \setminus \{c\}$ respectively.

Proportional justified representation (PJR). The PJR property was introduced in [21] for voters with unit vote strength. We present its natural generalization to arbitrary vote strengths. A committee $A \subseteq C$ of k members satisfies PJR if there is no group $N' \subseteq N$ of voters and integer $0 < r \leq k$ such that:

$$\text{a) } \sum_{n \in N'} s_n \geq \frac{r}{k} \sum_{n \in N} s_n, \quad \text{b) } |\cap_{n \in N'} C_n| \geq r, \quad \text{and c) } |A \cap (\cup_{n \in N'} C_n)| < r.$$

In words, if there is a group N' of voters with at least r commonly approved candidates, and enough aggregate vote strength to provide each of them with a vote support of value $\hat{t} := \sum_{n \in N} s_n / k$, then this group has a justified right to be represented by at least r members in committee A , though not necessarily commonly approved. Notice that \hat{t} is an upper bound on the average vote support that voter set N can possibly provide to any committee of k members.

Maximin support objective. For the given instance, we consider a solution consisting of a tuple (A, w) , where $A \subseteq C$ is a committee of k elected candidates, and $w \in \mathbb{R}_{\geq 0}^E$ is a vector of non-negative edge weights that represents a fractional distribution of each voter's vote among her approved candidates.⁴ For instance, for voter n this distribution may assign a third of s_n to c_1 and two thirds of s_n to c_2 , where $c_1, c_2 \in C_n$. Vector w is considered *feasible*⁵ if

$$\sum_{c \in C_n} w_{nc} \leq s_n \quad \text{for each voter } n \in N. \quad (2)$$

In our analyses, we will also consider *partial* committees, with $|A| \leq k$. If $|A| = k$, we call it *full*. All solutions (A, w) in this paper are assumed to be feasible and full unless stated otherwise. Given a (possibly partial, unfeasible) solution (A, w) , we define the *support* over the committee members as

$$\text{supp}_w(c) := \sum_{n \in N_c} w_{nc} \quad \text{for each } c \in A, \quad \text{and} \quad \text{supp}_w(A) := \min_{c \in A} \text{supp}_w(c), \quad (3)$$

³A *polynomial time approximation scheme* (PTAS) for an optimization problem is an algorithm that, for any constant $\varepsilon > 0$ and any given instance, returns a $(1 + \varepsilon)$ -factor approximation in polynomial time.

⁴This weight vector is related to the notions of *support distribution function* in [22] and *price system* in [16]. In particular, all voting rules considered in this paper are *priceable*, as defined in [16].

⁵Intuitively, a feasible solution (A, w) should also observe $w_{nc} = 0$ for each edge nc with $c \notin A$. However, as this constraint can always be enforced in post-computation, we ignore it so that the feasibility of a vector w is independent of any committee.

where we use the convention that $\text{supp}_w(\emptyset) = \infty$ for any weight vector $w \in \mathbb{R}_{\geq 0}^E$. The maximin support objective, introduced in [22], asks to maximize the least member support $\text{supp}_w(A)$ over all feasible full solutions (A, w) .

Balanced solutions. For a fixed committee A , a feasible weight vector $w \in \mathbb{R}_{\geq 0}^E$ that maximizes $\text{supp}_w(A)$ can be found efficiently. In this paper we seek additional desirable properties on a weight vector which can still be achieved efficiently. We say that a feasible $w \in \mathbb{R}_{\geq 0}^E$ is *balanced* for A , or that (A, w) is a balanced solution, if

1. it maximizes the sum of member supports, $\sum_{c \in A} \text{supp}_w(c)$, over all feasible weight vectors, and
2. it minimizes the sum of supports squared, $\sum_{c \in A} (\text{supp}_w(c))^2$, over all vectors that observe the point above.

In other words, a balanced weight vector maximizes the sum of supports and then minimizes their variance. In the next lemma, whose proof is delayed to Appendix E, we establish some key properties that we exploit in our analyses.

Lemma 4. *Let (A, w) be a balanced, possibly partial solution. Then,*

1. *for each $1 \leq r \leq |A|$, vector w simultaneously maximizes $\min_{A' \subseteq A, |A'|=r} \sum_{c \in A'} \text{supp}_{w'}(c)$ over all feasible weight vectors $w' \in \mathbb{R}_{\geq 0}^E$;*
2. *for each voter $n \in \cup_{c \in A} N_c$, it must hold that $\sum_{c \in A \cap C_n} w_{nc} = s_n$; and*
3. *for each voter $n \in \cup_{c \in A} N_c$ and each candidate $c \in A \cap C_n$ with $w_{nc} > 0$, it must hold that $\text{supp}_w(c) = \text{supp}_w(A \cap C_n) := \min_{c' \in A \cap C_n} \text{supp}_w(c')$.*

Furthermore, a feasible solution (A, w) is balanced if and only if it observes properties 2 and 3 above.

Notice that by setting $r = 1$ on the first point, we obtain that balanced vector w indeed maximizes the least member support $\text{supp}_w(A)$ over all feasible weight vectors. More generally, for each r the quantity defined in the first point defines a lower bound on the cost for an adversary to get r representatives in the validator committee in NPoS, so maximizing these objectives for all thresholds r aligns with our security objective as it makes any attack as costly as possible. The second point follows from the fact that the sum of member supports is maximal, so all the available voters' vote must be distributed to candidates in A . The third point is a consequence of having the supports as evenly distributed as possible within A : if $c, c' \in A \cap C_n$ and candidate c has a higher support than c' , then none of n 's vote can go to c , as all of it must be assigned to c' or other members with low support.

In Appendix A we present new algorithms for computing a balanced weight vector for a given committee A . In particular, we prove that one can be found in time $O(|E| \cdot k + k^3)$ using parametric flow techniques, which to the best of our knowledge is the current fastest algorithm in the literature even for the simpler problem of maximizing $\text{supp}_w(A)$.

Remark 5. *In the remainder of the paper, we denote by Bal the time complexity of finding a balanced weight vector, which will depend on the precise algorithm used.*

Equivalence of objectives. We now establish that our security objective (1) is indeed equivalent to maximin support. For this we use the fact that, in view of point 1 in Lemma 4, in the maximin support objective we can reduce the solution space to only balanced solutions without loss of generality. The proof of the next lemma is delayed to Appendix E.

Lemma 6. *If (A, w) is a balanced solution, then*

$$\text{supp}_w(A) = \min_{\emptyset \neq A' \subseteq A} \frac{1}{|A'|} \sum_{n \in \cup_{c \in A'} N_c} s_n.$$

Consequently, maximin support, the problem of maximizing the left-hand side over all balanced full solutions (A, w) , is equivalent to the problem of maximizing the right-hand side over all full committees A . Furthermore, this equivalence preserves approximations, as any balanced solution (A, w) provides the same objective value to both problems.

Network flows. In many proofs we deal with a vector $f \in \mathbb{R}^E$ of edge weights over the input graph $G = (N \cup C, E)$, which we regard as a vector of flows with positive signs considered to be flow directed toward C , and negative signs as flow directed toward N . Consequently, the *excess* of a voter $n \in N$ relative to f is $f(n) := \sum_{c \in C_n} f_{nc}$, and the excess of a candidate $c \in C$ is $f(c) := -\sum_{n \in N_c} f_{nc}$. A set of vertices $S \subseteq N \cup C$ has *net excess* if $\sum_{x \in S} f(x) > 0$, and it has *net demand* if $\sum_{x \in S} f(x) < 0$. A vector $f' \in \mathbb{R}^E$ is a *sub-flow* of f if a) for each edge $e \in E$ with $f_e \neq 0$, flows f'_e and f_e have the same sign and $|f'_e| \leq |f_e|$, and b) for each vertex $x \in N \cup C$ with $f'(x) \neq 0$, excesses $f'(x)$ and $f(x)$ have the same sign and $|f'(x)| \leq |f(x)|$. The proof of the next lemma is delayed to Appendix E.

Lemma 7. *If weight vectors $w, w' \in \mathbb{R}_{\geq 0}^E$ are non-negative and feasible for the given instance, and $f' \in \mathbb{R}^E$ is a sub-flow of $f := w' - w$, then both $w + f'$ and $w' - f'$ are non-negative and feasible as well.*

Remark 8. *In all algorithms analyzed, we assume that all numerical operations take constant time.*

3 Complexity results for maximin support

Consider an instance $(G = (N \cup C, E), s, k)$ of a multiwinner election as defined in Section 2. In this section we present an analysis of the complexity of the maximin support problem, including new results both on approximability and on hardness. As our ultimate goal is to develop efficient algorithms that provide guarantees for both the PJR property and the maximin support problem, we also comment on how the relevant heuristics in the literature of the former objective fare for the latter.

The maximin support problem was introduced in [22], where it was observed to be NP-hard. We start by showing a stronger hardness result for this problem, which rules out the existence of a PTAS.

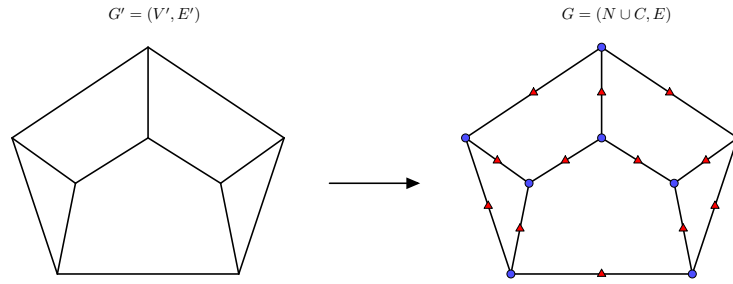


Figure 1: Reduction of an instance of the k -independent set problem on cubic graphs to an instance of the maximin support problem. Set N of voters is represented by triangles and set C of candidates by circles.

Lemma 9. *For any constant $\varepsilon > 0$, it is NP-hard to approximate the unweighted maximin support problem within a factor of $\alpha = 1.2 - \varepsilon$.*

Proof. We present a reduction from the k -independent set problem on cubic graphs, which is known to be NP-hard [9]. In this problem, one is given a graph $G' = (V', E')$ where every vertex has degree exactly 3, and a parameter k' , and one must decide whether there is a vertex subset $I \subseteq V'$ of size k' such that no two vertices in I are adjacent, i.e. I is an independent set. Given such an instance, we define an instance $(G = (N \cup C, E), s, k)$ of maximin support where $k = k'$, $C = V'$ (each vertex in V' corresponds to a candidate), and $N = E'$ with $s_n = 1$ and $C_n = n$ for each $n \in N$ (each edge in E' corresponds to a voter with unit vote that approves of the two candidates on its endpoints); see Figure 1. Notice that in this instance, each candidate is approved by exactly 3 voters, and two candidates c, c' have an approving voter in common if and only if c and c' are adjacent in V' .

Hence, if there is an independent set I of size k in G' , the same committee of validators in G can be assigned the full vote of each of its three approving voters, so that each receives a support of 3 units, which is clearly maximal. On the other hand, if there is no independent set of size k in G' , then for any solution (A, w) of the maximin support instance there must be two committee members $c, c' \in A$ who have an approving voter in common. These two members have at most five voters approving either of them, so one of them must have a support of at most $5/2$. This shows that $\text{supp}_w(A) \leq 5/2$ for any feasible solution (A, w) . Finally, notice that the ratio between the objective values 3 and $5/2$ is $6/5 = 1.2 > \alpha$, so the assumed α -approximation algorithm for maximin support would allow us to distinguish between these two cases and decide whether such an independent set I exists. This completes the proof. \square

The seqPhragmen heuristic [5] achieves the PJR property and is very fast, with a running time of $O(|E| \cdot k)$. However, in Appendix B we prove that it does not offer a constant-factor approximation for the maximin support problem.

Lemma 10. *For the maximin support problem, the approximation ratio offered by seqPhragmen is no better than the k -th harmonic number $H_k := \sum_{i=1}^k 1/i = \Theta(\log k)$.*

In contrast, we prove next that MMS [22], which also achieves the PJR property, provides a 2-factor approximation for maximin support, although with a considerably slower runtime of $O(\text{Bal} \cdot |C| \cdot k)$, where we recall that Bal is the time

complexity of computing a balanced weight vector; see Remark 5. In simple terms, MMS (Algorithm 1) starts with an empty committee A and adds to it one candidate throughout k iterations; in each iteration, it computes a balanced weight vector for each possible augmented committee that can be obtained from adding an unelected candidate, and then inserts the candidate whose corresponding augmented committee has the highest least member support.

Data: Bipartite approval graph $G = (N \cup C, E)$, vector s of vote strengths, target committee size k .

Initialize $A = \emptyset$ and $w = 0 \in \mathbb{R}_{\geq 0}^E$;

for $i = 1, 2, \dots, k$ **do**

for each $c \in C \setminus A$ **do** Compute a balanced⁶ edge weight vector w_c for $A + c$;
Find $c_i \in \arg \max_{c \in C \setminus A} \text{supp}_{w_c}(A + c)$;
Update $A \leftarrow A + c_i$ and $w \leftarrow w_{c_i}$;

end

return (A, w) ;

Algorithm 1: MMS, proposed in [22]

Theorem 11. *The MMS algorithm provides a 2-approximation for maximin support.*

We will need the following technical result, whose proof is delayed momentarily.

Lemma 12. *If (A^*, w^*) is an optimal solution to the given instance of maximin support, and (A, w) is a partial solution with $|A| \leq k$ and $A \neq A^*$, then there is a candidate $c' \in A^* \setminus A$ and a feasible solution $(A + c', w')$ such that*

$$\text{supp}_{w'}(A + c') \geq \min \left\{ \text{supp}_w(A), \frac{1}{2} \text{supp}_{w^*}(A^*) \right\}.$$

Proof of Theorem 11. Let (A_i, w_i) be the partial solution at the end of the i -th round of MMS, and let (A^*, w^*) be an optimal full solution. We prove by induction on i that $\text{supp}_{w_i}(A_i) \geq \frac{1}{2} \text{supp}_{w^*}(A^*)$, where the base case for $i = 0$ holds trivially as we use the convention that $\text{supp}_w(\emptyset) = \infty$. Assuming now that the inequality holds for i , an application of Lemma 12 for (A_i, w_i) and (A^*, w^*) implies that there is a candidate $c' \in A^* \setminus A_i$ and a feasible solution $(A_i + c', w')$ such that

$$\text{supp}_{w'}(A_i + c') \geq \min \left\{ \text{supp}_{w_i}(A_i), \frac{1}{2} \text{supp}_{w^*}(A^*) \right\} = \frac{1}{2} \text{supp}_{w^*}(A^*).$$

As the algorithm is bound to inspect candidate c' in round $i + 1$, and compute for it a balanced weight vector w_c which maximizes the support of $A_i + c'$ (by Lemma 4), the solution (A_{i+1}, w_{i+1}) at the end of round $i + 1$ must have an even higher support, i.e.

$$\text{supp}_{w_{i+1}}(A_{i+1}) \geq \text{supp}_{w_c}(A_i + c) \geq \text{supp}_{w'}(A_i + c) \geq \frac{1}{2} \text{supp}_{w^*}(A^*).$$

This completes the proof. \square

Proof of Lemma 12. Let (A, w) and (A^*, w^*) be as in the statement, let $t^* := \text{supp}_{w^*}(A^*)$, and let $t := \min\{\text{supp}_w(A), t^*/2\}$. To prove the lemma, it suffices to find a candidate $c' \in A^* \setminus A$ and a feasible weight vector $w' \in \mathbb{R}_{\geq 0}^E$ such that $\text{supp}_{w'}(A + c') \geq t$.

By decreasing some components in w and w^* , we can assume without loss of generality that $\text{supp}_w(c) = t$ for each $c \in A$, and $\text{supp}_{w^*}(c) = t^*$ for each $c \in A^*$. Define now the flow vector $f := w^* - w \in \mathbb{R}^E$. We partition the network nodes into four sets: relative to f , we have that a) N has a net excess of $|A^*| \cdot t^* - |A| \cdot t$, b) $A \setminus A^*$ has a net excess of $|A \setminus A^*| \cdot t$, c) $A^* \setminus A$ has a net demand of $|A^* \setminus A| \cdot t^*$, and d) $A \cap A^*$ has a net demand of $|A \cap A^*| \cdot (t^* - t)$. Now, using the flow decomposition theorem, we can decompose flow f into circulations and simple paths, where each path starts in a vertex with net excess and ends in a vertex with net demand. If we define f' to be the sub-flow of f that contains only the simple paths that start in N and end in $A^* \setminus A$, then

$$\begin{aligned} \text{net demand in } A^* \setminus A \text{ wrt } f' &\geq \text{net demand in } A^* \setminus A \text{ wrt } f - \text{net excess in } A \setminus A^* \text{ wrt } f \\ &= |A^* \setminus A| \cdot t^* - |A \setminus A^*| \cdot t \\ &\geq |A^* \setminus A| \cdot (t^* - t) \geq |A^* \setminus A| \cdot t, \end{aligned}$$

⁶The original algorithm in [22] does not compute balanced weight vectors, but any vector w that maximizes $\text{supp}_w(A)$, which is indeed sufficient for our analysis. However, we propose the use of balanced vectors here as they achieve further desirable properties (Lemmas 4 and 6) and because adding such requirement does not seem to cause any additional overhead in complexity.

where the last two inequalities follow from $|A^*| \geq |A|$ and $t \leq t^*/2$, respectively. By an averaging argument, this implies that there is a candidate $c' \in A^* \setminus A$ with a demand of at least t relative to f' . Finally, we define weight vector $w' := w + f'$: by Lemma 7, w' is non-negative and feasible. Furthermore, it provides the same support as w to each committee member $c \in A$, namely t , and a support of at least t to candidate c' . Hence, $\text{supp}_{w'}(A + c') \geq t$. \square

MMS is a standard greedy algorithm. To conclude the section, we mention that a "lazy" version of it can save a factor $\Theta(k)$ in the runtime while keeping the approximation guarantee virtually unchanged. In particular, in Appendix C we prove the following result.

Theorem 13. *There is an algorithm LazyMMS that, for any $\varepsilon > 0$, offers a $(2 + \varepsilon)$ -approximation for the maximin support problem, satisfies the PJR property, and executes in time $O(\text{Bal} \cdot |C| \cdot \log(1/\varepsilon))$.*

4 A new heuristic

In the previous section we established that the efficient seqPhragmen heuristic [5] fails to provide a good guarantee for the maximin support objective, whereas MMS [22] guarantees a 2-factor approximation albeit with a considerably worse running time. In this section we introduce Phragmms, a new heuristic that is inspired in seqPhragmen and maintains a comparable runtime, yet lends itself to more robust analyses both for maximin support and for PJR.

4.1 Inserting one candidate to a partial solution

We start with a brief analysis of the approaches taken in MMS and seqPhragmen. Both are iterative greedy algorithms that start with an empty committee and add to it a new candidate over k iterations, following some specific rule for candidate selection. For a given partial solution, MMS (Algorithm 1) computes a balanced weight vector for each possible augmented committee resulting from adding a candidate, and keeps the one that offers the largest support. Such a heuristic offers strong guarantees for maximin support, but is relatively slow as computing balanced vectors is costly. On the other hand, the seqPhragmen heuristic (Algorithm 5 in Appendix B) forgoes balancing and replaces it with a "lazy" version that performs minimal modifications to the current weight vector, making it balanced only in the neighborhood around the newly inserted candidate. The Phragmms heuristic takes a similar approach, but uses a slightly less lazy version of balancing, with a corresponding slight increase in runtime.

In all algorithms described in this section, we assume that there is a known background instance $(G = (N \cup C, E), s, k)$ that does not need to be passed as input. Rather, the input is a partial solution (A, w) with $|A| \leq k$. We also assume that the list of committee member supports $(\text{supp}_w(c))_{c \in A}$ is implicitly passed by reference and updated in every algorithm, so it does not need to be recomputed every time. Let $c' \in C \setminus A$ be a candidate that we consider adding to (A, w) . To do so, we need to modify weight vector w into a new feasible vector w' that redirects towards c' some of the votes of voters in $N_{c'}$, in turn decreasing the support of other committee members approved by these voters. Now, for a given threshold $t \geq 0$, we want to make sure not to reduce the support of any member c below t , assuming it starts above t , and not to reduce it at all otherwise. A simple rule to ensure this is as follows: for each voter n in $N_{c'}$ and each member $c \in A \cap C_n$, reduce the weight on edge nc from w_{nc} to $w_{nc} \cdot \min\{1, t/\text{supp}_w(c)\}$, and assign the difference to edge nc' . That way, it is clear that even if all edges incident to a member c are so reduced in weight, the support of c is scaled by a factor at most $\min\{1, t/\text{supp}_w(c)\}$ and hence its new support does not fall below t .

Thus, if for each $n \in N$ and $t \geq 0$ we define that voter's *slack* as

$$\text{slack}_{(A,w)}(n, t) := s_n - \sum_{c \in A \cap C_n} w_{nc} \cdot \min\{1, t/\text{supp}_w(c)\} \quad (4)$$

and for each $c' \in C \setminus A$ and $t \geq 0$ we define that candidate's *pre-score* as

$$\text{prescore}_{(A,w)}(c', t) := \sum_{n \in N_{c'}} \text{slack}_{(A,w)}(n, t), \quad (5)$$

then we can add c' to the solution with a support of $\text{prescore}_{(A,w)}(c', t)$, while not making any other member's support decrease below threshold t . The resulting weight modification rule is formalized in Algorithm 2. The next lemma easily follows from the previous exposition and its proof is skipped.

Lemma 14. *For a feasible partial solution (A, w) , candidate $c' \in C \setminus A$ and threshold $t \geq 0$, Algorithm Insert(A, w, c', t) executes in time $O(|E|)$ and returns a feasible solution $(A + c', w')$ where $\text{supp}_{w'}(c') = \text{prescore}_{(A,w)}(c', t)$ and $\text{supp}_{w'}(c) \geq \min\{\text{supp}_w(c), t\}$ for each member $c \in A$. In particular, if $\text{prescore}_{(A,w)}(c', t) \geq t$ then $\text{supp}_{w'}(A + c') \geq \min\{\text{supp}_w(A), t\}$.*

Data: Partial feasible solution (A, w) , candidate $c' \in C \setminus A$, threshold $t \geq 0$.
Initialize $w' \leftarrow w$;
for each voter $n \in N_{c'}$ **do**
 Set $w'_{nc'} \leftarrow s_n$;
 for each member $c \in A \cap C_n$ **do**
 if $\text{supp}_w(c) > t$ **then** update $w'_{nc} \leftarrow w'_{nc} \cdot \frac{t}{\text{supp}_w(c)}$;
 Update $w'_{nc'} \leftarrow w'_{nc'} - w'_{nc}$;
 end
end
return $(A + c', w')$;

Algorithm 2: Insert(A, w, c', t)

Whenever partial solution (A, w) is clear from context, we drop the subscript from our notation of slack and pre-score. When we add the new candidate c' to the solution, we want to ensure that inequality $\text{prescore}(c', t) \geq t$ holds, as we want to avoid increasing the number of validators with support below threshold t . Thus, for each unelected candidate $c' \in C \setminus A$ we define its *score* to be the highest value of t such that $\text{prescore}(c', t) \geq t$ holds, i.e.

$$\text{score}_{(A, w)}(c') := \max\{t \geq 0 : \text{prescore}_{(A, w)}(c', t) \geq t\}, \quad (6)$$

where again we drop the subscript if (A, w) is clear from context. Our heuristic now becomes apparent.

Heuristic (Phragmms). *Given a partial solution (A, w) , find a candidate $c_{\max} \in C \setminus A$ with highest score $t_{\max} = \max_{c' \in C \setminus A} \text{score}(c')$, and execute Insert(A, w, c_{\max}, t_{\max}) so that its output solution $(A + c_{\max}, w')$ observes*

$$\forall c \in A, \text{supp}_{w'}(c) \geq \min\{\text{supp}_w(c), t_{\max}\}, \quad \text{and} \quad \text{supp}_{w'}(A + c_{\max}) \geq \min\{\text{supp}_w(A), t_{\max}\}.$$

In Appendix D we describe efficient algorithms to find the candidate with highest pre-score for a given threshold t , as well as the candidate with overall highest score.

Theorem 15. *For a partial solution (A, w) and threshold $t \geq 0$, there is an algorithm MaxPrescore(A, w, t) that executes in time $O(|E|)$ and returns a tuple (c_t, p_t) such that $c_t \in C \setminus A$ and $p_t = \text{prescore}(c_t, t) = \max_{c' \in C \setminus A} \text{prescore}(c', t)$. Furthermore, there is an algorithm MaxScore(A, w) that runs in time $O(|E| \cdot \log k)$ and returns a tuple (c_{\max}, t_{\max}) such that $t_{\max} = \text{score}(c_{\max}) = \max_{c' \in C \setminus A} \text{score}(c')$.*

We remark that our heuristic, which finds a candidate with highest score and adds it to the current partial solution (Algorithm MaxScore followed by Insert) executes in time $O(|E| \cdot \log k)$. It thus matches up to a logarithmic term the running time of seqPhragmen which is $O(|E|)$ per iteration; see Appendix B. In Appendix D we also draw parallels between the seqPhragmen and Phragmms heuristics, and explain how the latter can be seen as a natural complication of the former which always grants higher score values to candidates and thus inserts them with higher supports.

4.2 A faster constant-factor approximation algorithm for maximin support

We proved in Section 3 that a 2-approximation algorithm for maximin support can be computed in time $O(\text{Bal} \cdot |C| \cdot k)$ or $O(\text{Bal} \cdot |C|)$ (Theorems 11 and 13 respectively). We use the Phragmms heuristic to develop a 3.15-approximation algorithm that runs in time $O(\text{Bal} \cdot k)$, and satisfies PJR as well. We highlight that this is the fastest known algorithm to achieve a constant-factor guarantee for maximin support, and that gains in speed are of paramount importance for our application to validator election in NPoS, where there are hundreds of candidates and a massive number of voters.

We propose BalPhragmms (Algorithm 3), an iterative greedy algorithm that starts with an empty committee and alternates between inserting a new candidate with the Phragmms heuristic, and *rebalancing* the weight vector, i.e. replacing it with a balanced one. This constitutes a middle ground between the approach taken in seqPhragmen where the weight vector is never rebalanced, and the approach in MMS where $O(|C|)$ balanced vectors are computed in each iteration. We formalize the procedure below. Notice that running the Insert procedure (Algorithm 2) before rebalancing is optional, but the step simplifies the analysis, and provides a good starting point to the balancing algorithm.

Theorem 16. *BalPhragmms (Algorithm 3) offers a 3.15-approximation for the maximin support problem, satisfies the PJR property, and executes in time $O(\text{Bal} \cdot k)$, assuming that $\text{Bal} = \Omega(|E| \cdot \log k)$.*

This in turn proves Theorem 1. The claim on runtime is straightforward: we established in Theorem 15 that the Phragmms heuristic runs in time $O(|E| \cdot \log k)$, so each iteration of BalPhragmms has a runtime of $O(|E| \cdot \log k +$

Data: Approval graph $G = (N \cup C, E)$, vector s of vote strengths, committee size k .
Initialize $A = \emptyset$ and $w = 0 \in \mathbb{R}_{\geq 0}^E$;

for i from 1 to k **do**

 Let $(c_{\max}, t_{\max}) \leftarrow \text{MaxScore}(A, w)$ // the candidate with highest score, and its score ;
 Update $(A, w) \leftarrow \text{Insert}(A, w, c_{\max}, t_{\max})$ // or optionally just update $A \leftarrow A + c_{\max}$;
 Replace w with a balanced weight vector for A ;

end

return (A, w) ;

Algorithm 3: BalPhragmms

$\text{Bal}) = O(\text{Bal})$, where in the last equality we assume that $\text{Bal} = \Omega(|E| \cdot \log k)$. In fact, in Appendix D we improve upon this analysis and prove that each iteration can be made to run in time $O(|E| + \text{Bal})$. Next, in order to prove PJR we need the following technical lemmas.

Lemma 17. *If (A, w) and (A', w') are two balanced partial solutions with $A \subseteq A'$, then $\text{supp}_w(c) \geq \text{supp}_{w'}(c)$ for each candidate $c \in A$. Furthermore, for each $c' \in C \setminus A'$ we have that $\text{score}_{(A, w)}(c') \geq \text{score}_{(A', w')}(c')$.*

Lemma 18. *If $\text{supp}_w(A) \geq \max_{c' \in C \setminus A} \text{score}(c')$ holds for a full solution (A, w) , then A satisfies PJR.*

Lemma 17 formalizes the intuition that as more candidates are added to committee A , supports and scores can only decrease, never increase; its proof is delayed to Appendix E. Lemma 18 provides a sense of local optimality to solution (A, w) , because if the corresponding inequality did not hold we could attempt to improve the solution by swapping the member with least support for the unelected candidate with highest score; its proof is delayed to Section 5.1 where we explore this notion of local optimality and its relation with the PJR property. We prove now that the output of BalPhragmms satisfies the condition in Lemma 18, and hence satisfies PJR.

Lemma 19. *At the end of each one of the k iterations of Algorithm BalPhragmms, if (A, w) is the current partial balanced solution, we have that $\text{supp}_w(A) \geq \max_{c' \in C \setminus A} \text{score}_{(A, w)}(c')$.*

Proof. Let (A_i, w_i) be the partial solution at the end of the i -th iteration. We prove the claim by induction on i , where the base case for $i = 0$ holds trivially as we use the convention that $\text{supp}_{w_0}(\emptyset) = \infty$ for any w_0 . For $i \geq 1$, suppose that on iteration i we insert a candidate c_i with highest score, and let w' be the weight vector obtained from running $\text{Insert}(A_{i-1}, w_{i-1}, c_i, \text{score}_{(A_{i-1}, w_{i-1})}(c_i))$ (Algorithm 2). Then

$$\begin{aligned}
\text{supp}_{w_i}(A_i) &\geq \text{supp}_{w'}(A_i) && \text{(as } w_i \text{ is balanced for } A_i) \\
&\geq \min\{\text{supp}_{w_{i-1}}(A_{i-1}), \text{score}_{(A_{i-1}, w_{i-1})}(c_i)\} && \text{(by Lemma 14)} \\
&\geq \max_{c' \in C \setminus A_{i-1}} \text{score}_{(A_{i-1}, w_{i-1})}(c') && \text{(by induction hyp. and by choice of } c_i) \\
&\geq \max_{c' \in C \setminus A_i} \text{score}_{(A_i, w_i)}(c'). && \text{(by Lemma 17)}
\end{aligned}$$

This completes the proof. \square

It remains to prove the claimed approximation guarantee of Algorithm BalPhragmms. To do that, we need the following technical result which, informally speaking, establishes that if a partial solution is balanced, then there must be a subset of voters with large slacks, and neighboring unelected candidates with high pre-scores and scores.

Lemma 20. *Let (A^*, w^*) be an optimal solution to the maximin support instance, $t^* = \text{supp}_{w^*}(A^*)$, and let (A, w) be a balanced solution with $|A| \leq k$ and $A \neq A^*$. For each $0 \leq a \leq 1$, there is a subset $N(a) \subseteq N$ of voters such that*

1. *each voter $n \in N(a)$ has a neighbor in $A^* \setminus A$;*
2. *for each voter $n \in N(a)$, we have that $\text{supp}_w(A \cap C_n) := \min_{c \in A \cap C_n} \text{supp}_w(c) \geq at^*$;*
3. *$\sum_{n \in N(a)} s_n \geq |A^* \setminus A| \cdot (1 - a)t^*$; and*
4. *for any b with $a \leq b \leq 1$ we have that $N(b) \subseteq N(a)$, and for each $n \in N(a)$ we have that n is also in $N(b)$ if and only if property 2 above holds for n with parameter a replaced by b .*

Proof. Fix a parameter $0 \leq a \leq 1$ and define the set $N' := \{n \in N : \text{supp}_w(A \cap C_n) \geq at^*\}$, where $\text{supp}_w(\emptyset) = \infty$ by convention. If we define $N(a) \subseteq N'$ as those voters in N' that have a neighbor in $A^* \setminus A$, then properties 1, 2 and 4 become evident. Hence, it only remains to prove the third property.

We claim that there is no edge with non-zero weight between $N \setminus N'$ and $A' := \{c \in A : \text{supp}_w(c) \geq at^*\}$. Indeed, if there was a pair $n \in N \setminus N'$, $c \in A'$ with $w_{nc} > 0$, then by point 3 of Lemma 4 we would have that $\text{supp}_w(A \cap C_n) = \text{supp}_w(c) \geq at^*$, contradicting the fact that $n \notin N$. Thus, we get the inequality

$$\sum_{n \in N \setminus N'} s_n \leq \sum_{c \in A \setminus A'} \text{supp}_w(c) < |A \setminus A'| \cdot at^* < |A^* \setminus A'| \cdot at^*.$$

By reducing some components in vectors w and w^* , we can assume without loss of generality that $\text{supp}_{w^*}(c) = t^*$ if $c \in A^*$, zero otherwise, and $\text{supp}_w(c) = at^*$ if $c \in A^* \cap A'$, zero otherwise. Define $f := w^* - w \in \mathbb{R}^E$, which we interpret as a vector of flows over the network induced by $N \cup A^*$, with positive signs corresponding to flow leaving N , and vice-versa. We partition the network nodes into four sets: N' , $N \setminus N'$, $A^* \cap A'$, and $A^* \setminus A'$. Relative to f , we have that a) N has a net excess of $|A^*| \cdot t^* - |A^* \cap A'| \cdot at^*$, b) $N \setminus N'$ has a net excess of at most $\sum_{n \in N \setminus N'} s_n < |A^* \setminus A'| \cdot at^*$ (by the previous inequality), and c) $A^* \cap A'$ has a net demand of $|A^* \cap A'| \cdot (1-a)t^*$.

Using the flow decomposition theorem, we can decompose flow f into circulations and simple paths. If we define f' to be the sub-flow of f that contains only the simple paths that start in N' and end in $A^* \setminus A'$, then

$$\begin{aligned} \text{flow value in } f' &\geq (\text{net excess in } N - \text{net excess in } N \setminus N' - \text{net demand in } A^* \cap A') \text{ w.r.t. } f \\ &\geq |A^*| \cdot t^* - |A^* \cap A'| \cdot at^* - |A^* \setminus A'| \cdot at^* - |A^* \cap A'| \cdot (1-a)t^* \\ &= |A^* \setminus A'| \cdot (1-a)t^* \geq |A^* \setminus A| \cdot (1-a)t^*. \end{aligned}$$

A key observation now is that none of the flow in f' can pass by any node in $N \setminus N'$ or $A^* \cap A \setminus A'$; see Figure 2. This is because any path in f' starts in N' , nodes in N' have no neighbors in $A \setminus A'$ (by definitions of N' and A'), and furthermore there is no flow possible from $(A^* \setminus A) \cup (A^* \cap A')$ to $N \setminus N'$ in $f = w^* - w$, because w has no flow from $N \setminus N'$ toward A' (by our claim) nor toward $A^* \setminus A$ (by our assumption wlog on w). Therefore, the formula above is actually a lower bound on the flow going from N' to $A^* \setminus A$. Finally, we notice that for each path in f' , the last edge goes from N' to $A^* \setminus A$, so it originates in $N(a)$. This proves that $\sum_{n \in N(a)} s_n \geq \text{flow value in } f' \geq |A^* \setminus A| \cdot (1-a)t^*$, which is the third property. \square

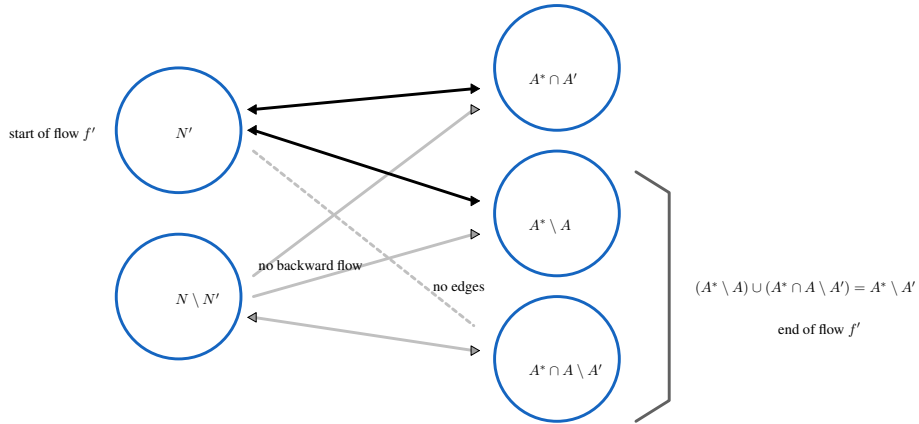


Figure 2: Flow f' can only visit nodes in N' , $A^* \cap A'$ and $A^* \setminus A$.

As a warm-up, we show how this last result easily implies a 4-approximation guarantee for BalPhragmms.

Lemma 21. *If (A, w) , (A^*, w^*) and t^* are as in Lemma 20, there is a candidate $c' \in A^* \setminus A$ with $\text{score}(c') \geq t^*/4$. Therefore, BalPhragmms provides a 4-approximation for the maximin support problem.*

Proof. Apply Lemma 20 with $a = 1/2$. In what follows we refer to the properties stated in that lemma. We have that

$$\begin{aligned}
\sum_{c' \in A^* \setminus A} \text{prescore}(c', t^*/4) &= \sum_{c' \in A^* \setminus A} \sum_{n \in N_{c'}} \text{slack}(n, t^*/4) \geq \sum_{n \in N(a)} \text{slack}(n, t^*/4) && \text{(by property 1)} \\
&\geq \sum_{n \in N(a)} \left[s_n - \frac{t^*}{4} \sum_{c \in A \cap C_n} \frac{w_{nc}}{\text{supp}_w(c)} \right] && \text{(by equation 4)} \\
&\geq \sum_{n \in N(a)} \left[s_n - \frac{1}{2} \sum_{c \in A \cap C_n} w_{nc} \right] && \text{(by property 2)} \\
&\geq \frac{1}{2} \sum_{n \in N(a)} s_n \geq \frac{1}{2} (|A^* \setminus A| \cdot t^*/2) = |A^* \setminus A| \cdot t^*/4. && \text{(by ineq. 2 and property 3)}
\end{aligned}$$

Therefore, by an averaging argument, there must be a candidate $c' \in A^* \setminus A$ with $\text{prescore}(c', t^*/4) \geq t^*/4$, which implies that $\text{score}(c') \geq t^*/4$, by definition of score. The 4-approximation guarantee for Algorithm BalPhragmms easily follows by induction on the k iterations, using Lemma 14 and the fact that rebalancing a partial solution never decreases its least member support. \square

To get a better approximation guarantee for Algorithm BalPhragmms and finish the proof of Theorem 16, we apply Lemma 20 with a different parameter a . For this, we will use the following result whose proof is delayed to Appendix E.

Lemma 22. *Consider a strictly increasing and differentiable function $f : \mathbb{R} \rightarrow \mathbb{R}$, with a unique root χ . For a finite sum $\sum_{i \in I} \alpha_i f(x_i)$ where $\alpha_i \in \mathbb{R}$ and $x_i \geq \chi$ for each $i \in I$, we have that*

$$\sum_{i \in I} \alpha_i f(x_i) = \int_{\chi}^{\infty} f'(x) \left(\sum_{i \in I: x_i \geq x} \alpha_i \right) dx.$$

Lemma 23. *If (A, w) , (A^*, w^*) and t^* are as in Lemma 20, there is a candidate $c' \in A^* \setminus A$ with $\text{score}(c') \geq t^*/3.15$. Therefore, BalPhragmms provides a 3.15-approximation for the maximin support problem.*

Proof. Again we apply Lemma 20 and its properties, for a parameter $0 \leq a \leq 1$ to be defined later. We have

$$\begin{aligned}
\sum_{c' \in A^* \setminus A} \text{prescore}(c', at^*) &= \sum_{c' \in A^* \setminus A} \sum_{n \in N_{c'}} \text{slack}(n, at^*) \geq \sum_{n \in N(a)} \text{slack}_w(n, at^*) && \text{(by property 1)} \\
&\geq \sum_{n \in N(a)} \left[s_n - at^* \sum_{c \in A \cap C_n} \frac{w_{nc}}{\text{supp}_w(c)} \right] && \text{(by equation 4)} \\
&\geq \sum_{n \in N(a)} \left[s_n - \frac{at^*}{\text{supp}_w(A \cap C_n)} \sum_{c \in A \cap C_n} w_{nc} \right] && \text{(by property 2)} \\
&\geq \sum_{n \in N(a)} s_n \left[1 - \frac{at^*}{\text{supp}_w(A \cap C_n)} \right], && \text{(by ineq. 2)}
\end{aligned}$$

where $\text{supp}_w(\emptyset) = \infty$ by convention. At this point, we apply Lemma 22 over function $f(x) := 1 - a/x$, which has the unique root $\chi = a$, and index set $I = N(a)$ with $\alpha_n = s_n$ and $x_n = \text{supp}_w(A \cap C_n)/t^*$. We obtain

$$\begin{aligned}
\sum_{c' \in A^* \setminus A} \text{prescore}(c', at^*) &\geq \int_a^{\infty} f'(x) \left(\sum_{n \in N(a): \text{supp}_w(A \cap C_n) \geq xt^*} s_n \right) dx \\
&= \int_a^{\infty} \frac{a}{x^2} \left(\sum_{n \in N(x)} s_n \right) dx && \text{(by property 4)} \\
&\geq \int_a^1 \frac{a}{x^2} \left(|A^* \setminus A| \cdot (1-x)t^* \right) dx && \text{(by property 3)} \\
&= |A^* \setminus A| \cdot at^* \int_a^1 \left(\frac{1}{x^2} - \frac{1}{x} \right) dx \\
&= |A^* \setminus A| \cdot at^* \left(\frac{1}{a} - 1 + \ln a \right).
\end{aligned}$$

If we now set $a = 1/3.15$, we have that $1/a - 1 + \ln a \geq 1$, and thus by an averaging argument there is a candidate $c' \in A^* \setminus A$ for which $\text{prescore}(c', at^*) \geq at^*$, and hence $\text{score}(c') \geq at^*$. The approximation guarantee for Algorithm BalPhragmms follows by induction on the k iterations, as before. \square

5 Exploiting local optimality

We start the section with an key property of algorithm BalPhragmms as motivation.

Theorem 24. *If (A, w) is a balanced solution such that $\text{supp}_w(A) \geq \max_{c' \in C \setminus A} \text{score}(c')$, then it satisfies PJR and guarantees a 3.15-factor approximation for the maximin support problem. Furthermore, testing the required conditions (feasibility, balancedness and the previous inequality) can be done in time $O(|E|)$. Finally, the output solution of the BalPhragmms algorithm is guaranteed to satisfy these conditions.*

Proof. The first statement follows from Lemmas 18 and 23, and the third one from Lemma 19. Feasibility (inequality 2) can clearly be checked in time $O(|E|)$, as can balancedness by Lemma 4, because properties 2 and 3 in that lemma can both be tested in this time. Finally, if $t := \text{supp}_w(A)$, the inequality in the statement is equivalent to $t \geq \max_{c' \in C \setminus A} \text{prescore}(c', t)$, which is tested with algorithm MaxPrescore(A, w, t) in time $O(|E|)$ by Theorem 15. \square

This in turn proves Theorem 2. A result such as the one above is essential in a scenario where a computationally limited user (the *verifier*) offloads a heavy task – in this case an election algorithm – to one or more external entities with more computational power (the *prover*). Yet, as the task is executed privately and the user does not trust the entities, the user must be in condition to provably and efficiently check the quality of the output. In the case of an election protocol over a decentralized blockchain, it would be very costly to run the full election algorithm as an *on-chain* process, meaning that validators must execute it simultaneously and the chain cannot progress until they have finished the execution and agreed on the output. Instead, a much more scalable solution is to execute the protocol *off-chain*, meaning that one or more parties run it privately and separate from block production, and then submit the output which is approved if it passes the verification test on-chain. The fact that algorithm BalPhragmms can have both of its guarantees efficiently verified on its outputs is in our opinion its most relevant feature.

The inequality $\text{supp}_w(A) \geq \max_{c' \in C \setminus A} \text{score}(c')$ mentioned in Lemma 18 and Theorem 24 corresponds to a notion of local optimality, for a local search variant of the Phragmms heuristic. In Section 5.1 we explore the relation between this notion and the PJR property. Then, in Section 5.2 we use the corresponding local search algorithm to devise a post-computation which takes an arbitrary solution (A, w) as input, and returns an output (A', w') which observes $\text{supp}_{w'}(A') \geq \text{supp}_w(A)$ and provably satisfies PJR.

5.1 A parametric version of proportional justified representation

We define next a parametric version of the PJR property that measures just how well represented the voters are by a given committee A . It is a generalization of the property which turns it from binary to quantitative.

Definition 25. *For any $t \in \mathbb{R}_{\geq 0}$, a committee $A \subseteq C$ (of any size) satisfies Proportional Justified Representation with parameter t (t -PJR for short) if there is no group $N' \subseteq N$ of voters and integer $0 < r \leq |A|$ such that:*

- a) $\sum_{n \in N'} s_n \geq r \cdot t$,
- b) $|\cap_{n \in N'} C_n| \geq r$, and
- c) $|A \cap (\cup_{n \in N'} C_n)| < r$.

In words, if there is a group N' of voters with at least r commonly approved candidates, who can afford to provide these candidates with a support of value t each, then this group is represented by at least r members in committee A , though not necessarily commonly approved. Notice that the standard version of PJR is equivalent to \hat{t} -PJR for $\hat{t} := \sum_{n \in N} s_n / |A|$ (see Section 2), and that if a committee satisfies t -PJR then it also satisfies t' -PJR for each $t' \geq t$, i.e. the property gets stronger as t decreases. Notice as well that this is in contrast to the maximin support objective, which implies a stronger property as it increases; we will exploit the opposite natures of these two parameters of quality in Section 5.2 to define a procedure that improves upon a given solution.

We remark that in [21] the related notion of *average satisfaction* is introduced, also to quantify the level of proportional representation achieved by a committee. Roughly speaking, that notion measures the average number of representatives in the committee that each voter in a group has, for any group of voters with sufficiently high vote strength and

cohesiveness level. In contrast, with parametric PJR we focus on providing sufficient representatives to the group as a whole and not to each individual voter, and we measure the required vote strength for a group to achieve it. Interestingly, the average satisfaction measure is closely linked to the property of extended justified representation (EJR) [2], and in [3] that measure is used to prove that a local search algorithm achieves EJR. Similarly, we use parametric PJR to prove that a local search algorithm achieves standard PJR.

Testing whether an arbitrary solution satisfies standard PJR is known to be coNP-complete [3], hence the same remains true for its parametric version. We provide next a sufficient condition for a committee to satisfy t -PJR, which is efficiently testable, based on our definitions of pre-score and score.

Lemma 26. *If for a feasible solution (A, w) there is a parameter $t \in \mathbb{R}_{\geq 0}$ such that $\max_{c' \in C \setminus A} \text{prescore}(c', t) < t$, or equivalently, such that $\max_{c' \in C \setminus A} \text{score}(c') < t$, then committee A satisfies t -PJR.*

Proof. We prove the contrapositive of the claim. If A does not satisfy t -PJR, there must be a subset $N' \subseteq N$ of voters and an integer $r > 0$ that observe points a) through c) in the definition above. By points b) and c), set $(\cap_{n \in N'} C_n) \setminus A$ must be non-empty: let c' be a candidate in it. We will prove that for any feasible weight vector $w \in \mathbb{R}_{\geq 0}^E$, it holds that $\text{prescore}(c', t) \geq t$, and consequently $\text{score}(c') \geq t$ by the definition of score. We have

$$\begin{aligned}
 \text{prescore}(c', t) &= \sum_{n \in N_{c'}} \text{slack}(n, t) \geq \sum_{n \in N'} \text{slack}(n, t) && (\text{as } N' \subseteq N_{c'}) \\
 &\geq \sum_{n \in N'} \left(s_n - t \cdot \sum_{c \in A \cap C_n} \frac{w_{nc}}{\text{supp}_w(c)} \right) && (\text{by definition 4}) \\
 &= \sum_{n \in N'} s_n - t \cdot \sum_{c \in A \cap (\cup_{n \in N'} C_n)} \frac{\sum_{n \in N' \cap N_c} w_{nc}}{\sum_{n \in N_c} w_{nc}} && (\text{where fraction is } \leq 1) \\
 &\geq t \cdot r - t \cdot |A \cap (\cup_{n \in N'} C_n)| && (\text{by a)}) \\
 &\geq t \cdot r - t \cdot (r - 1) = t && (\text{by c)}).
 \end{aligned}$$

This proves that $\text{prescore}(c', t) \geq t$, which is what we needed to show. \square

For a given solution (A, w) and parameter t , one can verify the condition above in time $O(|E|)$ by checking whether $\text{MaxPrescore}(A, w, t) < t$; see Theorem 15. Alternatively, from just solution (A, w) one can execute $\text{MaxScore}(A, w)$ in time $O(|E| \cdot \log k)$ to obtain the highest score t_{\max} and ascertain that A satisfies t -PJR for each $t > t_{\max}$. The proof of Lemma 18 now follows as a corollary.

Proof of Lemma 18. Let $t_{\max} := \max_{c' \in C \setminus A} \text{score}(c')$ and let $c_{\max} \in C \setminus A$ be a candidate with such highest score. If $\text{supp}_w(A) \geq t_{\max}$, it follows from Lemma 14 that if we execute $\text{Insert}(A, w, c_{\max}, t_{\max})$, we obtain a solution $(A + c_{\max}, w')$ with $\text{supp}_{w'}(A + c_{\max}) = t_{\max}$. Now, by feasibility of vector w' , we have the inequality $\sum_{n \in N} s_n \geq \sum_{c \in A + c_{\max}} \text{supp}_{w'}(c) \geq (k + 1) \cdot t_{\max}$, and thus $t_{\max} \leq \sum_{n \in N} s_n / (k + 1) < \sum_{n \in N} s_n / k = \hat{t}$. By Lemma 26 above, having $\max_{c' \in C \setminus A} \text{score}(c') < \hat{t}$ implies that A satisfies \hat{t} -PJR, which is standard PJR. \square

5.2 A local search algorithm

Suppose that we know of an efficient algorithm with good guarantees for maximin support but no guarantee for PJR, or we happen to know of a high quality solution in terms of maximin support but we ignore if it satisfies PJR. Can we use it to find a new solution of no lesser quality which is also guaranteed to satisfy PJR? And can we efficiently prove to a verifier that the new solution indeed satisfies PJR? We answer these questions in the positive for the first time.

We present a local search algorithm that takes an arbitrary feasible solution as input, and iteratively drops a member of least support and inserts a new candidate using the Phragmms heuristic. The procedure always maintains or increases the value of the least member support, hence the quality of the solution is preserved. Furthermore, as the solution converges to a local optimum, it is guaranteed to satisfy the PJR after a certain number of iterations. Therefore, when used as a post-computation, this procedure can make any approximation algorithm for maximin support also satisfy PJR in a black-box manner. We remark here that such an application of the Phragmms heuristic goes to show the robustness of the election rule; in particular, there is no evident way to build a similar post-computation from seqPhragmen [5], as the analysis of the PJR property in that rule is heavily dependent on the way the solution is constructed.

As we did in Section 4.1, in the following algorithm we assume that the background instance $(G = (N \cup C, E), s, k)$ is known and that does not need to be passed as input. Instead, the input is a feasible full solution (A, w) , and a parameter $\varepsilon > 0$. Our proposed algorithm PostPhragmms (Algorithm 4) is presented below.

Data: Full feasible solution (A, w) , parameter $\varepsilon > 0$.

Let $\hat{t} \leftarrow \sum_{n \in N} s_n / |A|$;

while True **do**

Find tuple (c_{\min}, t_{\min}) so that $c_{\min} \in A$ and $t_{\min} = \text{supp}_w(c_{\min}) = \text{supp}_w(A)$;
 Let $(c_{\max}, t_{\max}) \leftarrow \text{MaxScore}(A, w)$ // the candidate with highest score, and its score;
if $(t_{\max} < \min\{(1 + \varepsilon) \cdot t_{\min}, \hat{t}\})$ **then return** (A, w) ;
 Update $(A, w) \leftarrow \text{Insert}(A - c_{\min}, w, c_{\max}, t_{\max})$;

end

Algorithm 4: PostPhragmms(A, w, ε)

Theorem 27. For any parameter $\varepsilon > 0$ and a feasible full solution (A, w) , let (A', w') be the output solution to PostPhragmms(A, w, ε). Then:

1. solution (A', w') is feasible and full, and $\text{supp}_{w'}(A') \geq \text{supp}_w(A)$;
2. solution (A', w') satisfies the condition on Lemma 26 for parameter $t = \min\{(1 + \varepsilon) \cdot \text{supp}_{w'}(A'), \hat{t}\}$, where $\hat{t} = \sum_{n \in N} s_n / k$, so A' satisfies t -PJR and standard PJR;
3. if (A, w) has an α -approximation guarantee for the maximin support objective, for some $\alpha \geq 1$, then the algorithm performs at most $k \cdot (1 + \log_{1+\varepsilon} \alpha) + 1$ iterations, each taking time $O(|E| \cdot \log k)$; and
4. by setting $\varepsilon \rightarrow \infty$, the algorithm finds a solution satisfying standard PJR in at most $k + 1$ iterations.

This proves Theorem 3. Notice that point 3 establishes that PostPhragmms can be executed as a post-computation of any constant-factor approximation algorithm for maximin support in time $O(|E| \cdot k \log k)$. In particular, this complexity is dominated by that of all constant-factor approximations presented in this paper. By point 4, the algorithm can be sped up if we only care about standard PJR, or run further iterations to achieve a stronger parametric PJR guarantee. In the proof, we use the observation below whose proof is skipped as it follows from the definitions of slack and pre-score.

Lemma 28. For a feasible vector $w \in \mathbb{R}_{\geq 0}^E$, two committees $A \subseteq A'$, and any threshold $t \geq 0$ it holds that $\text{prescore}_{(A, w)}(c', t) \geq \text{prescore}_{(A', w)}(c', t)$ for each candidate $c' \in C \setminus A'$.

Proof of Theorem 27. We start with some notation. We use i as a superscript to indicate the value of the different variables at the beginning of the i -th iteration. We define $t_{\text{stop}}^i := \min\{(1 + \varepsilon) \cdot t_{\min}^i, \hat{t}\}$, so that the stopping condition reads $t_{\max}^i \stackrel{?}{<} t_{\text{stop}}^i$. Notice that $t_{\text{stop}}^i \geq t_{\min}^i$ always holds by feasibility of w^i and definition of \hat{t} .

We prove point 1 by induction on i . If the stopping condition does not hold then $t_{\max}^i \geq t_{\text{stop}}^i \geq t_{\min}^i$. Next, by Lemma 28 we have $\text{prescore}_{(A^i - c_{\min}^i, w^i)}(c_{\max}^i, t_{\max}^i) \geq \text{prescore}_{(A^i, w^i)}(c_{\max}^i, t_{\max}^i) = t_{\max}^i \geq t_{\min}^i$. On the other hand, it is evident that $\text{supp}_{w^i}(A^i - c_{\min}^i) \geq \text{supp}_{w^i}(A^i) = t_{\min}^i$. Therefore, by Lemma 14 we have that

$$\text{supp}_{w^{i+1}}(A^{i+1}) \geq \min\{\text{supp}_{w^i}(A^i - c_{\min}^i), \text{prescore}_{(A^i - c_{\min}^i, w^i)}(c_{\max}^i, t_{\max}^i)\} \geq t_{\min}^i,$$

and that (A^{i+1}, w^{i+1}) is feasible. This proves point 1. Point 2 is apparent from the stopping condition.

We continue to point 3. Each iteration is dominated in complexity by the call to Algorithm MaxScore, which takes time $O(|E| \cdot \log k)$ by Theorem 15. Hence, it remains to give a bound on the number T of total iterations. To do that, we analyze the evolution of the least member support $t_{\min}^i = \text{supp}_{w^i}(A^i)$. First, we argue that if ever $t_{\min}^i = \hat{t}$, then the algorithm terminates immediately, i.e. $i = T$; this is because in this case all members in A^i must have a support of exactly \hat{t} , all voters a zero slack for threshold \hat{t} , and all candidates in $C \setminus A^i$ a zero prescore for threshold \hat{t} and hence a score strictly below \hat{t} , and the stopping condition is fulfilled. Next, we claim that for any iteration i with $1 \leq i < T - k$, we have $t_{\min}^{i+k} \geq (1 + \varepsilon) \cdot t_{\min}^i$. This is because, by Lemma 14, in each iteration $j \geq i$ we are removing a member with least support while not increasing the number of members with support below

$$\text{prescore}_{(A^j - c_{\min}^j, w^j)}(c_{\max}^j, t_{\max}^j) \geq t_{\max}^j \geq t_{\text{stop}}^j \geq t_{\text{stop}}^i.$$

As there are only k candidates, we must have that $t_{\min}^{i+k} \geq t_{\text{stop}}^i = \min\{(1 + \varepsilon) \cdot t_{\min}^i, \hat{t}\}$. Thus t_{\min}^{i+k} is either at least $(1 + \varepsilon) \cdot t_{\min}^i$, or it is \hat{t} , but it cannot be the latter by the previous claim and the fact that $i + k < T$. This proves the new claim. Therefore, if the algorithm outputs a solution (A^T, w^T) and has an input solution (A^1, w^1) with an α -approximation guarantee for the maximin support objective, then

$$\alpha \cdot \text{supp}_{w^1}(A^1) \geq \text{supp}_{w^T}(A^T) \geq (1 + \varepsilon)^{\lfloor (T-1)/k \rfloor} \cdot \text{supp}_{w^1}(A^1).$$

Hence, $\alpha \geq (1 + \varepsilon)^{\lfloor (T-1)/k \rfloor}$, and $T \leq k \cdot (1 + \log_{1+\varepsilon} \alpha) + 1$. This completes the proof of point 3.

For point 4, we consider again the previous inequality $t_{\min}^{i+k} \geq t_{\text{stop}}^i = \min\{(1 + \varepsilon) \cdot t_{\min}^i, \hat{t}\}$. By setting $i = 1$ and $\varepsilon \rightarrow \infty$, we have that either the stopping condition is triggered before the $(k + 1)$ -st iteration, or it must be the case that $t_{\min}^{k+1} \geq \hat{t}$, and the algorithm terminates immediately as argued previously. In either case, the claim on standard PJR follows from point 2. \square

6 A protocol for validator election in NPoS

In this section we provide high-level details of a proposal for a validator election protocol in an NPoS-based blockchain network. This proposal will be the basis for an implementation in the Polkadot network.

A committee of k validators is selected once per era, where an era is multiple hours or days long. In each era, a group of off-chain workers each privately runs the election algorithm for the next era, and submits its solution on-chain. We propose that each validator act as an off-chain worker and run such computation, but this need not be the case. More in detail, towards the end of each era there is an *election window* where the following events occur:

1. The chain fixes the (otherwise ever-evolving) set of validator candidates and nominators' preferences and stake to be considered for the election, ignoring any further changes in the remainder of the era.
2. All current validators trigger an off-chain execution of the BalPhragmms election rule, separate from block production and other duties.
3. Once the solutions are computed (after a few seconds of the start of the election window), validators submit them on-chain as a special type of transaction.
4. On the on-chain side, we only keep track of one solution at any given time, which is the current tentative winner. Specifically, if (A_t, w_t) is the tentative winner recorded on-chain, a block producer can include a new solution (A, w) to its block only if a) it is feasible, b) $\text{supp}_w(A) > \text{supp}_{w_t}(A_t)$, and c) it passes the test described in Theorem 24. If this is the case, (A, w) replaces (A_t, w_t) as the current tentative winner.
5. At the end of the election window, the current tentative winner is declared the official winner.

We make a few remarks about this protocol.

- By Theorem 24, the protocol will elect a solution that simultaneously satisfies the PJR property and a 3.15-factor approximation for the maximin support problem. These constitute strong and formal guarantees on security and proportionality.
- On top of guarantees above, we are bound to elect the best solution found by anyone. In particular, validators may find diverse solutions due to different tie-breaking rules, or more explicit deviations from the suggested algorithm. However, as long as we rank solutions objectively and pick the best one, any diversity in solutions can only improve the quality of the winning committee, and thus benefits the community.
- The fact that the guarantees on security and proportionality are verifiable protect the network against a *long range attack*, i.e. a scenario where an adversary creates a branch of the blockchain which is grown in secret, with the intention to eventually make it public and have it overtake the main chain. In particular, if the protocol limited itself to select the best solution submitted by validators without a verification of guarantees, an adversary could use a long range attack to create a branch in which, during an election window, it censors all solutions coming from honest validators and thus makes the protocol elect a committee where the adversary is heavily overrepresented.
- On the on-chain side, the election protocol runs in linear time $O(|E|)$ per block, by Theorem 24 and the fact that at most one solution is verified per block.

Next, we suggest additional variations and optimizations:

- We suggest that instead of gossiping a solution over the network as a transaction, each validator waits for its turn to be a block producer to submit its solution. Doing so saves communication overhead, as election solutions make for rather heavy transactions. In this case, it is important that the election window is long enough so that with high probability each validator gets to produce a block, and thus is not censored. The system can still receive solutions as transactions from arbitrary users, but in this case the transaction fee must be high to protect against spamming attacks.
- For any weight vector $w \in \mathbb{R}_{\geq 0}^E$, let E_w denote its edge support, i.e. $E_w := \{e \in E : w_e > 0\}$, and notice that the size of a solution (A, w) is dominated by the size of its edge support $|E_w|$. However, by removing circulations, from any feasible vector w one can compute another feasible vector $w' \in \mathbb{R}_{\geq 0}^E$ such that a) all supports are preserved, i.e. $\text{supp}_w(c) = \text{supp}_{w'}(c)$ for each $c \in C$, and b) $E_{w'}$ is a forest, and hence $|E_{w'}| < |N| + k = O(|N|)$, where we assume that $|N| \geq k$. Therefore, validators can run this post-computation off-chain to reduce the size of any solution to $O(|N|)$, thus saving space on the block.
- Recall that on the on-chain side, for a new solution (A, w) we check a) feasibility, b) its objective value (and whether it beats the current tentative winner), and c) the test in Theorem 24. Notice that the third check runs in time $O(|E|)$ and thus is considerably slower than the first two checks, which run in time $O(|E_w|)$, with $O(|E_w|) = O(|N|)$ if the previous suggestion is implemented. A possible optimization is as follows: the chain requires the first submitted solution (A_0, w_0) to pass all three checks, and then skips the third check on all subsequent solutions, so that their processing time drops to $O(|N|)$. At the end of the election window, we check whether the current tentative winner (A_t, w_t) satisfies PJR with the condition on Lemma 26. We can expect that this check is always passed. In the unlikely case that (A_t, w_t) fails this condition, the election window is extended for a few more minutes and validators are asked to run the post-computation PostPhragmms over it off-chain. The window ends as soon as some block producer submits a feasible solution (A_T, w_T) such that $\text{supp}_{w_T}(A_T) \geq \text{supp}_{w_t}(A_t)$ and which satisfies PJR as attested by the condition on Lemma 26. With this optimization, the runtime drops to $O(|N|)$ per block, except for two or possibly three blocks with a runtime of $O(|E|)$, and we keep all guarantees, namely that the winning committee a) is at least as good as all submitted solutions, b) satisfies PJR, and c) is a 3.15-factor approximation for maximin support (as its objective value is at least as good as that of the first solution (A_0, w_0) , which passed the test in Theorem 24).
- In terms of incentives, the system can provide an economic reward to the submitter of each tentative winning solution, with larger rewards for the first and the last one to encourage early submissions and good submissions respectively. If the previous suggestion is implemented and a solution (A_t, w_t) fails the test on Lemma 26, a heavy fine should be charged to the submitter.

A Algorithms to compute a balanced solution

Recall from Section 2 that, for a given election instance $(G = (N \cup C, E), s, k)$ and a fixed committee $A \subseteq C$, a weight vector $w \in \mathbb{R}_{\geq 0}^E$ is balanced for A if a) it maximizes the sum of member supports, $\sum_{c \in A} \text{supp}_w(c)$, over all feasible weight vectors, and b) it minimizes the sum of supports squared, $\sum_{c \in A} (\text{supp}_w(c))^2$, over all vectors that observe the previous point.

We start by noticing that a balanced weight vector can be computed with numerical methods for quadratic convex programs. Let $E_A \subseteq E$ be the restriction of the input edge set over edges incident to committee A , and let $D \in \{0, 1\}^{A \times E_A}$ be the vertex-edge incidence matrix for A . For any weight vertex $w \in \mathbb{R}_{\geq 0}^{E_A}$, the support that w assigns to candidates in A is given by vector Dw , so that $\text{supp}_w(c) = (Dw)_c$ for each $c \in A$. We can now write the problem of finding a balanced weight vector as a convex program:

$$\begin{aligned}
& \text{Minimize} && \|Dw\|^2 \\
& \text{Subject to} && w \in \mathbb{R}_{\geq 0}^{E_A}, \\
& && \sum_{c \in C_n} w_{nc} \leq s_n \quad \text{for each } n \in N, \text{ and} \\
& && \mathbb{1}^\top Dw = \sum_{n \in \cup_{c \in A} N_c} s_n,
\end{aligned}$$

where the first two lines of constraints corresponds to non-negativity and feasibility conditions (see inequality 2), and the last line ensures that the sum of supports is maximized, where $\mathbb{1} \in \mathbb{R}^A$ is the all-ones vector.

However, there is a more efficient method using techniques for parametric flow, which we sketch now. Hochbaum and Hong [11, Section 6] consider a network resource allocation problem which generalizes the problem of finding

a balanced weight vector: given a network with a single source, single sink and edge capacities, minimize the sum of squared flows over the edges reaching the sink, over all maximum flows. They show that this is equivalent to a parametric flow problem called *lexicographically optimal flow*, studied by Gallo, Gregoriadis and Tarjan [8]. In turn, in this last paper the authors show that, even though a parametric flow problem usually requires solving several consecutive max-flow instances, this particular problem can be solved running a single execution of the FIFO preflow-push algorithm proposed by Goldberg and Tarjan [10].

Therefore, the complexity of finding a balanced weight vector is bounded by that of Goldberg and Tarjan's algorithm, which is $O(n^3)$ for a general n -node network. However, Ahuja et al. [1] showed how to optimize several popular network flow algorithms for the case of bipartite networks, where one of the partitions is considerably smaller than the other. Assuming the sizes of the bipartition are n_1 and n_2 with $n_1 \ll n_2$, they implement a two-edge push rule that allows one to "charge" most of the computation weight to the nodes on the small partition, and hence obtain algorithms whose running times depend on n_1 rather than n . In particular, they show how to adapt Goldberg and Tarjan's algorithm to run in time $O(e \cdot n_1 + n_1^3)$, where e is the number of edges. For our particular problem, which can be defined on a bipartite graph $(N \cup A, E_A)$ where $|A| \leq k \ll |N|$, we obtain thus an algorithm that runs in time $O(|E_A| \cdot k + k^3)$.

B A negative example for the sequential Phragmén heuristic

In this section we analyze the performance of seqPhragmen [5] with respect to the maximin support objective.

Data: Bipartite approval graph $G = (N \cup C, E)$, vector s of vote strengths, target committee size k .

Initialize $A = \emptyset$, $w = 0 \in \mathbb{R}_{\geq 0}^E$, $load(n) = 0$ for each $n \in N$, and $load(c') = 0$ for each $c' \in C$;

for $i = 1, 2, \dots, k$ **do**

for each candidate $c' \in C \setminus A$ **do** update $load(c') \leftarrow \frac{1 + \sum_{n \in N_{c'}} s_n \cdot load(n)}{\sum_{n \in N_{c'}} s_n}$;

Find $c_{\min} \in \arg \min_{c' \in C \setminus A} load(c')$;

Update $A \leftarrow A + c_{\min}$;

for each voter $n \in N_{c_{\min}}$ **do**

Update $w_{nc_{\min}} \leftarrow load(c_{\min}) - load(n)$;

Update $load(n) \leftarrow load(c_{\min})$;

end

end

Scaling edge weights: for each $nc \in E$ with $w_{nc} > 0$, update $w_{nc} \leftarrow w_{nc} \cdot \frac{s_n}{load(n)}$;

return (A, w) ;

Algorithm 5: seqPhragmen, proposed in [5]

The method proposed in [5] only considers unit votes, and outputs only a committee. In Algorithm 5 we present a generalization that admits weighted votes in the input, and outputs a feasible edge weight vector along with the committee, for ease of comparison with other algorithms in this paper. We remark however that the numerical example we present below can be easily converted into another with unit votes – where each voter n is replaced with a cluster of voters with unit vote strength and where the cluster size is proportional to s_n – that would lead to the same negative result for the method in [5], even when the output committee is coupled with a balanced weight vector.

We exhibit now a family of instances where the approximation ratio offered by the output of Algorithm 5 equals the k -th harmonic number $H_k := \sum_{i=1}^k 1/i = \Theta(\log k)$. This proves Lemma 10.

Proof of Lemma 10. For an arbitrarily small constant $\varepsilon > 0$ and a committee size k , consider an instance where $N = \{n_0, \dots, n_k\}$, $C = \{c_0, \dots, c_k\}$, $s_{n_j} = 1$ and $C_{n_j} = \{c_i : 1 \leq i \leq j\}$ for each $1 \leq j \leq k$, and $s_0 = 1/(H_k - \varepsilon)$ with $C_{n_0} = \{c_0\}$; see Figure 3. It is easy to see that if committee $\{c_1, \dots, c_k\}$ is selected, each member can be given a support of value 1. On the other hand, any committee that selects candidate c_0 can only provide it a support of value s_0 . We will prove that Algorithm 5 selects committee $\{c_0, c_1, \dots, c_{k-1}\}$, and thus its approximation ratio is $1/s_0 = H_k - \varepsilon$.

In the first round, we have that $load(c_0) = 1/s_0$ and $load(c_j) = \frac{1}{k+1-j}$ for each $j \geq 1$, so we add c_1 to the committee, which has $load(c_1) = 1/k = H_k - H_{k-1}$, and update $load(n_j) \leftarrow H_k - H_{k-1}$ for each $j \geq 1$. More generally, in the i -th round for $1 \leq i \leq k-1$, we have that $load(c_0) = s_0$ and $load(c_j) = \frac{1+(k+1-j)(H_k-H_{k+1-i})}{k+1-j} = \frac{1}{k+1-j} + H_k - H_{k+1-i}$ for each $j \geq i$, so we add c_i to the committee, which has $load(c_i) = H_k - H_{k-i}$, and update

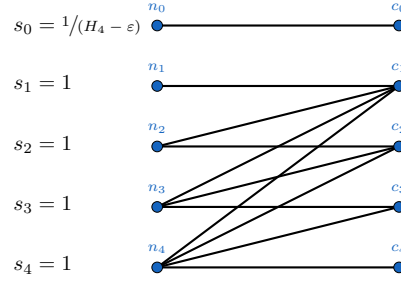


Figure 3: Instance of maximin support where seqPhragmen gives an approximation ratio of $H_k - \varepsilon$, for $k = 4$.

$load(n_j) \leftarrow H_k - H_{k_i}$ for each $j \geq i$. Finally, in the k -th round we have that $load(c_0) = 1/s_0 = H_k - \varepsilon$ and $load(c_k) = H_k$, thus we elect c_0 as the final committee member. This completes the proof. \square

C A lazy greedy algorithm

In this section we prove Theorem 13 and present LazyMMS (Algorithm 6), a variant of MMS (Algorithm 1) that is faster by a factor $\Theta(k)$ and offers virtually the same approximation guarantee.

Data: Approval graph $G = (N \cup C, E)$, vector s of vote strengths, committee size k , threshold support $t \geq 0$. Initialize $A = \emptyset$, $w = 0 \in \mathbb{R}_{\geq 0}^E$, and $U = C$ // U is the set of “uninspected” candidates;

```

while  $U \neq \emptyset$  do
    Find  $c_{\max} \in \arg \max_{c' \in U} score(c')$  // try the uninspected candidate with highest score ;
    Remove  $c_{\max}$  from  $U$ ;
    Compute a balanced edge weight vector  $w'$  for  $A + c_{\max}$ ;
    if  $supp_{w'}(A + c_{\max}) \geq t$  then // candidate is “good enough” to add
        Update  $A \leftarrow A + c_{\max}$  and  $w \leftarrow w'$ ;
        if  $|A| = k$  then return  $(A, w)$ ;
    end
end
return a failure message;

```

Algorithm 6: LazyMMS

Algorithm LazyMMS is lazier than MMS in the sense that for each candidate it inspects, it decides on the spot whether to add it to the current partial solution, if the candidate is “good enough”, or permanently reject it. In particular, each inspected candidate entails the computation of a single balanced weight vector, as opposed to $O(k)$ in MMS. Clearly, in each of the $O(|C|)$ iterations the complexity is dominated by this computation, which takes time Bal (the highest score in U can be computed in time $O(|E| \log k)$ with a variant of Algorithm MaxScore; see Theorem 15).

For a threshold $t \geq 0$ given as input, the algorithm either succeeds and returns a full solution (A, w) with $supp_w(A) \geq t$, or it returns a failure message. The idea is then to run trials of LazyMMS over several input thresholds t , where each trial executes in time $O(\text{Bal} \cdot |C|)$, performing binary search to converge to a value of t where it flips from failure to success, and return the output of the last successful trial. We start by proving that for low enough values of t , the algorithm is guaranteed to succeed. We highlight that in the following proof the order in which the candidate set is traversed (i.e. always selecting the uninspected candidate with highest score) is irrelevant.

Lemma 29. *If (A^*, w^*) is an optimal solution to the given instance of maximin support, and $t^* = supp_{w^*}(A^*)$, then for any input threshold t with $0 \leq t \leq t^*/2$, Algorithm LazyMMS is guaranteed to succeed.*

Proof. Assume by contradiction that for some input threshold $t \leq t^*/2$, LazyMMS fails. Thus, after traversing the whole candidate set C , the algorithm ends up with a partial solution (A, w) with $|A| < k$ and $supp_w(A) \geq t$. By Lemma 12, there must be a candidate $c' \in A^* \setminus A$ and a feasible solution $(A + c', w')$ such that $supp_{w'}(A + c') \geq t$. Notice as well that for any subset S of $A + c'$, vector w' provides a support of at least t , so any balanced weight vector for S also provides a support of at least t . This implies that at whichever point the algorithm inspected candidate c' , it

should have included it in the partial solution, which at that time was a subset of A . Hence, c' should be contained in A , and we reach a contradiction. \square

In the next lemma we establish the number of trials needed to achieve a solution whose value is within a factor $(2 + \varepsilon)$ from optimal for any $\varepsilon > 0$.

Lemma 30. *For any $\varepsilon > 0$, $O(\log(1/\varepsilon))$ trials of LazyMMS are sufficient to obtain a solution whose maximin support value is within a factor $(2 + \varepsilon)$ from optimal.*

Proof. First, we need to compute an α -factor estimation of the optimal objective value t^* . One way to do that is to use the BalPhragmms algorithm (Section 4.2), which provides an approximation guarantee of $\alpha = 3.15$ and runs in time $O(\text{Bal} \cdot k)$.⁷ If t is the objective value of its output, and we initialize the variables $t' \leftarrow t/2$, $t'' \leftarrow \alpha \cdot t$, then we have the properties that $t' < t''$ and that LazyMMS succeeds for threshold t' and fails for threshold t'' . We keep these properties as loop invariants as we perform binary search over Algorithm LazyMMS, in each iteration setting the new threshold value to the geometric mean of t' and t'' . This way, the ratio t''/t' starts with a constant value 2α , and is square-rooted in each iteration. By Lemma 29, to achieve a $(2 + \varepsilon)$ -factor guarantee it suffices to find threshold values $t' < t''$ such that LazyMMS succeeds for t' and fails for t'' and whose ratio is bounded by $t''/t' \leq 1 + \varepsilon/2$, and return the output for t' . If it takes $T + 1$ iterations for our binary search to bring this ratio below $(1 + \varepsilon/2)$, then $(2\alpha)^{1/2^T} > (1 + \varepsilon/2)$, so $T = O(\log(\varepsilon^{-1} \log(\alpha))) = O(\log(\varepsilon^{-1}))$. This completes the proof. \square

Finally, we prove that whenever Algorithm LazyMMS succeeds and returns a full solution, this solution satisfies PJR. For this, we exploit the order in which we inspected the candidates. This completes the proof of Theorem 13.

Lemma 31. *For any input threshold t , at the end of each iteration of Algorithm LazyMMS we have that if (A, w) is the current partial balanced solution, then $\text{supp}_w(A) \geq \max_{c' \in C \setminus A} \text{score}_{(A, w)}(c')$. Therefore, if threshold t is such that the algorithm succeeds and returns a full solution, this solution satisfies PJR.*

Proof. The second statement immediately follows from the first one together with Lemma 18, hence we focus on proving the first statement. Fix an input threshold t and some iteration of LazyMMS, and let (A, w) be the partial solution at the end of it. We consider three cases. Case 1: If all candidates inspected so far have been added to the solution, then up to this point the construction coincides with Algorithm BalPhragmms, and the claim follows by Lemma 19. Case 2: Suppose the last iteration was the first to reject a candidate, and let c' be this candidate. Then, c' has the highest score in $C \setminus A$, and we claim that this score must be below threshold t , and hence below $\text{supp}_w(A)$. Otherwise, by Lemma 14 we have that Algorithm Insert(A, w, c', t) could find a weight vector that gives $A + c'$ a support above t , so a balanced weight vector would also give $A + c'$ a support above t which contradicts the fact that c' was rejected. Case 3: If a candidate was rejected in a previous iteration, then at the time of the first rejection we had that the highest score in $C \setminus A$ was below t , and this inequality must continue to hold true by Lemma 17, because scores can only decrease in future iterations. This completes the proof. \square

D Algorithmic considerations of the new heuristic

The goal of this section is threefold. First, we prove Theorem 15 and describe efficient algorithms related to the Phragmms heuristic proposed in Section 4. Second, we compare it to seqPhragmen (Algorithm 5) and argue that the new heuristic can be seen as a natural progression in terms of algorithmic complication. Finally, we improve the runtime analysis of Algorithm BalPhragmms and show that each iteration can be executed in time $O(\text{Bal} + |E|)$, down from $O(\text{Bal} + |E| \cdot \log k)$ as was proved in Section 4.2.

As we did in Section 4.1, we assume in the following that the election instance $(G = (N \cup C, E), s, k)$ is known and does not need to be given as input. Instead, the input is a partial solution (A, w) with $|A| \leq k$. The list of committee member supports $(\text{supp}_w(c))_{c \in A}$ is implicitly passed by reference and updated in every algorithm. We start with Algorithm 7, which shows how to find the candidate with highest pre-score for a given threshold t .

Lemma 32. *For a partial solution (A, w) and threshold $t \geq 0$, MaxPrescore(A, w, t) executes in time $O(|E|)$, and returns a tuple (c_t, p_t) such that $c_t \in C \setminus A$ and $p_t = \text{prescore}(c_t, t) = \max_{c' \in C \setminus A} \text{prescore}(c', t)$.*

Proof. The correctness of the algorithm directly follows from the definitions of slack and pre-score. The running time is $O(|E|)$ because each edge in the approval graph $G = (N \cup V, E)$ is inspected at most once in each of the two loops. The first loop also inspects each voter, but we have $|N| = O(|E|)$ since we assume that G has no isolated vertices. \square

⁷In fact, it can be checked that the BalPhragmms algorithm is equivalent to an execution of LazyMMS with threshold $t = 0$.

Data: Partial solution (A, w) , threshold $t \geq 0$.
for each voter $n \in N$ **do** compute $slack(n, t) = s_n - \sum_{c \in A \cap C_n} w_{nc} \cdot \min\{1, t/supp_w(c)\}$;
for each candidate $c' \in C \setminus A$ **do** compute $prescore(c', t) = \sum_{n \in N_{c'}} slack(n, t)$;
Find a candidate $c_t \in \arg \max_{c' \in C \setminus A} prescore(c', t)$;
return $(c_t, prescore(c_t, t))$;

Algorithm 7: MaxPrescore(A, w, t)

For a fixed partial solution (A, w) and for each candidate $c' \in C \setminus A$, consider the function $f_{c'}(t) := prescore(c', t) - t$ in the interval $[0, \infty)$. Notice from the definition of pre-score (5) that this function is convex, continuous and strictly decreasing with no lower bound, and that $f_{c'}(0) \geq 0$; hence it has a unique root which corresponds precisely to $score(c')$. We could approximate this root via binary search – however, we can do better. Function $f_{c'}(t)$ is piece-wise linear: if we sort the member supports $\{supp_w(c) : c \in A\} = \{t_1, \dots, t_r\}$ so that $t_1 < \dots < t_r$ for some $r \leq |A|$, then $f_{c'}(t)$ is linear in each of the intervals $[0, t_1), [t_1, t_2), \dots, [t_r, \infty)$.

Similarly, function $f(t) := \max_{c' \in C \setminus A} f_{c'}(t) = \max_{c' \in C \setminus A} prescore(c', t) - t$ is continuous and strictly decreasing in the interval $[0, \infty)$, with a unique root $t_{\max} = \max_{c' \in C \setminus A} score(c')$. Unfortunately, this function is in general not linear within each of the intervals above.⁸ Still, it will be convenient to use binary search to identify the interval that contains t_{\max} . We do so in Algorithm 8. The next lemma follows from our exposition and its proof is skipped.

Data: Partial solution (A, w) .

Sort the member supports to obtain $0 = t_0 < t_1 < \dots < t_r$, where $\{t_1, \dots, t_r\} = \{supp_w(c) : c \in A\}$;

if $p_{t_r} \geq t_r$ **where** $(c_{t_r}, p_{t_r}) \leftarrow \text{MaxPrescore}(A, w, t_r)$ **then return** t_r ;

Let $j_{lo} = 0, j_{hi} = r - 1$;

while $j_{lo} < j_{hi}$ **do**

 Let $j = \lceil (j_{lo} + j_{hi})/2 \rceil$;

if $p_{t_j} \geq t_j$ **where** $(c_{t_j}, p_{t_j}) \leftarrow \text{MaxPrescore}(A, w, t_j)$ **then** Set $j_{lo} \leftarrow j$ **else** Set $j_{hi} \leftarrow j - 1$;

end

return $t_{j_{lo}}$;

Algorithm 8: FindInterval(A, w)

Lemma 33. For a partial solution (A, w) , Algorithm FindInterval(A, w) makes $O(\log |A|)$ calls to MaxPrescore, and thus runs in time $O(|E| \cdot \log k)$. It returns a value t' so that $t' \leq t_{\max} := \max_{c' \in C \setminus A} score(c')$, and for each candidate $c' \in C \setminus A$, the value of $prescore(c', t)$ is linear in t within the interval $[t', t_{\max}]$.

Once we have identified such a value t' , we exploit the fact that for each $c' \in C \setminus A$, function $f_{c'}(t)$ is linear inside the interval $[t', t_{\max}]$. If we fix a candidate c' and linearize function $f_{c'}(t)$ by extending its linear behavior within $[t', t_{\max}]$ onto the full domain $[0, \infty)$, and we denote its corresponding unique root by $t_{c'}$, then we have

$$\begin{aligned} 0 &= f_{c'}(t_{c'})|_{\text{linearized around } [t', t_{\max}]} \\ &= prescore(c', t_{c'})|_{\text{linearized around } [t', t_{\max}]} - t_{c'} \\ &= \sum_{n \in N_{c'}} slack(n, t)|_{\text{linearized around } [t', t_{\max}]} - t_{c'} \\ &= \sum_{n \in N_{c'}} \left(s_n - \sum_{c \in A \cap C_n: supp_w(c) < t'} w_{nc} - \sum_{c \in A \cap C_n: supp_w(c) \geq t'} w_{nc} \cdot t_{c'} / supp_w(c) \right) - t_{c'}, \end{aligned}$$

where we used the definitions of pre-score and slack. Solving for $t_{c'}$, we obtain

$$t_{c'} = \frac{\sum_{n \in N_{c'}} \left(s_n - \sum_{c \in A \cap C_n: supp_w(c) < t'} w_{nc} \right)}{1 + \sum_{n \in N_{c'}} \sum_{c \in A \cap C_n: supp_w(c) \geq t'} \frac{w_{nc}}{supp_w(c)}} =: linscore_{(A, w)}(c', t').$$

Notice that the expression above depends only on t' and not on t_{\max} . Replacing t' by x in the expression above, we define the linearized score of c' around point x (denoted by $linscore_{(A, w)}(c', x)$ and shortened to $linscore(c', t')$)

⁸In each of these $O(k)$ intervals, function $f(t)$ is piece-wise linear with $O(|C|)$ pieces. We could thus find the root of $f(t)$ via binary search by performing $O(\log k + \log |C|)$ calls to MaxPrescore. However, our approach only requires $O(\log k)$ such calls.

for each candidate $c' \in C \setminus A$ and each $x \geq 0$. Since $f_{c'}(t)$ is a convex decreasing function over $[0, \infty)$ and its linearization around x is a straight line tangential to it (meeting at point x), then the root of this linearization must fall to the left of its own root, i.e. it holds that

$$\text{lin score}(c', x) \leq \text{score}(c') \quad \text{for each } c' \in C \setminus A \text{ and each } x \geq 0.$$

On the other hand, for the candidate c_{\max} with highest score t_{\max} , and for $x = t'$ as in Lemma 33, these two roots must coincide, i.e. $t_{\max} = \text{score}(c_{\max}) = \text{lin score}(c_{\max}, t')$. Consequently, c_{\max} also has the highest linearized score around t' among all candidates in $C \setminus A$, and we can exploit this fact to find it. We formalize these observations in Algorithm 9 and the lemma below.

Data: Partial solution (A, w) .

Let $t' \leftarrow \text{FindInterval}(A, w)$;

for each voter $n \in N$ **do**

 Compute $p_n := s_n - \sum_{c \in A \cap C_n: \text{supp}_w(c) < t'} w_{nc}$;

 Compute $q_n := \sum_{c \in A \cap C_n: \text{supp}_w(c) \geq t'} w_{nc} / \text{supp}_w(c)$;

end

for each candidate $c' \in C \setminus A$ **do** compute $\text{lin score}(c', t') = \frac{\sum_{n \in N_{c'}} p_n}{1 + \sum_{n \in N_{c'}} q_n}$;

Find a candidate $c_{\max} \in \arg \max_{c' \in C \setminus A} \text{lin score}(c', t')$;

return $(c_{\max}, \text{lin score}(c_{\max}, t'))$;

Algorithm 9: MaxScore(A, w)

Lemma 34. *For a partial solution (A, w) , Algorithm MaxScore(A, w) runs in time $O(|E| \cdot \log k)$ and returns a tuple (c_{\max}, t_{\max}) such that $t_{\max} = \text{score}(c_{\max}) = \max_{c' \in C \setminus A} \text{score}(c')$.*

Proof. The correctness of the algorithm follows from the definition of linearized score and the arguments above. Each of the **for** loops executes in time $O(|E|)$ because in each one of them each edge is examined at most once. Thus, the running time is dominated by the call to function FindInterval(A, w), which takes time $O(|E| \cdot \log k)$. \square

This completes the proof of Theorem 15. We highlight again that an iteration of the Phragmms executes in time $O(|E| \cdot \log k)$, almost matching the complexity of seqPhragmen which is $O(|E|)$ time per iteration.

Next, we discuss the similarities and differences between the Phragmms and seqPhragmen heuristics. To this end, consider an execution of Phragmms in the case that the input partial solution (A, w) is balanced. In this case, by point 3 of Lemma 4, for each voter $n \in N$ and each member $c \in A \cap C_n$ such that $w_{nc} > 0$ it must hold that $\text{supp}_w(c) = \text{supp}_w(A \cap C_n)$. If we then compute the linearized score of a candidate $c' \in C \setminus A$ around the origin (setting $x = 0$) we obtain

$$\begin{aligned} \text{lin score}(c', 0) &= \frac{\sum_{n \in N_{c'}} s_n}{1 + \sum_{n \in N_{c'}} \frac{1}{\text{supp}_w(A \cap C_n)} \sum_{c \in A \cap C_n} w_{nc}} \\ &= \frac{\sum_{n \in N_{c'}} s_n}{1 + \sum_{n \in N_{c'}} \frac{s_n}{\text{supp}_w(A \cap C_n)}}, \end{aligned}$$

where in the second line we used the fact that $\sum_{c \in A \cap C_n} w_{nc} = s_n$ must hold for each voter n for which $A \cap C_n \neq \emptyset$, by point 2 of Lemma 4 (if $A \cap C_n = \emptyset$ then $\text{supp}_w(A \cap C_n) = \infty$ by convention and the corresponding term in the denominator vanishes). This linearized score corresponds precisely to the inverse of the candidate *load* function being minimized in the seqPhragmen heuristic (see Algorithm 5), with a corresponding voter load function set to the inverse of $\text{supp}_w(A \cap C_n)$. Therefore, the two heuristics coincide in the specific case that the input partial solution is balanced and the linearized score around the origin is used. Put differently, we can say that the new heuristic provides two main advantages: First, by using edge weights it defines a more robust notion of loads, which enables it to deal with an arbitrary input partial solution without any guarantees on how it was constructed. Second, by considering further linearizations of the pre-score function it grants new candidates higher scores and thus adds them with higher supports.

Finally, we reconsider the complexity of BalPhragmms (Algorithm 3). At the start of each iteration with current partial solution (A, w) , notice by Lemma 19 that the highest score t_{\max} must be lower than the least member support $t_1 = \text{supp}_w(A)$. So, t_{\max} lies in the interval $[0, t_1]$, and we can skip the computation of Algorithm FindInterval(A, w) as we know that it would return $t' = 0$. Without this computation, MaxScore(A, w) (Algorithm 9) runs in time

$O(|E|)$, so the runtime of a full iteration of BalPhragmms can be performed in time $O(\text{Bal} + |E|)$, down from $O(\text{Bal} + |E| \cdot \log k)$ as was established in Section 4.2.

E Delayed proofs

Lemma 35. *Let $w \in \mathbb{R}_{\geq 0}^E$ be a feasible weight vector for a given instance, let $c, c' \in C$ be two candidates with $\text{supp}_w(c) > \text{supp}_w(c')$, and suppose there is a path $p \in \mathbb{R}^E$ that carries some non-zero flow from c to c' , with zero excess elsewhere. If $w + p$ is non-negative and feasible, then w is not balanced for any committee A that contains c' .*

Proof. Fix a committee $A \subseteq C$ that contains c' . If c is not in A , then $w + p$ provides a greater sum of member supports over A than w , so the latter is not balanced as it does not maximize this sum. Now suppose both c and c' are in A . Let $\lambda > 0$ be the flow value carried by p , let $\varepsilon := \min\{\lambda, (\text{supp}_w(c) - \text{supp}_w(c'))/2\} > 0$, and let p' be the scalar multiple of p whose flow value is ε . By an application of Lemma 7 over w and $w' := w + p$, and the fact that p' is a sub-flow of $p = w' - w$, we have that $w + p'$ is non-negative and feasible. Moreover, both w and $w + p'$ clearly provide the same sum of member supports over A . Finally, if we compare their sums of member supports squared, we have that

$$\begin{aligned} \sum_{d \in A} \text{supp}_w^2(d) - \sum_{d \in A} \text{supp}_{w+p'}^2(d) &= \text{supp}_w^2(c) + \text{supp}_w^2(c') - \text{supp}_{w+p'}^2(c) - \text{supp}_{w+p'}^2(c') \\ &= \text{supp}_w^2(c) + \text{supp}_w^2(c') - (\text{supp}_w(c) - \varepsilon)^2 - (\text{supp}_w(c) + \varepsilon)^2 \\ &= 2\varepsilon \cdot (\text{supp}_w(c) - \text{supp}_w(c') - \varepsilon) \geq 2\varepsilon \cdot (2\varepsilon - \varepsilon) = 2\varepsilon^2 > 0. \end{aligned}$$

Therefore, w is not balanced for A , as it does not minimize the sum of member supports squared. \square

Proof of Lemma 4. Fix a balanced partial solution (A, w) . The first statement says that function $F_r(w') := \min_{A' \subseteq A, |A'|=r} \sum_{c \in A'} \text{supp}_{w'}(c)$ is maximized by vector w over all feasible vectors $w' \in \mathbb{R}_{\geq 0}^E$, for all $1 \leq r \leq |A|$. Assume by contradiction that there is a parameter r and feasible w' such that $F_r(w') > F_r(w)$. We can also assume without loss of generality that

1. $\sum_{c \in A} \text{supp}_{w'}(c) = \sum_{c \in A} \text{supp}_w(c)$, i.e. w' also maximizes the sum of member supports, as does w by definition of balancedness; and
2. we enumerate the candidates in $A = \{c_1, \dots, c_{|A|}\}$ so that whenever $i < j$ we have that $\text{supp}_w(c_i) \leq \text{supp}_w(c_j)$, and in case of a tie, $\text{supp}_w(c_i) = \text{supp}_w(c_j)$, we have that $\text{supp}_{w'}(c_i) \leq \text{supp}_{w'}(c_j)$.

With a candidate enumeration as above, it follows that $F_r(w) = \sum_{i=1}^r \text{supp}_w(c_i)$, while for vector w' we have the inequality $F_r(w') \leq \sum_{i=1}^r \text{supp}_{w'}(c_i)$. Thus, by our assumption by contradiction,

$$\begin{aligned} \sum_{i=1}^r \text{supp}_{w'}(c_i) &\geq F_r(w') > F_r(w) = \sum_{i=1}^r \text{supp}_w(c_i), \quad \text{and} \\ \sum_{i=r+1}^{|A|} \text{supp}_{w'}(c_i) &= \sum_{i=1}^{|A|} \text{supp}_{w'}(c_i) - \sum_{i=1}^r \text{supp}_{w'}(c_i) \\ &= \sum_{i=1}^{|A|} \text{supp}_w(c_i) - \sum_{i=1}^r \text{supp}_{w'}(c_i) \\ &< \sum_{i=1}^{|A|} \text{supp}_w(c_i) - \sum_{i=1}^r \text{supp}_w(c_i) = \sum_{i=r+1}^{|A|} \text{supp}_w(c_i). \end{aligned}$$

Now define the edge vector $f := w' - w \in \mathbb{R}^E$ and consider it as a flow over the network $(N \cup A, E)$. We have that f has a zero net demand over set N , by our first assumption, and the previous two inequalities show that f has a positive net demand over set $\{c_1, \dots, c_r\}$ and a positive net excess over set $\{c_{r+1}, \dots, c_{|A|}\}$. Thus, by the flow

decomposition theorem, f can be decomposed into circulations and simple paths, where every path starts in a vertex with positive demand and ends in a vertex with positive excess, and there must be a simple path p carrying non-zero flow from c_j to c_i for some $1 \leq i \leq r < j \leq |A|$. Moreover, by our second assumption, it must be the case that $\text{supp}_w(c_i) < \text{supp}_w(c_j)$, because in case of a tie we would have that $\text{supp}_{w'}(c_i) < \text{supp}_{w'}(c_j)$ and so f would have a net excess on c_i and net demand on c_j , and the c_i -to- c_j path would not exist in the flow decomposition. But now, by Lemma 7, vector $w + p$ is non-negative and feasible, and by Lemma 35, w is not balanced for A , and we reach a contradiction.

The second statement follows directly from the fact that w maximizes the sum of member supports, and thus all of the vote strength of all represented voters (i.e. all voters in $\cup_{c \in A} N_c$) must be directed to members of A . We move on to the third statement. Assume by contradiction that there is a voter $n \in N$ and two candidates $c, c' \in A \cap C_n$ such that $w_{nc} > 0$ and $\text{supp}_w(c) > \text{supp}_w(c')$. Let $\varepsilon := \min\{w_{nc}, (\text{supp}_w(c) - \text{supp}_w(c'))/2\} > 0$, and define a path $p \in \mathbb{R}^E$ carrying flow ε from c to c' via n , i.e. $p_{nc'} = -p_{nc} = \varepsilon$ and $p_e = 0$ for every other edge $e \in E$. It can be checked that $w + p$ is non-negative and feasible, so by Lemma 35 w is not balanced for A , which is a contradiction.

Finally, we prove that if a feasible weight vector satisfies conditions 2 and 3, then it is balanced for A . In fact, we claim that all such weight vectors provide the same list of member supports $(\text{supp}_w(c))_{c \in A}$, and hence are all balanced. Let $w, w' \in \mathbb{R}_{\geq 0}^E$ be two such weight vectors. It easily follows from feasibility (inequality 2) and condition 2 that both provide the same sum of member supports, namely $\sum_{c \in A} \text{supp}_w(c) = \sum_{c \in A} \text{supp}_{w'}(c) = \sum_{n \in \cup_{c \in A} N_c} s_n$. Now, assume by contradiction and wlog that there is a candidate $c \in A$ for which $\text{supp}_w(c) > \text{supp}_{w'}(c)$, and consider the flow vector $f := w' - w$ over the network induced by $N \cup A$. Clearly, all nodes in N have zero excess, while c has positive excess. By the flow decomposition theorem, f can be decomposed into circulations and single paths, where every path starts in a node with net excess and ends in a node with net demand; so, there must be non-zero path p that starts in c and ends in a candidate $c' \in A$ with net demand. Now, path p alternates between candidates and voters, and there must be three consecutive nodes c_1, n, c_2 in it, with $c_1, c_2 \in A \cap C_n$, such that c_1 has positive excess and c_2 does not, i.e. $\text{supp}_{w'}(c_1) < \text{supp}_w(c_2)$ and $\text{supp}_{w'}(c_2) \geq \text{supp}_w(c_2)$, which in turn implies that either $\text{supp}_{w'}(c_1) < \text{supp}_{w'}(c_2)$ or $\text{supp}_w(c_2) < \text{supp}_w(c_1)$ (or both). If $\text{supp}_{w'}(c_1) < \text{supp}_{w'}(c_2)$, we reach a contradiction to the fact that w' satisfies condition 3 and that w'_{nc_2} must be positive since the flow in p moves from n to c_2 ; and similarly if $\text{supp}_w(c_2) < \text{supp}_w(c_1)$ we reach a contradiction to the fact that w satisfies condition 3 and that w_{nc_1} must be positive since the flow in p moves from c_1 to n . This completes the proof of the lemma. \square

Proof of Lemma 6. Let $A' \subseteq A$ be the non-empty subset that minimizes the expression $\frac{1}{|A'|} \sum_{n \in \cup_{c \in A'} N_c} s_n$. Hence,

$$\begin{aligned} \text{supp}_w(A) &\leq \text{supp}_w(A') \leq \frac{1}{|A'|} \sum_{c \in A'} \text{supp}_w(c) && \text{(by an averaging argument)} \\ &= \frac{1}{|A'|} \sum_{c \in A'} \sum_{n \in N_c} w_{nc} \leq \frac{1}{|A'|} \sum_{n \in \cup_{c \in A'} N_c} \sum_{c \in C_n} w_{nc} \\ &\leq \frac{1}{|A'|} \sum_{n \in \cup_{c \in A'} N_c} s_n && \text{(by 2).} \end{aligned}$$

This proves one inequality. To prove the opposite inequality, we use the fact that (A, w) is a balanced solution. Let $A_{\min} \subseteq A$ be the set of committee members with least support, i.e. those $c \in A$ with $\text{supp}_w(c) = \text{supp}_w(A)$. Then,

$$\begin{aligned} \text{supp}_w(A_{\min}) &= \frac{1}{|A_{\min}|} \sum_{c \in A_{\min}} \text{supp}_w(c) \\ &= \frac{1}{|A_{\min}|} \sum_{c \in A_{\min}} \sum_{n \in N_c} w_{nc} = \frac{1}{|A_{\min}|} \sum_{n \in \cup_{c \in A_{\min}} N_c} \sum_{c \in C_n \cap A_{\min}} w_{nc} \\ &= \frac{1}{|A_{\min}|} \sum_{n \in \cup_{c \in A_{\min}} N_c} \left(\sum_{c \in C_n \cap A} w_{nc} - \sum_{c \in C_n \cap (A \setminus A_{\min})} w_{nc} \right) \\ &= \frac{1}{|A_{\min}|} \sum_{n \in \cup_{c \in A_{\min}} N_c} s_n \geq \min_{\emptyset \neq A' \subseteq A} \frac{1}{|A'|} \sum_{n \in \cup_{c \in A'} N_c} s_n, \end{aligned}$$

where we used the fact that for each voter $n \in \cup_{c \in A_{\min}} N_c$, the term $\sum_{c \in C_n \cap A} w_{nc}$ equals s_n by point 2 of Lemma 4, and the term $\sum_{c \in C_n \cap (A \setminus A_{\min})} w_{nc}$ vanishes by point 3 of Lemma 4 and the definition of set A_{\min} . This proves the second inequality and completes the proof. \square

Proof of Lemma 7. We prove the claim only for $w + f'$, as the proof for $w' - f'$ is similar. For each edge $e \in E$, we have that $(w + f')_e$ is a value between w_e and $(w + f)_e = w'_e$. As both of these values are non-negative, the same holds for $(w + f')_e$. Notice now from inequality (2) that proving feasibility corresponds to proving that the excess $(w + f')(n)$ is at most s_n for each voter $n \in N$. We have

$$(w + f')(n) = \sum_{c \in C_n} (w + f')_{nc} = \sum_{c \in C_n} w_{nc} + \sum_{c \in C_n} f'_{nc} = w(n) + f'(n).$$

If excess $f'(n)$ is non-positive, then $(w + f')(n) \leq w(n) \leq s_n$, since w is feasible. Otherwise, $f'(n) \leq f(n)$, and thus $(w + f')(n) \leq w(n) + f(n) = (w + f)(n) = w'(n) \leq s_n$, since w' is feasible. This completes the proof. \square

Proof of Lemma 17. The second statement follows directly from the first one and the definitions of slack, pre-score and score in Section 4. Hence we focus on the first statement, i.e. that $\text{supp}_w(c) \geq \text{supp}_{w'}(c)$ for each candidate $c \in A$.

Consider vector $f := w' - w \in \mathbb{R}^E$ as a flow vector over the network induced by $N \cup A'$. We need to prove that there is no candidate in A with positive demand relative to f . Assume by contradiction that there is such a candidate $c' \in A$. By the flow decomposition theorem, f can be decomposed into circulations and simple paths, where each path starts in a vertex with positive excess and ends in a vertex with positive demand. By our assumption, there must be a path p carrying non-zero flow to c' . Now, where does path p start? It cannot start in N nor in $A' \setminus A$, as otherwise weight vector $w + p$ is feasible by Lemma 7, and has a greater sum of candidate supports over A than w , which contradicts the fact that w is balanced for A and thus maximizes this sum. Therefore, it must start in another candidate c in A with positive excess. The fact that f has positive excess in c and positive demand in c' implies that

$$\text{supp}_w(c') - \text{supp}_{w'}(c') = f(c') < 0 < f(c) = \text{supp}_w(c) - \text{supp}_{w'}(c),$$

which in turn implies that either $\text{supp}_w(c) > \text{supp}_w(c')$ or $\text{supp}_{w'}(c') > \text{supp}_{w'}(c)$, or both. If $\text{supp}_w(c) > \text{supp}_w(c')$, then Lemma 35 implies that w is not balanced for A . Similarly, if $\text{supp}_{w'}(c') > \text{supp}_{w'}(c)$, notice that $w' - p$ is non-negative and balanced by Lemma 7, so again Lemma 35 applied over vector w' and path $-p$ implies that w' is not balanced for A' . Hence, in either case we reach a contradiction. \square

Proof of Lemma 22. Recall that for any set $A \subseteq \mathbb{R}$, the indicator function $1_A : \mathbb{R} \rightarrow \mathbb{R}$ is defined as $1_A(t) = 1$ if $t \in A$, and 0 otherwise. For any $i \in I$, we can write

$$\alpha_i f(x_i) = \alpha_i \int_0^{f(x_i)} dt = \alpha_i \int_0^{\lim_{x \rightarrow \infty} f(x)} 1_{(-\infty, f(x_i)]}(t) dt,$$

and thus

$$\sum_{i \in I} \alpha_i f(x_i) = \int_0^{\lim_{x \rightarrow \infty} f(x)} \left(\sum_{i \in I} \alpha_i 1_{(-\infty, f(x_i)]}(t) \right) dt = \int_0^{\lim_{x \rightarrow \infty} f(x)} \left(\sum_{i \in I: f(x_i) \geq t} \alpha_i \right) dt.$$

This is a Lebesgue integral over the measure with weights α_i . Now, conditions on function $f(x)$ are sufficient for its inverse $f^{-1}(t)$ to exist, with $f^{-1}(0) = \chi$. Substituting with the new variable $x = f^{-1}(t)$ on the formula above, where $t = f(x)$ and $dt = f'(x)dx$, we finally obtain

$$\sum_{i \in I} \alpha_i f(x_i) = \int_{\chi}^{\infty} \left(\sum_{i \in I: x_i \geq x} \alpha_i \right) (f'(x) dx).$$

\square

References

- [1] R. K. Ahuja, J. B. Orlin, C. Stein, and R. E. Tarjan. Improved algorithms for bipartite network flow. *SIAM Journal on Computing*, 23(5):906–933, 1994.
- [2] H. Aziz, M. Brill, V. Conitzer, E. Elkind, R. Freeman, and T. Walsh. Justified representation in approval-based committee voting. *Social Choice and Welfare*, 48(2):461–485, 2017.
- [3] H. Aziz, E. Elkind, S. Huang, M. Lackner, L. Sánchez-Fernández, and P. Skowron. On the complexity of extended and proportional justified representation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

- [4] H. Aziz, S. Gaspers, J. Gudmundsson, S. Mackenzie, N. Mattei, and T. Walsh. Computational aspects of multi-winner approval voting. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.
- [5] M. Brill, R. Freeman, S. Janson, and M. Lackner. Phragmén’s voting methods and justified representation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [6] J. Burdges, A. Cevallos, P. Czaban, R. Habermeier, S. Hosseini, F. Lama, H. K. Alper, X. Luo, F. Shirazi, A. Stewart, et al. Overview of Polkadot and its design considerations. *arXiv preprint arXiv:2005.13456*, 2020.
- [7] J. R. Douceur. The Sybil attack. In *International workshop on peer-to-peer systems*, pages 251–260. Springer, 2002.
- [8] G. Gallo, M. D. Grigoriadis, and R. E. Tarjan. A fast parametric maximum flow algorithm and applications. *SIAM Journal on Computing*, 18(1):30–55, 1989.
- [9] M. R. Garey and D. S. Johnson. *Computers and intractability: A guide to the theory of NP-completeness*, volume 1. WH Freeman San Francisco, 1979.
- [10] A. V. Goldberg and R. E. Tarjan. A new approach to the maximum-flow problem. *Journal of the ACM (JACM)*, 35(4):921–940, 1988.
- [11] D. S. Hochbaum and S.-P. Hong. About strongly polynomial time algorithms for quadratic optimization over submodular constraints. *Mathematical programming*, 69(1-3):269–309, 1995.
- [12] S. Janson. Phragmén’s and Thiele’s election methods. *arXiv preprint arXiv:1611.08826*, 2016.
- [13] M. Lackner and P. Skowron. Approval-based committee voting: Axioms, algorithms, and applications. *arXiv preprint arXiv:2007.01795*, 2020.
- [14] S. Nakamoto. Bitcoin: A peer-to-peer electronic cash system. Technical report, 2008.
- [15] M. Pease, R. Shostak, and L. Lamport. Reaching agreement in the presence of faults. *Journal of the ACM (JACM)*, 27(2):228–234, 1980.
- [16] D. Peters and P. Skowron. Proportionality and the limits of welfarism. *arXiv preprint arXiv:1911.11747*, 2019.
- [17] E. Phragmén. *Sur une méthode nouvelle pour réaliser, dans les élections, la représentation proportionnelle des partis*. 1894.
- [18] E. Phragmén. *Proportionella Val: en valteknisk studie*. Hökerberg, 1895.
- [19] E. Phragmén. Sur la théorie des élections multiples. *Öfversigt af Kongliga Vetenskaps-Akademiens Förhandlingar*, 53:181–191, 1896.
- [20] E. Phragmén. Till frågan om en proportionell valmetod. *Statsvetenskaplig Tidskrift*, 2(2):297–305, 1899.
- [21] L. Sánchez-Fernández, E. Elkind, M. Lackner, N. Fernández, J. A. Fisteus, P. B. Val, and P. Skowron. Proportional justified representation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [22] L. Sánchez-Fernández, N. Fernández, J. A. Fisteus, and M. Brill. The maximin support method: An extension of the d’hondt method to approval-based multiwinner elections. *arXiv preprint arXiv:1609.05370*, 2016.
- [23] P. Skowron, P. Faliszewski, and J. Lang. Finding a collective set of items: From proportional multirepresentation to group recommendation. *Artificial Intelligence*, 241:191–216, 2016.
- [24] T. N. Thiele. Om flerfolksvalg. *Oversigt over det Kongelige Danske Videnskabernes Selskabs Forhandlinger*, 1895:415–441, 1895.