




ON ASYMPTOTIC FAIRNESS IN VOTING WITH GREEDY SAMPLING

ABRAHAM GUTIERREZ,^{*} *Graz University of Technology*
SEBASTIAN MÜLLER ^{**}, *Aix-Marseille Université and IOTA Foundation*
STJEPAN ŠEBEK,^{***} *University of Zagreb*

Abstract

The basic idea of voting protocols is that nodes query a sample of other nodes and adjust their own opinion throughout several rounds based on the proportion of the sampled opinions. In the classic model, it is assumed that all nodes have the same weight. We study voting protocols for heterogeneous weights with respect to fairness. A voting protocol is *fair* if the influence on the eventual outcome of a given participant is linear in its weight. Previous work used sampling with replacement to construct a fair voting scheme. However, it was shown that using greedy sampling, i.e., sampling with replacement until a given number of distinct elements is chosen, turns out to be more robust and performant.

In this paper, we study fairness of voting protocols with greedy sampling and propose a voting scheme that is asymptotically fair for a broad class of weight distributions. We complement our theoretical findings with numerical results and present several open questions and conjectures.

Keywords: Asymptotic fairness; consensus protocol; voting scheme; heterogeneous network; Sybil protection

2020 Mathematics Subject Classification: Primary 60F99
Secondary 94A20; 91A20; 68M14

1. Introduction

This article focuses on fairness in binary voting protocols. As early as 1785, Condorcet [4] studied the principle of voting from a probabilistic point of view. One of his famous results is Condorcet's jury theorem, which states the following. Let us suppose there is a large population of voters, and each of them independently votes 'correctly' with probability $p > 1/2$. Then the probability that the outcome of a majority vote is 'correct' grows with the sample size and converges to one. In many applications, for instance, distributed computing, it is not feasible for every node to query every other participant, and a centralized entity that collects the votes of all participants and communicates the final result is not desired. Natural decentralized solutions with low message complexity are the so-called voting consensus protocols. Nodes query other

Received 7 October 2021; revision received 13 September 2022.

^{*} Postal address: Institute of Discrete Mathematics, Graz University of Technology, Graz, Austria. Email address: schnirelmann@gmail.com

^{**} Postal address: Aix-Marseille Université, CNRS, Centrale Marseille, I2M, UMR 7373, 13453 Marseille, France; IOTA Foundation, 10405 Berlin, Germany. Email address: sebastian.muller@univ-amu.fr

^{***} Postal address: Department of Applied Mathematics, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. Email address: stjepan.sebek@fer.hr

© The Author(s), 2023. Published by Cambridge University Press on behalf of Applied Probability Trust.

nodes (only a sample of the entire population) about their current opinions and adjust their own opinion throughout several rounds based on the proportion of other opinions they have observed.

These protocols may achieve good performance in noiseless and undisturbed networks. However, their performances significantly decrease with noise [6, 7] or errors [10] and may completely fail in a Byzantine setting [2]. Recently, [14] introduced a variant of the standard voting protocol, the so-called fast probabilistic consensus (FPC), that is robust in Byzantine environments. The performance of FPC was then studied using Monte Carlo simulations in [2]. The above voting protocols are tailored for homogeneous networks where all votes have equal weight. In [11, 12] FPC was generalized to heterogeneous settings. These studies also revealed that how votes are sampled does have a considerable impact on the quality of the protocol.

The aim is to choose a ‘representative’ sample from a given population. If we allow elements to occur multiple times in the sample, there are three different ways of sampling:

1. Choose with replacement until one has $m \in \mathbb{N}$ (not necessarily distinct) elements.
2. Choose with replacement until one has $k \in \mathbb{N}$ distinct elements.
3. Choose without replacement until one has $k = m$ distinct elements.

The first method is usually referred to as sampling with replacement. Although in the 1950s, e.g., in [16], the second way was called sampling without replacement, sampling without replacement nowadays usually refers to the third possibility. To avoid any confusion, in this paper we call the second possibility ‘greedy sampling’.

Most voting protocols assume that every participant has the same weight. In heterogeneous situations, this does not reflect possible differences in weight or influence of the participants. An essential way in which weights improve voting protocols is by ensuring that the voting protocol is fair in the sense that the influence of a node on another node’s opinion is proportional to its weight. This fairness is an essential feature of a voting protocol both for technical reasons (e.g., defense against Sybil attacks) and social reasons (e.g., participants may decide to leave the network if the voting protocol is unfair). Moreover, an unfair situation may incentivize participants to split their weight among several participants or increase their weight by pooling with other participants. These incentives may lead to undesired effects such as fragility against Sybil attacks and centralization.

The construction of fair voting consensus protocols with weights was recently discussed in [11, 12]. We consider a network with N nodes (or participants), identified with a finite set \mathcal{N} . The weights of the nodes are described by $(m_i)_{i \in \mathcal{N}}$ with $\sum_{i \in \mathcal{N}} m_i = 1$, $m_i \geq 0$ being the weight of the node i . Every node i has an initial state or opinion $s_i \in \{0, 1\}$. Then, at each (discrete) time step, each node chooses $k \in \mathbb{N}$ random nodes from the network and queries their opinions. This sampling can be done in one of the three ways described above. For instance, [11] studied fairness in the case of sampling with replacement. The mathematical treatment of this case is the easiest of the three possibilities. However, simulations in [12] strongly suggest that the performance of some consensus protocols are considerably better in the case of greedy sampling. The third method, sampling without replacement, is not explicitly mentioned in [11, 12]. A possible explanation is that in the case of heterogeneous weight distributions, the consensus protocol is not as robust towards differences in the perception of the weights. Nevertheless, it might be interesting to study this sampling scheme for fairness. The main object of our work is the mathematical analysis of weighted greedy sampling with respect to fairness.

The weights of the node may enter at two points during the voting: in sampling and in weighting the collected votes or opinions. We consider a first weighting function $f : [0, \infty) \rightarrow [0, \infty)$ that describes the weight of a node in the sampling. More precisely, a node i is chosen with probability

$$p_i := \frac{f(m_i)}{\sum_{j=1}^N f(m_j)}. \tag{1.1}$$

We call this function f the sampling weight function. A natural weight function is $f \equiv id$; a node is chosen proportional to its weight.

As discussed later in the paper, we are interested in how the weights influence the voting if the number of nodes in the network tends to infinity. Therefore, we often consider the situation with an infinite number of nodes. The weights of these nodes are again described by $(m_i)_{i \in \mathcal{N}}$ with $\sum_{i \in \mathcal{N}} m_i = 1$. A network of N nodes is then described by setting $m_i = 0$ for all but N nodes.

Once a node has chosen k distinct elements, by greedy sampling, it calculates a weighted mean opinion of these nodes. Let us denote by S_i the multi-set of the sample for a given node i . The mean opinion of the sample is

$$\eta_i := \frac{\sum_{j \in S_i} g(m_j)s_j}{\sum_{j \in S_i} g(m_j)}, \tag{1.2}$$

where $g : [0, \infty) \rightarrow [0, \infty)$ is a second weight function that we dub the averaging weight function. The pair (f, g) of the two weight functions is called a voting scheme.

In standard majority voting every node adjusts its opinion as follows: if $\eta_i < 1/2$ it updates its own opinion s_i to 0, and if $\eta_i > 1/2$ it updates its opinion to 1. The case of a draw, $\eta_i = 1/2$, may be solved by randomization or by deterministically choosing one of the options. After the opinion update, every node re-samples; this procedure is continued until some stopping condition is verified. In general, the aim of such a protocol is for all nodes to ultimately agree on one opinion—in other words, to find consensus. As mentioned above, this kind of protocol works well in a non-faulty environment. However, if some nodes do not follow the rules, or even try to hinder the other nodes from reaching consensus, then the protocol may fail to yield consensus. In this case, one speaks of honest nodes, the nodes which follow the protocol, and malicious nodes, the nodes that try to interfere. An additional feature was introduced by [14] that makes this kind of consensus protocol robust to some given proportion of malicious nodes in the network.

Let us briefly explain this crucial feature. As in [2, 11, 12] we consider a basic version of the FPC introduced in [14]. Let $U_t, t = 1, 2, \dots$, be independent and identically distributed random variables with law $\text{Unif}([\beta, 1 - \beta])$ for some parameter $\beta \in [0, 1/2]$. Every node i has an opinion or state. We denote by $s_i(t)$ the opinion of the node i at time t . Opinions take values in $\{0, 1\}$. Every node i has an initial opinion $s_i(0)$. The update rules for the opinion of a node i are then given by

$$s_i(1) = \begin{cases} 1 & \text{if } \eta_i(1) \geq \tau, \\ 0 & \text{otherwise,} \end{cases}$$

for some $\tau \in [0, 1]$. For $t \geq 1$,

$$s_i(t + 1) = \begin{cases} 1 & \text{if } \eta_i(t + 1) > U_t, \\ 0 & \text{if } \eta_i(t + 1) < U_t, \\ s_i(t) & \text{otherwise.} \end{cases}$$

Note that if $\tau = \beta = 0.5$, FPC reduces to a standard majority consensus. It is important that the above sequence of random variables U_t is the same for all nodes. The randomness of the threshold effectively reduces the capabilities of an attacker to control the opinions of honest nodes, and it also increases the rate of convergence in the case of honest nodes only. Since in this paper we focus our attention mainly on the construction and analysis of the voting schemes (f, g) , we refer to [2, 11, 12] for more details on FPC.

We concentrate mostly on the case $f \equiv id$ and $g \equiv 1$. For the voting scheme with sampling with replacement, it was shown in [11, Theorem 1] that for $g \equiv 1$, i.e., when the opinions of different nodes are not additionally weighted after the nodes are sampled, the voting scheme (f, g) is fair (see Definition 3) if and only if $f \equiv id$. For $f \equiv id$, the probability of sampling a node j satisfies $p_j = m_j$, because we assumed that $\sum_{i \in \mathcal{N}} m_i = 1$. In many places we use m_j and p_j interchangeably, with both variables referring simultaneously to the weight of the node j and the probability that the node j is sampled.

Our primary goal is to verify whether the voting scheme $(id, 1)$ is fair in the case of greedy sampling. We show in Proposition 4 that the voting scheme $(id, 1)$ is in general not fair. For this reason, we introduce the notion of asymptotic fairness; see Definition 5. Even though the definition of asymptotic fairness is very general, the best example to keep in mind is when the number of nodes grows to infinity. An important question related to the robustness of the protocol against Sybil attacks is whether the gain in influence on the voting obtained by splitting one node into ‘infinitely’ many nodes is limited.

We find a sufficient condition on the sequence of weight distributions $\{(m_i^{(n)})_{i \in \mathcal{N}}\}_{n \in \mathbb{N}}$ for asymptotic fairness; see Theorem 1. In particular, this ensures robustness against Sybil attacks for wide classes of weight distributions. However, we also note that there are situations that are not asymptotically fair; see Corollary 2 and Remark 5.

A key ingredient of our proof is a preliminary result on greedy sampling. This is a generalization of some of the results of [16]. More precisely, we obtain a formula for the joint distribution of the random vector $(A_k(i), v_k)$. Here, the random variable v_k , defined in (2.1), counts the number of samplings needed to sample k different elements, and the random variable $A_k(i)$, defined in (2.2), counts how many times in those v_k samplings the node i is sampled. The result on asymptotic fairness, Corollary 2, relies on a stochastic coupling that compares the nodes’ influence before and after splitting. We use this coupling again in the simulations in Section 5; it considerably improves the convergence of our simulations by reducing the variance.

Fairness plays a prominent role in many areas of science and applications. It is therefore not astonishing that it plays a part also in distributed ledger technologies. For instance, proof-of-work in Nakamoto consensus ensures that the probability of creating a new block is proportional to the computational power of a node; see [3] for an axiomatic approach to block rewards and further references. In proof-of-stake blockchains, the probability of creating a new block is usually proportional to the node’s balance. However, this does not always have to be the optimal choice; see [8, 13].

Our initial motivation for this paper was to show that the consensus protocol described in [15] is robust against splitting and merging. Both effects are undesirable in a decentralized and permissionless distributed system. We refer to [11, 12] for more details. Besides this, we believe that the study of the different voting schemes is of theoretical interest and that many natural questions are still open; see Section 5.

We organize the article as follows. Section 2 defines the key concepts of this paper: voting power, fairness, and asymptotic fairness. We also recall Zipf’s law, which we use to model the weight distribution of the nodes. Even though our results are obtained in a general setting, we discuss in several places how these results apply to the case of Zipf’s law; see Subsection 2.2

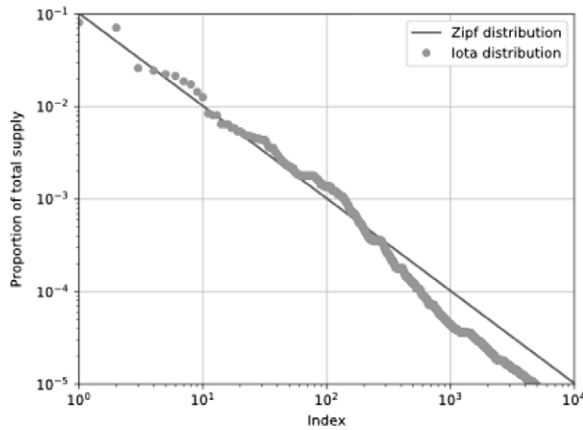


FIGURE 1. Relative distribution of the top 10,000 IOTA addresses with a fitted Zipf distribution with $s = 1.1$, July 2020.

and Figure 1. Section 3 is devoted to studying greedy sampling on its own. We find the joint probability distribution of sample size and occurrences of the nodes, $(A_k(i), v_k)$, and develop several asymptotic results that we use in the rest of the paper. In Section 4 we show that the voting scheme $(id, 1)$ is in general not fair. However, we give a sufficient condition on the sequence of weight distributions that ensures asymptotic fairness. We provide an example where, without this condition, the voting scheme $(id, 1)$ is not asymptotically fair. Section 5 contains a short simulation study. Besides illustrating the theoretical results developed in the paper, we investigate the cases when some of the assumptions we impose in our theoretical results are not met. Last but not least, we present some open problems and conjectures in Section 5. To keep the presentation as clear as possible, we present some technical results in Appendices A and B.

2. Preliminaries

2.1. Main definitions

We now introduce this paper’s key concepts: greedy sampling, voting scheme, voting power, fairness, and asymptotic fairness.

We start with the definition of greedy sampling. We consider a probability distribution $P = (p_i)_{i \in \mathcal{N}}$ on \mathcal{N} and an integer $k \in \mathbb{N}$. We sample with replacement until k different nodes have been chosen. The number of samplings needed to choose k different nodes is given by the random variable

$$v_k := v_k^{(P)} := \text{the number of samplings with replacement from the distribution } P \text{ until } k \text{ different nodes have been sampled.} \tag{2.1}$$

The outcome of a sampling will be denoted by the multi-set

$$S := \{a_1, a_2, \dots, a_{v_k}\};$$

here the a_i take values in \mathcal{N} . Furthermore, for any $i \in \mathcal{N}$, let

$$A_k(i) := A_k^{(P)}(i) := \#\{j \in \{1, 2, \dots, v_k\} : a_j = i\} \tag{2.2}$$

be the number of occurrences of i in the multi-set $S = \{a_1, a_2, \dots, a_{v_k}\}$.

Every node i is assigned a weight m_i . Together with a function $f : [0, \infty) \rightarrow [0, \infty)$ that we call the sampling weight function, the weights define a probability distribution $P = (p_i)_{i \in \mathcal{N}}$ on \mathcal{N} by

$$p_i = \frac{f(m_i)}{\sum_{j \in \mathcal{N}} f(m_j)}.$$

We consider a second weight function $g : [0, \infty) \rightarrow [0, \infty)$, the averaging weight function, that weights the sample's opinions; see Equation (1.2). The couple (f, g) is called a voting scheme. We first consider general voting schemes but later focus on the voting scheme (f, g) with $f \equiv id$ and $g \equiv 1$.

Let us denote by S_i the multi-set of the sample for a given node i . To define the voting powers of the nodes, we recall the definition of the mean opinion, Equation (1.2),

$$\eta_i = \frac{\sum_{j \in S_i} g(m_j) s_j}{\sum_{j \in S_i} g(m_j)},$$

where s_j is the opinion of node j . The multi-set S_i is random, and taking expectation leads to

$$\mathbb{E}[\eta_i] = \mathbb{E} \left[\frac{\sum_{j \in \mathcal{N}} \frac{g(m_j) A_k(j) s_j}{\sum_{\ell \in \mathcal{N}} g(m_\ell) A_k(\ell)}}{\sum_{j \in \mathcal{N}} \frac{g(m_j) A_k(j) s_j}{\sum_{\ell \in \mathcal{N}} g(m_\ell) A_k(\ell)}} \right].$$

Hence, the influence of the node j on another node's mean opinion is measured by the corresponding coefficient in the above series.

Definition 1. (*Voting power.*) The voting power of a node $i \in \mathcal{N}$ is defined as

$$V_k(i) := \mathbb{E} \left[\frac{g(m_i) A_k(i)}{\sum_{\ell \in \mathcal{N}} g(m_\ell) A_k(\ell)} \right].$$

If $g \equiv 1$, the voting power reduces to

$$V_k(i) = V_k^{(P)}(i) = \mathbb{E} \left[\frac{A_k(i)}{v_k} \right].$$

Definition 2. (*r-splitting.*) Let $(m_i)_{i \in \mathcal{N}}$ be the weight distribution of the nodes. We fix some node $i \in \mathcal{N}$ and $r \in \mathbb{N}$. We say that $m_{i_1^{(r)}}, \dots, m_{i_r^{(r)}} > 0$ is an r -splitting of node i if $m_i = \sum_{j=1}^r m_{i_j^{(r)}}$. With an r -splitting of node i , the probability distribution $P = (p_i)_{i \in \mathcal{N}}$ given in (1.1) changes to the probability distribution $\widehat{P}_{r,i}$ on $\mathcal{N} \setminus \{i\} \cup \{i_1^{(r)}, \dots, i_r^{(r)}\}$ defined by

$$\widehat{p}_j = \frac{f(m_j)}{\sum_{u \in \mathcal{N} \setminus \{i\}} f(m_u) + \sum_{u=1}^r f(m_{i_u^{(r)}})}, \quad j \neq i,$$

$$\widehat{p}_{i_j^{(r)}} = \frac{f(m_{i_j^{(r)}})}{\sum_{u \in \mathcal{N} \setminus \{i\}} f(m_u) + \sum_{u=1}^r f(m_{i_u^{(r)}})}, \quad j \in \{1, 2, \dots, r\}.$$

Definition 3. (*Fairness.*) (i) We say that a voting scheme (f, g) is robust to splitting into r nodes if for all nodes $i \in \mathcal{N}$ and all r -splittings $m_{i_1^{(r)}}, \dots, m_{i_r^{(r)}}$ we have

$$V_k^{(P)}(i) \geq \sum_{j=1}^r V_k^{(\widehat{P}_{r,i})}(i_j^{(r)}). \tag{2.3}$$

(ii) We say that a voting scheme (f, g) is robust to merging of r nodes if for all nodes $i \in \mathcal{N}$ and all r -splittings $m_{i_1}^{(r)}, \dots, m_{i_r}^{(r)}$ we have

$$V_k^{(P)}(i) \leq \sum_{j=1}^r V_k^{(\widehat{P}_{r,i})}(i_j^{(r)}). \tag{2.4}$$

If the relation (2.3) holds for every $r \in \mathbb{N}$, we say that the voting scheme (f, g) is robust to splitting, and if the relation (2.4) holds for every $r \in \mathbb{N}$, we say that the voting scheme (f, g) is robust to merging. If a voting scheme (f, g) is both robust to splitting and robust to merging, that is, if for every node $i \in \mathcal{N}$, every $r \in \mathbb{N}$, and every r -splitting $m_{i_1}^{(r)}, \dots, m_{i_r}^{(r)} > 0$ it holds that

$$V_k^{(P)}(i) = \sum_{j=1}^r V_k^{(\widehat{P}_{r,i})}(i_j^{(r)}),$$

then we say that the voting scheme (f, g) is fair.

The existence of fair voting schemes for greedy sampling is an open question; see also Question 5. In Section 4 we show that the natural voting scheme $(id, 1)$ is not fair for $k = 2$, but the question for $k > 2$ remains open.

To generalize the above definitions to sequences of weights and to define asymptotic fairness, we first define sequences of r -splittings.

Definition 4. (*Sequence of r -splittings.*) Let $k \in \mathbb{N}$ be a positive integer and let $\{(m_i^{(n)})_{i \in \mathcal{N}}\}_{n \in \mathbb{N}}$ be a sequence of weight distributions. For a fixed positive integer $r \in \mathbb{N}$ and a fixed node i , we say that $m_{i_1}^{(n)}, \dots, m_{i_r}^{(n)} > 0$ is a sequence of r -splittings of node r if $m_i^{(n)} = \sum_{j=1}^r m_{i_j}^{(n)}$. We define the sequence of probability distributions $(\widehat{P}_{r,i}^{(n)})_{n \in \mathbb{N}}$ on the set $\mathcal{N} \setminus \{i\} \cup \{i_1^{(r)}, \dots, i_r^{(r)}\}$ by

$$\begin{aligned} \widehat{p}_j^{(n)} &= \frac{f(m_j^{(n)})}{\sum_{u \in \mathcal{N} \setminus \{i\}} f(m_u^{(n)}) + \sum_{u=1}^r f(m_{i_u}^{(n)})}, & j \neq i, \\ \widehat{p}_{i_j}^{(n)} &= \frac{f(m_{i_j}^{(n)})}{\sum_{u \in \mathcal{N} \setminus \{i\}} f(m_u^{(n)}) + \sum_{u=1}^r f(m_{i_u}^{(n)})}, & j \in \{1, 2, \dots, r\}. \end{aligned}$$

Definition 5. (*Asymptotic fairness.*) We say that a voting scheme (f, g) is asymptotically fair for the sequence $\{(m_i^{(n)})_{i \in \mathcal{N}}\}_{n \in \mathbb{N}}$ of weight distributions if, for all $r \in \mathbb{N}$ and all nodes $i \in \mathcal{N}$,

$$\left| \sum_{j=1}^r V_k^{(\widehat{P}_{r,i}^{(n)})}(i_j^{(r)}) - V_k^{(P^{(n)})}(i) \right| \xrightarrow{n \rightarrow \infty} 0$$

for all sequences of r -splittings of node i .

Remark 1. The canonical class of examples of the sequence $\{(m_i^{(n)})_{i \in \mathcal{N}}\}_{n \in \mathbb{N}}$ of weight distributions is the one with $\mathcal{N} = \mathbb{N}$ and $m_i^{(n)} = 0$ for all $i > N_n$ for some strictly increasing sequence $(N_n)_{n \in \mathbb{N}}$. With weight distributions of this type, we can model the scenario where the number of nodes in the network grows to infinity.

2.2. Zipf’s law

We do not assume any particular weight distribution in our theoretical results. However, for examples and numerical simulations, it is essential to consider specific weight distributions. Probably the most appropriate modelings of weight distributions rely on universality phenomena. The most famous example of this universality phenomenon is the central limit theorem. While the central limit theorem is suited to describing statistics where values are of the same order of magnitude, it is not appropriate for modeling more heterogeneous situations where the values might differ by several orders of magnitude. A heterogeneous weight distribution may instead be described by a Zipf law. Zipf’s law was first observed in quantitative linguistics, stating that any word’s frequency is inversely proportional to its rank in the corresponding frequency table. Nowadays, many fields claim that specific data fit a Zipf law; examples include city populations, internet traffic data, the formation of peer-to-peer communities, company sizes, and scientific citations. We refer to [9] for a brief introduction and more references, and to [1] for the appearance of Zipf’s law in the internet and computer networks. We also refer to [17] for a more mathematical introduction to this topic.

There is a rule of thumb for situations when a Zipf law may govern the asymptotic distribution of data or statistics: variables

- (1) take values as positive numbers,
- (2) range over many different orders of magnitude,
- (3) arise from a complicated combination of largely independent factors, and
- (4) have not been artificially rounded, truncated, or otherwise constrained in size.

We consider a situation with N elements or nodes and set $\mathcal{N} = \{1, \dots, N\}$. Zipf’s law predicts that the (normalized) frequency of the node of rank k is given by

$$y(k) := \frac{k^{-s}}{\sum_{i=1}^n i^{-s}}, \tag{2.5}$$

where $s \in [0, \infty)$ is the Zipf parameter. Since the value $y(k)$ in (2.5) only depends on two parameters, s and n , this provides a convenient model for investigating the performance of a voting protocol in a wide range of network situations. For instance, nodes with equal weight can be modeled by choosing $s = 0$, while more centralized networks can be described with parameters $s > 1$.

A convenient way to observe a Zipf law is by plotting the data on a log–log graph, with the axes being log(rank order) and log(value). The data conform to a Zipf law to the extent that the plot is linear, and the value of s may be estimated using linear regression. We note that this visual inspection of the log–log plot of the ranked data is not a rigorous procedure. We refer to the literature on how to detect systematic modulation of the basic Zipf law and on how to fit more accurate models. In this work, we deal with distributions that are ‘Zipf-like’ without verifying certain test conditions.

For instance, Figure 1 shows the distribution of IOTA for the top 10,000 richest addresses with a fitted Zipf law. Based on the universality phenomenon, the plausibility of the hypotheses (1)–(4) above, and Figure 1, we assume the weight distribution to follow a Zipf law if we want to specify a weight distribution. To be more precise, we assume that for every $n \in \mathbb{N}$ and some parameter $s > 0$,

$$p_j^{(n)} := \begin{cases} \frac{1/j^s}{\sum_{i=1}^n (1/i^s)}, & j \leq n, \\ 0, & j > n, \end{cases} \tag{2.6}$$

where $P^{(n)} = (p_j^{(n)})_{j \in \mathbb{N}}$ is the weight distribution among the nodes in the network when the total number of nodes is n . Notice that, for a fixed j , the sequence $(p_j^{(n)})_{n \in \mathbb{N}}$ is decreasing in n . Furthermore, since $\sum_{i=1}^{\infty} (1/i^s)$ diverges for $s \leq 1$, the sequence $(p_j^{(n)})_{n \in \mathbb{N}}$ converges to 0 in this case (when n goes to infinity). On the other hand, if the parameter s is strictly larger than 1, the sequence $(p_j^{(n)})_{n \in \mathbb{N}}$ converges to a positive number (when n goes to infinity).

3. Greedy weighted sampling

In this section, we study fundamental properties of greedy sampling that are necessary for the rest of the paper. In particular, we identify the joint probability distribution of sample size and occurrences of the nodes and develop several related asymptotic results. Recall that in greedy sampling, we consider sampling with replacement until k different elements have been chosen. The actual size of the sample is described by the random variable v_k .

Proposition 1. *Let $P = (p_i)_{i \in \mathcal{N}}$ be a probability distribution on \mathcal{N} , $k \in \mathbb{N}$ a positive integer, and $v_k = v_k^{(P)}$ the random variable defined in (2.1). For every $v \in \{k, k + 1, k + 2, \dots\}$ we have*

$$\mathbb{P}(v_k = v) = \sum_{i \in \mathcal{N}} p_i \sum_{\substack{x_1 + \dots + x_{k-1} = v-1 \\ x_1, \dots, x_{k-1} \geq 1}} \binom{v-1}{x_1, \dots, x_{k-1}} \sum_{\substack{A \subset \mathcal{N} \setminus \{i\} \\ |A|=k-1}} (p_{a_1})^{x_1} \dots (p_{a_{k-1}})^{x_{k-1}}, \quad (3.1)$$

where

$$\binom{v-1}{x_1, x_2, \dots, x_{k-1}} = \begin{cases} \frac{(v-1)!}{x_1! x_2! \dots x_{k-1}!}, & x_1 + x_2 + \dots + x_{k-1} = v-1, \\ 0, & \text{otherwise.} \end{cases} \quad (3.2)$$

Proof. We are sampling from the distribution P until we have sampled k different nodes. A first observation is that the last node will be sampled only once. Any of the nodes that appear before the last one can be sampled more than once. We can construct such a sampling in the following way: first we choose a node $i \in \mathcal{N}$ that will be sampled last, then we choose $k - 1$ different nodes a_1, a_2, \dots, a_{k-1} from the set $\mathcal{N} \setminus \{i\}$ that will appear in the sequence before the last node, and we choose positive integers $x_1, x_2, \dots, x_{k-1} \in \mathbb{N}$ that represent how many times each of the $k - 1$ nodes from the set $\{a_1, a_2, \dots, a_{k-1}\}$ will appear in the sampled sequence. Notice that $\sum_{i=1}^{k-1} x_i$ has to be equal to $v - 1$, because the total length of the sequence, including the last node i , has to be v . The last thing we need to choose is the permutation of the first $v - 1$ elements in the sequence, which can be done in $\binom{v-1}{x_1, x_2, \dots, x_{k-1}}$ ways. Summarizing, the probability of sampling a sequence where the last node is i , the first $k - 1$ nodes are a_1, a_2, \dots, a_{k-1} , and they appear x_1, x_2, \dots, x_{k-1} times is

$$p_i \binom{v}{x_1, x_2, \dots, x_{k-1}} (p_{a_1})^{x_1} (p_{a_2})^{x_2} \dots (p_{a_{k-1}})^{x_{k-1}}.$$

Now we need to sum this up with respect to all the possible values of the element i , all the possible sequences of $k - 1$ positive integers x_1, x_2, \dots, x_{k-1} that sum up to $v - 1$ (i.e., all the partitions of the integer $v - 1$ into $k - 1$ parts), and all the subsets of $\mathcal{N} \setminus \{i\}$ of cardinality $k - 1$. This gives us exactly the expression from Equation (3.1). \square

Remark 2. The random variable $v_k^{(P)}$ was studied in [16] in the case where the population is finite and elements have equal weight. Therefore, (3.1) is a generalization of [16, Formula (16)].

Another random variable studied in [16] is the number of different elements in a sample with replacement of a fixed size. To be precise, let $k \in \mathbb{N}$ be a positive integer and let $P = (p_i)_{i \in \mathcal{N}}$ be a probability distribution on \mathcal{N} . Let

$$u_k^{(P)} = \text{the number of different nodes sampled in } k \text{ samplings with replacement from the distribution } P.$$

The authors of [16] calculated the distribution of the random variable $u_k^{(P)}$, but again under the assumptions that the set from which the elements are sampled is finite and that all the elements are sampled with the same probability. Using analogous reasoning as in the proof of Proposition 1, for $u \in \{1, 2, \dots, k\}$ we get

$$\mathbb{P}(u_k^{(P)} = u) = \sum_{\substack{x_1 + \dots + x_u = k \\ x_1, \dots, x_u \geq 1}} \binom{k}{x_1, \dots, x_u} \sum_{\substack{A \subset \mathcal{N} \\ |A|=u}} (p_{a_1})^{x_1} \dots (p_{a_u})^{x_u}.$$

This formula generalizes [16, Formula (8)].

Using Proposition 1, we now find the distribution of the random vector $(A_k(i), v_k)$ for all $i \in \mathbb{N}$.

Proposition 2. *Let $P = (p_i)_{i \in \mathcal{N}}$ be a probability distribution on \mathcal{N} , $k \in \mathbb{N}$ a positive integer, $v_k = v_k^{(P)}$ the random variable defined in (2.1), and $A_k(i) = A_k^{(P)}(i)$ the random variable defined in (2.2). The support of the random vector $(A_k(i), v_k)$ is*

$$\{(0, v) : v \geq k\} \cup \{(\ell, v) : 1 \leq \ell \leq v - k + 1\}.$$

For every node $i \in \mathcal{N}$ and every (ℓ, v) in the support of $(A_k(i), v_k)$ we have

$$\mathbb{P}(A_k(i) = \ell, v_k = v) = \begin{cases} \sum_{\substack{j \in \mathcal{N} \\ j \neq i}} p_j \sum_{\substack{x_1 + \dots + x_{k-1} = v-1 \\ x_1, \dots, x_{k-1} \geq 1}} \binom{v-1}{x_1, \dots, x_{k-1}} \sum_{\substack{A \subset \mathcal{N} \setminus \{i, j\} \\ |A|=k-1}} \prod_{r=1}^{k-1} (p_{a_r})^{x_r}, & \ell = 0, \\ \sum_{\substack{j \in \mathcal{N} \\ j \neq i}} p_j \left(\sum_{\substack{x_1 + \dots + x_{k-2} = v-2 \\ x_1, \dots, x_{k-2} \geq 1}} \binom{v-1}{x_1, \dots, x_{k-2}, 1} p_i \sum_{\substack{A \subset \mathcal{N} \setminus \{i, j\} \\ |A|=k-2}} \prod_{r=1}^{k-2} (p_{a_r})^{x_r} \right) + \\ p_i \sum_{\substack{x_1 + \dots + x_{k-1} = v-1 \\ x_1, \dots, x_{k-1} \geq 1}} \binom{v-1}{x_1, \dots, x_{k-1}} \sum_{\substack{A \subset \mathcal{N} \setminus \{i\} \\ |A|=k-1}} \prod_{r=1}^{k-1} (p_{a_r})^{x_r}, & \ell = 1, \\ \sum_{\substack{j \in \mathcal{N} \\ j \neq i}} p_j \sum_{\substack{x_1 + \dots + x_{k-2} = v-\ell-1 \\ x_1, \dots, x_{k-2} \geq 1}} \binom{v-1}{x_1, \dots, x_{k-2}, \ell} (p_i)^\ell \sum_{\substack{A \subset \mathcal{N} \setminus \{i, j\} \\ |A|=k-2}} \prod_{r=1}^{k-2} (p_{a_r})^{x_r}, & \ell \geq 2. \end{cases} \tag{3.3}$$

Proof. Notice first that $(0, v)$ is in the support of $(A_k(i), v_k)$. Now, if $\ell \geq 1$, then for all $v < \ell + k - 1$ we have that $\mathbb{P}(A_k(i) = \ell, v_k = v) = 0$, since we need at least $\ell + k - 1$ samplings to sample node i ℓ times and to sample the other $k - 1$ different nodes at least once.

Let us consider separately different values of the non-negative integer $\ell \in \mathbb{N} \cup \{0\}$.

$\ell = 0$: This case is an immediate consequence of Proposition 1. We just need to restrict the set of all nodes that can be sampled to $\mathcal{N} \setminus \{i\}$.

$\ell = 1$: Here we need to distinguish two disjoint scenarios. The first is when the node i is not sampled as the last node (i.e., the node i is not the k th different node that has been sampled). This means that the node i was sampled during the first $v - 1$ samplings. Hence, we first choose a node $j \in \mathcal{N} \setminus \{i\}$ that will be sampled last. Then we choose $k - 2$ different nodes a_1, a_2, \dots, a_{k-2} from the set $\mathcal{N} \setminus \{i, j\}$ that will appear (together with the node i) in the sampled sequence before the last node, and we choose positive integers $x_1, x_2, \dots, x_{k-2} \in \mathbb{N}$ to represent how many times each of the $k - 2$ nodes in the set $\{a_1, a_2, \dots, a_{k-2}\}$ will appear in the sampled sequence. Notice that $\sum_{i=1}^{k-2} x_i$ has to be equal to $v - 2$, because the total length of the sequence, including the one appearance of node j (in the last place) and one appearance of node i (somewhere in the first $v - 1$ samplings), has to be v . The last thing we need to choose is the permutation of the first $v - 1$ nodes in the sequence, which can be done in $\binom{v-1}{x_1, \dots, x_{k-2}, 1}$ ways (taking into consideration that node i appears only once). Summarizing, the probability of sampling a sequence where the last node is $j \neq i$, the node i appears exactly once in the first $v - 1$ sampled nodes, and the remaining $k - 2$ nodes that appear together with the node i before the last node j are a_1, a_2, \dots, a_{k-2} and appear x_1, x_2, \dots, x_{k-2} times is

$$p_j \binom{v-1}{x_1, \dots, x_{k-2}, 1} p_i \prod_{r=1}^{k-2} (p_{a_r})^{x_r}.$$

As in Proposition 1, we now sum this up with respect to all the possible values of the node j , all the possible sequences of $k - 2$ positive integers x_1, x_2, \dots, x_{k-2} that sum up to $v - 2$, and all the subsets of $\mathcal{N} \setminus \{i, j\}$ of cardinality $k - 2$. In this way we obtain the first term in the expression for $\mathbb{P}(A_k(i) = 1, v_k = v)$.

The second scenario is the one where the node i is sampled last. Here the situation is much simpler. The last node is fixed to be $i \in \mathcal{N}$; we then choose the $k - 1$ nodes that appear before the node i , and the number of times they appear, analogously as in Proposition 1. We immediately get the second term in the expression for $\mathbb{P}(A_k(i) = 1, v_k = v)$.

$\ell \geq 2$: Notice that in this case we do not have two different scenarios, because it is impossible for the node i to be the last node sampled. As we explained in Proposition 1, the last node can be sampled only once, since we terminate sampling when we have reached k different nodes. Now we reason analogously as in the first scenario of the case $\ell = 1$. The only difference is that here the node i appears ℓ times (in the first $v - 1$ samplings), so the integers x_1, x_2, \dots, x_{k-2} have to satisfy $\sum_{i=1}^{k-2} x_i = v - \ell - 1$. Together with ℓ appearances of the node i and one appearance of the last node, this gives v sampled nodes in total. \square

Let $b = (b_i)_{i \in \mathcal{N}} \subset \mathbb{R}$ be a sequence of real numbers. Denote by $\|b\|_\infty = \sup_{i \in \mathcal{N}} |b_i|$ the supremum norm of the sequence b . The next result shows that if the probabilities of sampling each of the nodes converge uniformly to zero, then the number of samplings needed to sample k different elements converges to k .

Lemma 1. *Let $(P^{(n)})_{n \in \mathbb{N}}, P^{(n)} = (p_i^{(n)})_{i \in \mathcal{N}}$, be a sequence of probability distributions on \mathcal{N} , and let $(k_n)_{n \in \mathbb{N}}$ be a sequence of positive integers such that $k_n^2 \|P^{(n)}\|_\infty \xrightarrow{n \rightarrow \infty} 0$. Then*

$$v_{k_n}^{(P^{(n)})} - k_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}} 0.$$

In particular, if for some fixed positive $k \in \mathbb{N}$ we have $k_n = k$ for all $n \in \mathbb{N}$, and $\|P^{(n)}\|_\infty \xrightarrow{n \rightarrow \infty} 0$, then we have that $v_k^{(P^{(n)})} \xrightarrow[n \rightarrow \infty]{\mathbb{P}} k$.

Proof. For simplicity, we define $v_{k_n}^{(n)} := v_{k_n}^{(P^{(n)})}$. Since $v_{k_n}^{(n)}$ is larger than or equal to k_n , it is sufficient to show that $\mathbb{P}(v_{k_n}^{(n)} > k_n) \xrightarrow{n \rightarrow \infty} 0$. Denote by $X_i^{(n)}$ the random variable representing the node sampled in the i th sampling. Since the event $\{v_{k_n}^{(n)} > k_n\}$ happens if and only if some of the nodes sampled in the first k_n samplings appear more than once, we have

$$\begin{aligned} &\mathbb{P}(v_{k_n}^{(n)} > k_n) \\ &= \mathbb{P}(X_1^{(n)} = X_2^{(n)}) + \mathbb{P}(X_1^{(n)} \neq X_2^{(n)}, X_3^{(n)} \in \{X_1^{(n)}, X_2^{(n)}\}) + \dots \\ &\quad \dots + \mathbb{P}(X_1^{(n)} \neq X_2^{(n)}, X_3^{(n)} \notin \{X_1^{(n)}, X_2^{(n)}\}, \dots, X_{k_n}^{(n)} \in \{X_1^{(n)}, X_2^{(n)}, \dots, X_{k_n-1}^{(n)}\}) \\ &= \sum_{i_1 \in \mathcal{N}} (p_{i_1}^{(n)})^2 + \sum_{i_1 \in \mathcal{N}} \sum_{\substack{i_2 \in \mathcal{N} \\ i_2 \neq i_1}} p_{i_1}^{(n)} p_{i_2}^{(n)} (p_{i_1}^{(n)} + p_{i_2}^{(n)}) \\ &\quad + \dots + \sum_{i_1 \in \mathcal{N}} \sum_{\substack{i_2 \in \mathcal{N} \\ i_2 \neq i_1}} \dots \sum_{\substack{i_{k_n-1} \in \mathcal{N} \\ \dots \\ i_{k_n-1} \neq i_1}} p_{i_1}^{(n)} p_{i_2}^{(n)} \dots p_{i_{k_n-1}}^{(n)} \sum_{j=1}^{k_n-1} p_{i_j}^{(n)} \\ &\leq \|P^{(n)}\|_\infty \sum_{i_1 \in \mathcal{N}} p_{i_1}^{(n)} + 2\|P^{(n)}\|_\infty \sum_{i_1 \in \mathcal{N}} \sum_{i_2 \in \mathcal{N}} p_{i_1}^{(n)} p_{i_2}^{(n)} \\ &\quad + \dots + (k_n - 1)\|P^{(n)}\|_\infty \sum_{i_1 \in \mathcal{N}} \sum_{i_2 \in \mathcal{N}} \dots \sum_{i_{k_n-1} \in \mathcal{N}} p_{i_1}^{(n)} p_{i_2}^{(n)} \dots p_{i_{k_n-1}}^{(n)} \\ &= \|P^{(n)}\|_\infty (1 + 2 + \dots + k_n - 1) \leq k_n^2 \|P^{(n)}\|_\infty. \end{aligned}$$

By the assumption, the last term converges to zero when n goes to infinity, which is exactly what we wanted to prove. □

Remark 3. Let us investigate what happens when the sequence $(P^{(n)})_{n \in \mathbb{N}}$ is defined by a Zipf law (see (2.6)) with parameter $s > 0$. Since each of the sequences $(p_i^{(n)})_{i \in \mathbb{N}}$ is decreasing in i , we have

$$\|P^{(n)}\|_\infty = p_1^{(n)} = \frac{1}{\sum_{i=1}^n \frac{1}{i^s}}.$$

Notice that for all $s \leq 1$ we have $\|P^{(n)}\|_\infty \xrightarrow{n \rightarrow \infty} 0$, because the series $\sum_{i=1}^\infty \frac{1}{i^s}$ diverges for those values of the parameter s . Hence, for a fixed integer $k \in \mathbb{N}$, we have that $v_k^{(P^{(n)})} \xrightarrow[n \rightarrow \infty]{\mathbb{P}} k$ whenever $s \leq 1$. Another important example is when the sequence $(k_n)_{n \in \mathbb{N}}$ is given by

$$k_n = \lfloor \log(n) \rfloor,$$

where, for $x \in \mathbb{R}$, $\lfloor x \rfloor$ is the largest integer less than or equal to x . Using

$$\sum_{i=1}^n \frac{1}{i^s} \sim \begin{cases} n^{1-s}, & s < 1, \\ \log(n), & s = 1, \end{cases}$$

we get $k_n^2 \|P^{(n)}\|_\infty \xrightarrow{n \rightarrow \infty} 0$ for $s < 1$, so we can apply Lemma 1 for this particular choice of sequences $(k_n)_{n \in \mathbb{N}}$ and $(P^{(n)})_{n \in \mathbb{N}}$.

In Lemma 1 we dealt with the behavior of the sequence of random variables $(v_{k_n}^{P^{(n)}})_{n \in \mathbb{N}}$ if the sequence $(P^{(n)})_{n \in \mathbb{N}}$ satisfies $\|P^{(n)}\|_\infty \xrightarrow{n \rightarrow \infty} 0$. Next, we study the case when the sequence $(P^{(n)})_{n \in \mathbb{N}}$ converges in the supremum norm to another probability distribution $P^{(\infty)}$ on \mathcal{N} . As before, for $b = (b_i)_{i \in \mathcal{N}} \subset \mathbb{R}$, we use the notation $\|b\|_\infty = \sup_{i \in \mathcal{N}} |b_i|$ and we write $\|b\|_1 = \sum_{i \in \mathcal{N}} |b_i|$.

Proposition 3. *Let $(P^{(n)})_{n \in \mathbb{N}}$, $P^{(n)} = (p_i^{(n)})_{i \in \mathcal{N}}$, be a sequence of probability distributions on \mathcal{N} , and let $P^{(\infty)} = (p_i^{(\infty)})_{i \in \mathcal{N}}$ be a probability distribution on \mathcal{N} . If*

$$\|P^{(n)} - P^{(\infty)}\|_\infty = \sup_{i \in \mathcal{N}} |p_i^{(n)} - p_i^{(\infty)}| \xrightarrow{n \rightarrow \infty} 0,$$

then, for all $i \in \mathcal{N}$ and all fixed $k \in \mathbb{N}$,

$$(A_k^{(P^{(n)})}(i), v_k^{(P^{(n)})}) \xrightarrow[n \rightarrow \infty]{(d)} (A_k^{(P^{(\infty)})}(i), v_k^{(P^{(\infty)})}),$$

where $\xrightarrow{(d)}$ denotes convergence in distribution.

Proof. For simplicity, we write $v_k^{(n)} := v_k^{(P^{(n)})}$, $v_k^{(\infty)} := v_k^{(P^{(\infty)})}$, $A_k^{(n)}(i) := A_k^{(P^{(n)})}(i)$, and $A_k^{(\infty)}(i) := A_k^{(P^{(\infty)})}(i)$. Since we consider discrete random variables, the statement

$$(A_k^{(n)}(i), v_k^{(n)}) \xrightarrow[n \rightarrow \infty]{(d)} (A_k^{(\infty)}(i), v_k^{(\infty)})$$

is equivalent to

$$\mathbb{P}(A_k^{(n)}(i) = \ell, v_k^{(n)} = v) \xrightarrow{n \rightarrow \infty} \mathbb{P}(A_k^{(\infty)}(i) = \ell, v_k^{(\infty)} = v)$$

for all $\ell \in \mathbb{N} \cup \{0\}$ and all $v \in \mathbb{N}$. As in the proof of Proposition 2, we consider separately different values of the non-negative integer $\ell \in \mathbb{N} \cup \{0\}$.

$\ell = 0$: Using Proposition 2, we have

$$\begin{aligned} & |\mathbb{P}(A_k^{(n)}(i) = 0, v_k^{(n)} = v) - \mathbb{P}(A_k^{(\infty)}(i) = 0, v_k^{(\infty)} = v)| \\ &= \left| \sum_{\substack{j \in \mathcal{N} \\ j \neq i}} p_j^{(n)} \sum_{\substack{x_1 + \dots + x_{k-1} = v-1 \\ x_1, \dots, x_{k-1} \geq 1}} \binom{v-1}{x_1, \dots, x_{k-1}} \sum_{\substack{A \subset \mathcal{N} \setminus \{i, j\} \\ |A| = k-1}} (p_{a_1}^{(n)})^{x_1} \dots (p_{a_{k-1}}^{(n)})^{x_{k-1}} \right. \\ &\quad \left. - \sum_{\substack{j \in \mathcal{N} \\ j \neq i}} p_j^{(\infty)} \sum_{\substack{x_1 + \dots + x_{k-1} = v-1 \\ x_1, \dots, x_{k-1} \geq 1}} \binom{v-1}{x_1, \dots, x_{k-1}} \sum_{\substack{A \subset \mathcal{N} \setminus \{i, j\} \\ |A| = k-1}} (p_{a_1}^{(\infty)})^{x_1} \dots (p_{a_{k-1}}^{(\infty)})^{x_{k-1}} \right| \\ &\leq \sum_{\substack{x_1 + \dots + x_{k-1} = v-1 \\ x_1, \dots, x_{k-1} \geq 1}} \binom{v-1}{x_1, \dots, x_{k-1}} \left| \left(\sum_{\substack{j \in \mathcal{N} \\ j \neq i}} p_j^{(n)} \sum_{\substack{A \subset \mathcal{N} \setminus \{i, j\} \\ |A| = k-1}} (p_{a_1}^{(n)})^{x_1} \dots (p_{a_{k-1}}^{(n)})^{x_{k-1}} \right) \right. \\ &\quad \left. - \left(\sum_{\substack{j \in \mathcal{N} \\ j \neq i}} p_j^{(\infty)} \sum_{\substack{A \subset \mathcal{N} \setminus \{i, j\} \\ |A| = k-1}} (p_{a_1}^{(\infty)})^{x_1} \dots (p_{a_{k-1}}^{(\infty)})^{x_{k-1}} \right) \right| \\ &= \sum_{\substack{x_1 + \dots + x_{k-1} = v-1 \\ x_1, \dots, x_{k-1} \geq 1}} \binom{v-1}{x_1, \dots, x_{k-1}} |I^{(n)}(x_1, \dots, x_{k-1}) - I^{(\infty)}(x_1, \dots, x_{k-1})|, \end{aligned}$$

with

$$I^{(n)}(x_1, \dots, x_{k-1}) := \sum_{\substack{j \in \mathcal{N} \\ j \neq i}} p_j^{(n)} \sum_{\substack{A \subset \mathcal{N} \setminus \{i,j\} \\ |A|=k-1}} (p_{a_1}^{(n)})^{x_1} \dots (p_{a_{k-1}}^{(n)})^{x_{k-1}},$$

$$I^{(\infty)}(x_1, \dots, x_{k-1}) := \sum_{\substack{j \in \mathcal{N} \\ j \neq i}} p_j^{(\infty)} \sum_{\substack{A \subset \mathcal{N} \setminus \{i,j\} \\ |A|=k-1}} (p_{a_1}^{(\infty)})^{x_1} \dots (p_{a_{k-1}}^{(\infty)})^{x_{k-1}}.$$

It remains to prove that $|I^{(n)}(x_1, \dots, x_{k-1}) - I^{(\infty)}(x_1, \dots, x_{k-1})| \xrightarrow{n \rightarrow \infty} 0$, uniformly for all possible values of positive integers x_1, x_2, \dots, x_{k-1} . This is sufficient since the number of partitions of the integer $v - 1$ into $k - 1$ parts is finite and independent of n . Notice that for every $i, j \in \mathcal{N}$,

$$\sum_{\substack{A \subset \mathcal{N} \setminus \{i,j\} \\ |A|=k-1}} (p_{a_1}^{(\infty)})^{x_1} \dots (p_{a_{k-1}}^{(\infty)})^{x_{k-1}} \leq \sum_{\substack{A \subset \mathcal{N} \\ |A|=k-1}} p_{a_1}^{(\infty)} \dots p_{a_{k-1}}^{(\infty)} \leq \sum_{a_1 \in \mathcal{N}} p_{a_1}^{(\infty)} \dots \sum_{a_{k-1} \in \mathcal{N}} p_{a_{k-1}}^{(\infty)} = 1. \tag{3.4}$$

Clearly the same is true when, instead of the distribution $P^{(\infty)}$, we consider the distribution $P^{(n)}$. Thanks to the convergence of these series, we can rewrite

$$I^{(n)}(x_1, \dots, x_{k-1}) - I^{(\infty)}(x_1, \dots, x_{k-1}) = \left(\sum_{\substack{j \in \mathcal{N} \\ j \neq i}} (p_j^{(n)} - p_j^{(\infty)}) \sum_{\substack{A \subset \mathcal{N} \setminus \{i,j\} \\ |A|=k-1}} (p_{a_1}^{(n)})^{x_1} \dots (p_{a_{k-1}}^{(n)})^{x_{k-1}} \right) + \left(\sum_{\substack{j \in \mathcal{N} \\ j \neq i}} p_j^{(\infty)} \sum_{\substack{A \subset \mathcal{N} \setminus \{i,j\} \\ |A|=k-1}} \left((p_{a_1}^{(n)})^{x_1} \dots (p_{a_{k-1}}^{(n)})^{x_{k-1}} - (p_{a_1}^{(\infty)})^{x_1} \dots (p_{a_{k-1}}^{(\infty)})^{x_{k-1}} \right) \right) =: T_1^{(n)}(x_1, \dots, x_{k-1}) + T_2^{(n)}(x_1, \dots, x_{k-1}).$$

For simplicity, we write $T_1^{(n)} = T_1^{(n)}(x_1, \dots, x_{k-1})$ and $T_2^{(n)} = T_2^{(n)}(x_1, \dots, x_{k-1})$. It remains to prove that $T_1^{(n)}$ and $T_2^{(n)}$ converge to 0 when n goes to infinity. Using the inequality (3.4) and Proposition 5 we have that

$$|T_1^{(n)}| \leq \sum_{j \in \mathcal{N}} |p_j^{(n)} - p_j^{(\infty)}| \cdot 1 \xrightarrow{n \rightarrow \infty} 0.$$

To treat the term $T_2^{(n)}$, we use Lemma 2 in the second line and Proposition 5 in the last line to obtain

$$|T_2^{(n)}| \leq \sum_{\substack{j \in \mathcal{N} \\ j \neq i}} p_j^{(\infty)} \sum_{\substack{A \subset \mathcal{N} \setminus \{i,j\} \\ |A|=k-1}} \left| (p_{a_1}^{(n)})^{x_1} \dots (p_{a_{k-1}}^{(n)})^{x_{k-1}} - (p_{a_1}^{(\infty)})^{x_1} \dots (p_{a_{k-1}}^{(\infty)})^{x_{k-1}} \right| \leq \sum_{j \in \mathcal{N}} p_j^{(\infty)} \sum_{\substack{A \subset \mathcal{N} \\ |A|=k-1}} \left| \sum_{r=1}^{k-1} (p_{a_1}^{(n)})^{x_1} \dots (p_{a_{r-1}}^{(n)})^{x_{r-1}} \left((p_{a_r}^{(n)})^{x_r} - (p_{a_r}^{(\infty)})^{x_r} \right) \cdot (p_{a_{r+1}}^{(\infty)})^{x_{r+1}} \dots (p_{a_{k-1}}^{(\infty)})^{x_{k-1}} \right|$$

$$\begin{aligned}
 &\leq \sum_{r=1}^{k-1} \sum_{a_1 \in \mathcal{N}} \dots \sum_{a_{k-1} \in \mathcal{N}} (p_{a_1}^{(n)})^{x_1} \dots (p_{a_{r-1}}^{(n)})^{x_{r-1}} \left| (p_{a_r}^{(n)})^{x_r} - (p_{a_r}^{(\infty)})^{x_r} \right| \\
 &\qquad \qquad \qquad \cdot (p_{a_{r+1}}^{(\infty)})^{x_{r+1}} \dots (p_{a_{k-1}}^{(\infty)})^{x_{k-1}} \\
 &\leq \sum_{r=1}^{k-1} \sum_{a_r \in \mathcal{N}} \left| (p_{a_r}^{(n)})^{x_r} - (p_{a_r}^{(\infty)})^{x_r} \right| \\
 &= \sum_{r=1}^{k-1} \sum_{a_r \in \mathcal{N}} \left| p_{a_r}^{(n)} - p_{a_r}^{(\infty)} \right| \left| (p_{a_r}^{(n)})^{x_r-1} + (p_{a_r}^{(n)})^{x_r-2} (p_{a_r}^{(\infty)}) + \dots + (p_{a_r}^{(\infty)})^{x_r-1} \right| \\
 &\leq (k-1)x_r \sum_{j \in \mathcal{N}} \left| p_j^{(n)} - p_j^{(\infty)} \right| \leq vk \sum_{j \in \mathcal{N}} \left| p_j^{(n)} - p_j^{(\infty)} \right| \xrightarrow{n \rightarrow \infty} 0.
 \end{aligned}$$

$\ell \geq 2$: Again using Proposition 2, we have

$$\begin{aligned}
 &|\mathbb{P}(A_k^{(n)}(i) = \ell, v_k^{(n)} = v) - \mathbb{P}(A_k^{(\infty)}(i) = \ell, v_k^{(\infty)} = v)| \\
 &= \left| \sum_{\substack{j \in \mathcal{N} \\ j \neq i}} p_j^{(n)} \sum_{\substack{x_1 + \dots + x_{k-2} = v - \ell - 1 \\ x_1, \dots, x_{k-2} \geq 1}} \binom{v-1}{x_1, \dots, x_{k-2}, \ell} (p_i^{(n)})^\ell \sum_{\substack{A \subset \mathcal{N} \setminus \{i, j\} \\ |A| = k-2}} (p_{a_1}^{(n)})^{x_1} \dots (p_{a_{k-2}}^{(n)})^{x_{k-2}} \right. \\
 &\quad \left. - \sum_{\substack{j \in \mathcal{N} \\ j \neq i}} p_j^{(\infty)} \sum_{\substack{x_1 + \dots + x_{k-2} = v - \ell - 1 \\ x_1, \dots, x_{k-2} \geq 1}} \binom{v-1}{x_1, \dots, x_{k-2}, \ell} (p_i^{(\infty)})^\ell \sum_{\substack{A \subset \mathcal{N} \setminus \{i, j\} \\ |A| = k-2}} (p_{a_1}^{(\infty)})^{x_1} \dots (p_{a_{k-2}}^{(\infty)})^{x_{k-2}} \right| \\
 &\leq \sum_{\substack{x_1 + \dots + x_{k-2} = v - \ell - 1 \\ x_1, \dots, x_{k-2} \geq 1}} \binom{v-1}{x_1, \dots, x_{k-2}, \ell} \times \\
 &\quad \left| \sum_{\substack{j \in \mathcal{N} \\ j \neq i}} (p_j^{(n)} (p_i^{(n)})^\ell - p_j^{(\infty)} (p_i^{(\infty)})^\ell) \sum_{\substack{A \subset \mathcal{N} \setminus \{i, j\} \\ |A| = k-2}} (p_{a_1}^{(n)})^{x_1} \dots (p_{a_{k-2}}^{(n)})^{x_{k-2}} \right. \\
 &\quad \left. + \sum_{\substack{j \in \mathcal{N} \\ j \neq i}} p_j^{(\infty)} (p_i^{(\infty)})^\ell \sum_{\substack{A \subset \mathcal{N} \setminus \{i, j\} \\ |A| = k-2}} \left((p_{a_1}^{(n)})^{x_1} \dots (p_{a_{k-2}}^{(n)})^{x_{k-2}} - (p_{a_1}^{(\infty)})^{x_1} \dots (p_{a_{k-2}}^{(\infty)})^{x_{k-2}} \right) \right|.
 \end{aligned}$$

To show that the above expression converges to zero as n tends to infinity, it remains to verify that

$$\sum_{\substack{j \in \mathcal{N} \\ j \neq i}} |p_j^{(n)} (p_i^{(n)})^\ell - p_j^{(\infty)} (p_i^{(\infty)})^\ell| \xrightarrow{n \rightarrow \infty} 0.$$

To obtain this, we can use the same arguments as in the previous case. Again, introducing a middle term leads to

$$\begin{aligned}
 &\sum_{\substack{j \in \mathcal{N} \\ j \neq i}} |p_j^{(n)} (p_i^{(n)})^\ell - p_j^{(\infty)} (p_i^{(\infty)})^\ell| \\
 &= \sum_{\substack{j \in \mathcal{N} \\ j \neq i}} |p_j^{(n)} (p_i^{(n)})^\ell - p_j^{(n)} (p_i^{(\infty)})^\ell + p_j^{(n)} (p_i^{(\infty)})^\ell - p_j^{(\infty)} (p_i^{(\infty)})^\ell|
 \end{aligned}$$

$$\begin{aligned} &\leq |(p_i^{(n)})^\ell - (p_i^{(\infty)})^\ell| \sum_{j \in \mathcal{N}} p_j^{(n)} + (p_i^{(\infty)})^\ell \sum_{j \in \mathcal{N}} |p_j^{(n)} - p_j^{(\infty)}| \\ &\leq |p_i^{(n)} - p_i^{(\infty)}| (p_i^{(n)})^{\ell-1} + (p_i^{(n)})^{\ell-2} p_i^{(\infty)} + \dots + (p_i^{(\infty)})^{\ell-1} + \sum_{j \in \mathcal{N}} |p_j^{(n)} - p_j^{(\infty)}| \\ &\leq \ell \cdot \|P^{(n)} - P^{(\infty)}\|_\infty + \|P^{(n)} - P^{(\infty)}\|_1. \end{aligned}$$

Applying Proposition 5 again, we get the desired result.

$\ell = 1$: The difference of the first terms in the expressions for $\mathbb{P}(A_k^{(n)}(i) = 1, v_k^{(n)} = v)$ and $\mathbb{P}(A_k^{(\infty)}(i) = 1, v_k^{(\infty)} = v)$ (see (3.3)) goes to zero. The difference of the second terms can be handled similarly as in the case $\ell = 0$; the situation is even simpler owing to the absence of the initial sum. This concludes the proof of this proposition. \square

Corollary 1. Let $(P^{(n)})_{n \in \mathbb{N}}, P^{(n)} = (p_i^{(n)})_{i \in \mathcal{N}}$, be a sequence of probability distributions on \mathcal{N} , and let $P^{(\infty)} = (p_i^{(\infty)})_{i \in \mathcal{N}}$ be a probability distribution on \mathcal{N} . We assume that $g \equiv 1$. If

$$\|P^{(n)} - P^{(\infty)}\|_\infty = \sup_{i \in \mathcal{N}} |p_i^{(n)} - p_i^{(\infty)}| \xrightarrow{n \rightarrow \infty} 0,$$

then for all $i \in \mathcal{N}$ and all fixed $k \in \mathbb{N}$,

$$A_k^{(P^{(n)})}(i) \xrightarrow[n \rightarrow \infty]{(d)} A_k^{(P^{(\infty)})}(i), \tag{3.5}$$

$$v_k^{(P^{(n)})} \xrightarrow[n \rightarrow \infty]{(d)} v_k^{(P^{(\infty)})}, \tag{3.6}$$

$$V_k^{(P^{(n)})}(i) \xrightarrow[n \rightarrow \infty]{} V_k^{(P^{(\infty)})}(i), \tag{3.7}$$

where $V_k^{(P^{(n)})}(i) = \mathbb{E}[A_k^{(P^{(n)})}(i)/v_k^{(P^{(n)})}]$, $n \in \mathbb{N} \cup \{\infty\}$, is the voting power of the node i in the case $g \equiv 1$.

Proof. Convergence in (3.5) and (3.6) follows directly from Proposition 3 using the continuous mapping theorem (see [5, Theorem 3.2.4]) applied to the projections $\Pi_1, \Pi_2 : \mathbb{R}^2 \rightarrow \mathbb{R}$, $\Pi_i(x_1, x_2) = x_i, i = 1, 2$. To prove the convergence in (3.7), let us first define the bounded and continuous function

$$\Phi : [0, \infty) \times [1, \infty) \rightarrow \mathbb{R}, \quad \Phi(x, y) = \min \left\{ \frac{x}{y}, 1 \right\}.$$

Then, combining Proposition 3 with [5, Theorem 3.2.3], we get

$$\mathbb{E} \left[\Phi \left(A_k^{(P^{(n)})}(i), v_k^{(P^{(n)})} \right) \right] \xrightarrow[n \rightarrow \infty]{} \mathbb{E} \left[\Phi \left(A_k^{(P^{(\infty)})}(i), v_k^{(P^{(\infty)})} \right) \right]. \tag{3.8}$$

Notice that we always have $A_k^{(P)}(i) \leq v_k^{(P)}$, since the random variable $A_k^{(P)}(i)$ counts the number of times the node i was sampled until k different nodes had been sampled, and the random variable $v_k^{(P)}$ counts the total number of samplings until k distinct elements had been sampled. Hence,

$$\Phi \left(A_k^{(P)}(i), v_k^{(P)} \right) = \frac{A_k^{(P)}(i)}{v_k^{(P)}}.$$

Combining the latter with (3.8) and using $V_k^{(P^{(n)})}(i) = \mathbb{E}[A_k^{(P^{(n)})}(i)/v_k^{(P^{(n)})}]$, $n \in \mathbb{N} \cup \{\infty\}$, we obtain (3.7). \square

4. Asymptotic fairness

We start this section with the case $k = 2$, i.e., we sample until we get two different nodes. This small choice of k allows us to perform analytical calculations and prove some facts rigorously. We prove that the voting scheme $(id, 1)$ is robust to merging but not fair. We also show that the more the node splits, the more voting power it can gain. However, with this procedure, the voting power does not grow to 1, but has a limit strictly less than 1.

Proposition 4. *We consider the voting scheme $(id, 1)$ and let $(m_i)_{i \in \mathcal{N}}$ be the weight distribution of the nodes. Let $P = (p_i)_{i \in \mathcal{N}}$ be the corresponding probability distribution on \mathcal{N} , let $r \in \mathbb{N}$, let $i \in \mathcal{N}$ be a node, and let $k = 2$. Then, for every r -splitting $m_{i_1^{(r)}}, \dots, m_{i_r^{(r)}} > 0$ of the node i , we have that*

$$V_k^{(P)}(i) < \sum_{j=1}^r V_k^{(\widehat{P}_{r,i})}(i_j^{(r)}). \tag{4.1}$$

In other words, the voting scheme $(id, 1)$ is robust to merging, but not robust to splitting. The difference of the voting power after splitting and before splitting reaches its maximum for

$$m_{i_j^{(r)}} = \frac{m_i}{r}, \quad j \in \{1, 2, \dots, r\}.$$

Furthermore, for this particular r -splitting, we have that the sequence

$$\left(\left(\sum_{j=1}^r V_k^{(\widehat{P}_{r,i})}(i_j^{(r)}) \right) - V_k^{(P)}(i) \right)_{r \in \mathbb{N}}$$

is strictly increasing and has a limit strictly less than 1.

Proof. Denote by $Y_i = A_2^{(P)}(i)$ the number of times that the node i was sampled from the distribution P until we had sampled 2 different nodes, and by $Y_{i_j} = A_2^{(\widehat{P}_{r,i})}(i_j^{(r)}), j \in \{1, 2, \dots, r\}$, the number of times the node $i_j^{(r)}$ was sampled from the distribution $\widehat{P}_{r,i}$ until we had sampled 2 different nodes. We also write $V(i) := V_k^{(P)}(i)$ and $V^{(r)}(i_j^{(r)}) := V_k^{(\widehat{P}_{r,i})}(i_j^{(r)})$. Using this notation, we have

$$\begin{aligned} V(i) &= \sum_{u \in \mathcal{N} \setminus \{i\}} \sum_{y_u=1}^{\infty} \mathbb{P}(Y_i = 1, Y_u = y_u) \cdot \frac{1}{1 + y_u} + \sum_{u \in \mathcal{N} \setminus \{i\}} \sum_{y_i=1}^{\infty} \mathbb{P}(Y_i = y_i, Y_u = 1) \cdot \frac{y_i}{1 + y_i} \\ &= p_i \sum_{u \in \mathcal{N} \setminus \{i\}} \sum_{y_u=1}^{\infty} p_u^{y_u} \cdot \frac{1}{1 + y_u} + \sum_{u \in \mathcal{N} \setminus \{i\}} p_u \sum_{y_i=1}^{\infty} p_i^{y_i} \cdot \frac{y_i}{1 + y_i} \\ &= p_i \sum_{u \in \mathcal{N} \setminus \{i\}} \left(\frac{-\log(1 - p_u)}{p_u} - 1 \right) + (1 - p_i) \cdot \frac{\frac{p_i}{1 - p_i} + \log(1 - p_i)}{p_i} \\ &= -p_i \sum_{u \in \mathcal{N} \setminus \{i\}} \left(\frac{\log(1 - p_u)}{p_u} + 1 \right) + \frac{(1 - p_i) \log(1 - p_i)}{p_i} + 1. \end{aligned}$$

Similarly, for $j \in \{1, 2, \dots, r\}$ we have

$$\begin{aligned}
 V^{(r)}(i_j^{(r)}) &= \sum_{u \in \mathcal{N} \setminus \{i\}} \sum_{y_u=1}^{\infty} \mathbb{P}(Y_{i_j} = 1, Y_u = y_u) \cdot \frac{1}{1 + y_u} \\
 &\quad + \sum_{\substack{\ell=1 \\ \ell \neq j}}^r \sum_{y_{i_\ell}=1}^{\infty} \mathbb{P}(Y_{i_j} = 1, Y_{i_\ell} = y_{i_\ell}) \cdot \frac{1}{1 + y_{i_\ell}} \\
 &\quad + \sum_{u \in \mathcal{N} \setminus \{i\}} \sum_{y_{i_j}=1}^{\infty} \mathbb{P}(Y_{i_j} = y_{i_j}, Y_u = 1) \cdot \frac{y_{i_j}}{1 + y_{i_j}} \\
 &\quad + \sum_{\substack{\ell=1 \\ \ell \neq j}}^r \sum_{y_{i_j}=1}^{\infty} \mathbb{P}(Y_{i_j} = y_{i_j}, Y_{i_\ell} = 1) \cdot \frac{y_{i_j}}{1 + y_{i_j}} \\
 &= -\widehat{p}_{i_j^{(r)}} \sum_{u \in \mathcal{N} \setminus \{i\}} \left(\frac{\log(1 - p_u)}{p_u} + 1 \right) - \widehat{p}_{i_j^{(r)}} \sum_{\substack{\ell=1 \\ \ell \neq j}}^r \left(\frac{\log(1 - \widehat{p}_{i_\ell^{(r)}})}{\widehat{p}_{i_\ell^{(r)}}} + 1 \right) \\
 &\quad + \frac{(1 - \widehat{p}_{i_j^{(r)}}) \log(1 - \widehat{p}_{i_j^{(r)}})}{\widehat{p}_{i_j^{(r)}}} + 1.
 \end{aligned}$$

Combining the above calculations, we obtain

$$\begin{aligned}
 &\sum_{j=1}^r V^{(r)}(i_j^{(r)}) - V(i) \\
 &= - \sum_{j=1}^r p_{i_j^{(r)}} \sum_{u \in \mathcal{N} \setminus \{i\}} \left(\frac{\log(1 - p_u)}{p_u} + 1 \right) - \sum_{j=1}^r \sum_{\substack{\ell=1 \\ \ell \neq j}}^r \frac{\widehat{p}_{i_j^{(r)}} \log(1 - \widehat{p}_{i_\ell^{(r)}})}{\widehat{p}_{i_\ell^{(r)}}} \\
 &\quad - \sum_{j=1}^r p_{i_j^{(r)}}(r - 1) + \sum_{j=1}^r \frac{(1 - \widehat{p}_{i_j^{(r)}}) \log(1 - \widehat{p}_{i_j^{(r)}})}{\widehat{p}_{i_j^{(r)}}} + r \\
 &\quad + p_i \sum_{u \in \mathcal{N} \setminus \{i\}} \left(\frac{\log(1 - p_u)}{p_u} + 1 \right) - \frac{(1 - p_i) \log(1 - p_i)}{p_i} - 1 \\
 &= \sum_{j=1}^r \frac{\log(1 - \widehat{p}_{i_j^{(r)}})}{\widehat{p}_{i_j^{(r)}}} \left(1 - \widehat{p}_{i_j^{(r)}} - \sum_{\substack{\ell=1 \\ \ell \neq j}}^r \widehat{p}_{i_\ell^{(r)}} \right) + r - 1 - p_i(r - 1) \\
 &\quad - \frac{(1 - p_i) \log(1 - p_i)}{p_i} \\
 &= (1 - p_i) \left[(r - 1) + \sum_{j=1}^r \frac{\log(1 - \widehat{p}_{i_j^{(r)}})}{\widehat{p}_{i_j^{(r)}}} - \frac{\log(1 - p_i)}{p_i} \right].
 \end{aligned}$$

We take $x_1, x_2, \dots, x_r \in (0, 1)$ such that $\sum_{j=1}^r x_j = 1$ and set

$$\widehat{p}_{i_j^{(r)}} = p_i \cdot x_j, \quad j \in \{1, \dots, r\}.$$

This gives us

$$\sum_{j=1}^r V^{(r)}(i_j^{(r)}) - V(i) = (1 - p_i) \left[(r - 1) + \sum_{j=1}^r \frac{\log(1 - p_i x_j)}{p_i x_j} - \frac{\log(1 - p_i)}{p_i} \right].$$

Define

$$\phi(x_1, \dots, x_r) := (r - 1) + \sum_{j=1}^r \frac{\log(1 - p_i x_j)}{p_i x_j} - \frac{\log(1 - p_i)}{p_i}.$$

First we need to show that $\phi(x_1, x_2, \dots, x_r) > 0$ for all $x_1, x_2, \dots, x_r \in (0, 1)$ such that $\sum_{j=1}^r x_j = 1$. Using Proposition 6 repeatedly ($r - 1$ times), we get

$$\begin{aligned} \phi(x_1, \dots, x_r) &= (r - 2) + \left(1 + \frac{\log(1 - p_i x_1)}{p_i x_1} + \frac{\log(1 - p_i x_2)}{p_i x_2} \right) \\ &\quad + \sum_{j=3}^r \frac{\log(1 - p_i x_j)}{p_i x_j} - \frac{\log(1 - p_i)}{p_i} \\ &> (r - 2) + \frac{\log(1 - p_i(x_1 + x_2))}{p_i(x_1 + x_2)} + \sum_{j=3}^r \frac{\log(1 - p_i x_j)}{p_i x_j} - \frac{\log(1 - p_i)}{p_i} \\ &= (r - 3) + \left(1 + \frac{\log(1 - p_i(x_1 + x_2))}{p_i(x_1 + x_2)} + \frac{\log(1 - p_i x_3)}{p_i x_3} \right) \\ &\quad + \sum_{j=4}^r \frac{\log(1 - p_i x_j)}{p_i x_j} - \frac{\log(1 - p_i)}{p_i} \\ &> (r - 3) + \frac{\log(1 - p_i(x_1 + x_2 + x_3))}{p_i(x_1 + x_2 + x_3)} + \sum_{j=4}^r \frac{\log(1 - p_i x_j)}{p_i x_j} - \frac{\log(1 - p_i)}{p_i} \\ &\quad \vdots \qquad \qquad \qquad \vdots \qquad \qquad \qquad \vdots \\ &> 1 + \frac{\log(1 - p_i \sum_{j=1}^{r-1} x_j)}{p_i \sum_{j=1}^{r-1} x_j} + \frac{\log(1 - p_i x_r)}{p_i x_r} - \frac{\log(1 - p_i)}{p_i} > 0. \end{aligned}$$

The second claim of this proposition is that the expression

$$\sum_{j=1}^r V^{(r)}(i_j^{(r)}) - V(i)$$

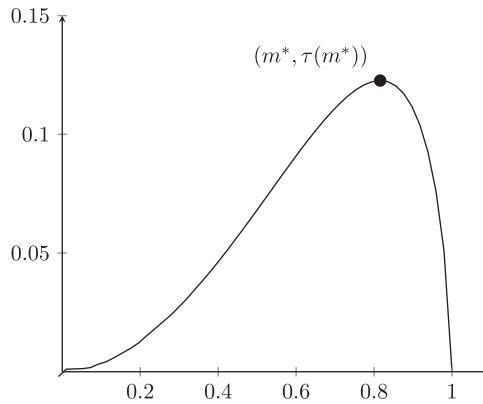


FIGURE 2. Graph of the function τ .

reaches its maximum for $\widehat{p}_{i_j}^{(r)} = \frac{p_i}{r}, j \in \{1, 2, \dots, r\}$. This follows directly from Lemma 3, where we show that ϕ attains its unique maximum for $(x_1, \dots, x_r) = (\frac{1}{r}, \dots, \frac{1}{r})$. Define

$$\begin{aligned} \tau_r(p_i) &= (1 - p_i)\phi\left(\frac{1}{r}, \dots, \frac{1}{r}\right) = (1 - p_i) \left[(r - 1) + \sum_{j=1}^r \frac{\log(1 - \frac{p_i}{r})}{\frac{p_i}{r}} - \frac{\log(1 - p_i)}{p_i} \right] \\ &= (1 - p_i) \left[r + \frac{r^2 \log(1 - \frac{p_i}{r})}{p_i} - \frac{\log(1 - p_i)}{p_i} - 1 \right]. \end{aligned}$$

By Proposition 7 we have that the sequence $(\tau_r(p_i))_{r \in \mathbb{N}}$ is strictly increasing and

$$\tau(p_i) = \lim_{r \rightarrow \infty} \tau_r(p_i) = (1 - p_i) \left(-\frac{p_i}{2} - \frac{\log(1 - p_i)}{p_i} - 1 \right). \quad \square$$

Remark 4. We consider the function $\tau : (0, 1) \rightarrow \mathbb{R}$ defined by

$$\tau(m) = (1 - m) \left(-\frac{m}{2} - \frac{\log(1 - m)}{m} - 1 \right).$$

This function describes the gain in voting power that a node with initial weight m can achieve by splitting up into infinitely many nodes. As Figure 2 shows, this maximal gain in voting power is bounded. The function τ attains its maximum at $m^* \approx 0.82$, and the maximum is $\tau(m^*) \approx 0.12$. This means that a node that initially has around 82% of the total amount of weight can obtain the biggest gain in voting power (by theoretically splitting into an infinite number of nodes), and this gain is approximately 0.12. Loosely speaking, if a voting power of a node increases by 0.12, this means that during the querying, the proportion of queries that are addressed to this particular node increases by around 12%.

Corollary 2. Let $(m^{(n)})_{n \in \mathbb{N}}$ be a sequence of weight distributions with corresponding probability distributions $(P^{(n)})_{n \in \mathbb{N}}, P^{(n)} = (p_i^{(n)})_{i \in \mathcal{N}}$, on \mathcal{N} . Let $m^{(\infty)}$ be a weight distribution such that for its corresponding probability distribution $P^{(\infty)} = (p_i^{(\infty)})_{i \in \mathcal{N}}$ we have that

$$\|P^{(n)} - P^{(\infty)}\|_{\infty} = \sup_{i \in \mathcal{N}} |p_i^{(n)} - p_i^{(\infty)}| \xrightarrow{n \rightarrow \infty} 0.$$

Furthermore, we consider a sequence of r -splittings $m_{i_1}^{(n)}, \dots, m_{i_r}^{(n)} > 0$ of a node i such that $m_{i_j}^{(n)} \xrightarrow{n \rightarrow \infty} m_{i_j}^{(\infty)}$, $j \in \{1, 2, \dots, r\}$, for some r -splitting $m^{(\infty)}$. Then for $k = 2$ we have

$$\begin{aligned} \lim_{n \rightarrow \infty} & \left(\left(\sum_{j=1}^r V_k^{\widehat{P}_{r,i}^{(n)}}(i_j^{(r)}) \right) - V_k^{P^{(n)}}(i) \right) \\ & = \left(\sum_{j=1}^r V_k^{\widehat{P}_{r,i}^{(\infty)}}(i_j^{(r)}) \right) - V_k^{P^{(\infty)}}(i) > 0. \end{aligned}$$

Proof. The convergence follows directly from Corollary 1, and the strict positivity of the limit follows from Proposition 4. □

Remark 5. Corollary 2 implies that if $k = 2$ and if the sequence of weight distributions $(P^{(n)})_{n \in \mathbb{N}}$ converges to a non-trivial probability distribution on \mathcal{N} , the voting scheme $(id, 1)$ is not asymptotically fair. Applying this result to the sequence of Zipf distributions defined in (2.6), we see that for $s > 1$ and $k = 2$, the voting scheme is not asymptotically fair. Simulations suggest (see Figures 4 and 8) that for higher values of k , the difference in voting power of the node i before and after the splitting does not converge to zero as the number of nodes in the network grows to infinity.

In the following proposition we give a condition on the sequence of weight distributions $(P^{(n)})_{n \in \mathbb{N}}$ under which the voting scheme $(id, 1)$ is asymptotically fair for any choice of the parameter k .

Theorem 1. Let k be a positive integer, and let $(m^{(n)})_{n \in \mathbb{N}}$ be a sequence of weight distributions with corresponding probability distributions $(P^{(n)})_{n \in \mathbb{N}}$, $P^{(n)} = (p_i^{(n)})_{i \in \mathcal{N}}$, on \mathcal{N} . We assume that $\|P^{(n)}\|_\infty \xrightarrow{n \rightarrow \infty} 0$. Furthermore, we consider a sequence of r -splittings $m_{i_1}^{(n)}, \dots, m_{i_r}^{(n)} > 0$ of a given node i such that

$$m_{i_j}^{(n)} \xrightarrow{n \rightarrow \infty} m_{i_j}^{(\infty)}, \quad j \in \{1, 2, \dots, r\},$$

for some r -splitting $m^{(\infty)}$. Then

$$\left| \left(\sum_{j=1}^r V_k^{\widehat{P}_{r,i}^{(n)}}(i_j^{(r)}) \right) - V_k^{P^{(n)}}(i) \right| \xrightarrow{n \rightarrow \infty} 0;$$

i.e., the voting scheme $(id, 1)$ is asymptotically fair if the sequence of weight distributions converges in the supremum norm to 0.

Proof. For simplicity, we write $P_r^{(n)} := \widehat{P}_{r,i}^{(n)}$, $v_k^{(n)} := v_k^{P^{(n)}}$, and $v_{k,r}^{(n)} := v_k^{P_r^{(n)}}$; recall that the random variable $v_k^{(P)}$ counts the number of samplings with replacement from the distribution P until k different elements have been sampled. The main idea of the proof is to couple the random variables $v_k^{(n)}$ and $v_{k,r}^{(n)}$. We sample simultaneously from the probability distributions $P^{(n)}$ and $P_r^{(n)}$ and construct two different sequences of elements that both terminate once they contain k different elements. We do that in the following way: we sample an element from the distribution $P^{(n)}$. If the sampled element is not i , we just add this element to both sequences that we are constructing and then sample the next element. If the element i is sampled, then we

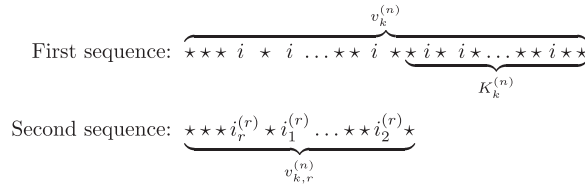


FIGURE 3. Coupling of random variables $v_k^{(n)}$ and $v_{k,r}^{(n)}$.

add i to the first sequence, but to the second sequence we add one of the elements $i_1^{(r)}, \dots, i_r^{(r)}$ according to the probability distribution $(P_{i_1^{(r)}}^{(n)}/P_i^{(n)}, \dots, P_{i_r^{(r)}}^{(n)}/P_i^{(n)})$. Now, the second sequence will terminate no later than the first one, since the second sequence always has at least as many different elements as the first sequence. This is a consequence of the fact that, each time the element i is sampled, we add one of the r elements $i_1^{(r)}, i_2^{(r)}, \dots, i_r^{(r)}$ to the second sequence, while we just add i to the first sequence; see Figure 3.

Define

$$K_k^{(n)} := v_k^{(n)} - v_{k,r}^{(n)}$$

Since $v_k^{(n)} \geq v_{k,r}^{(n)}$, we have $K_k^{(n)} \geq 0$. We also introduce the random variable

$$L_k^{(n)} := A_k^{(P^{(n)})}(i) - \sum_{j=1}^r A_k^{(P_j^{(n)})}(i_j^{(r)}),$$

where $A_k^{(P)}(i)$ is defined as in (2.2). The random variable $L_k^{(n)}$ measures the difference between the number of times the node i appears in the first sequence and the number of times the nodes $i_1^{(r)}, i_2^{(r)}, \dots, i_r^{(r)}$ appear in the second sequence. At the time when the second sequence terminates, the number of times the node i has appeared in the first sequence is the same as the number of times that the nodes $i_1^{(r)}, i_2^{(r)}, \dots, i_r^{(r)}$ have appeared in the second sequence; see Figure 3. Since the length of the first sequence is always larger than or equal to the length of the second sequence, it can happen that the element i is sampled again before the k th different element appears in the first sequence. Therefore, $L_k^{(n)} \geq 0$. Clearly, $L_k^{(n)} \leq K_k^{(n)}$, because $K_k^{(n)}$ counts all the extra samplings we need to sample k different elements in the first sequence, while $L_k^{(n)}$ counts only those extra samplings in which the element i was sampled. Notice that if the element i is not sampled before the k th different element appears, or if i itself is the k th different element, then $K_k^{(n)} = L_k^{(n)} = 0$.

Let

$$Y_k^{(n)} := A_k^{(P^{(n)})}(i) \quad \text{and} \quad Y_{k,r}^{(n)} := \sum_{j=1}^r A_k^{(P_j^{(n)})}(i_j^{(r)}).$$

Then

$$V_k^{(P^{(n)})}(i) = \mathbb{E} \left[\frac{Y_k^{(n)}}{v_k^{(n)}} \right], \quad \sum_{j=1}^r V_k^{(P_j^{(n)})}(i_j^{(r)}) = \mathbb{E} \left[\frac{Y_{k,r}^{(n)}}{v_{k,r}^{(n)}} \right] \quad \text{and} \quad Y_k^{(n)} = Y_{k,r}^{(n)} + L_k^{(n)}.$$

Combining this with $v_k^{(n)} = v_{k,r}^{(n)} + K_k^{(n)}$, we have

$$\begin{aligned} & \left| \left(\sum_{j=1}^r V_k^{(P_r^{(n)})}(i_j^{(r)}) \right) - V_k^{(P^{(n)})}(i) \right| = \left| \mathbb{E} \left[\frac{Y_{k,r}^{(n)}}{v_{k,r}^{(n)}} - \frac{Y_k^{(n)}}{v_k^{(n)}} \right] \right| \\ &= \left| \mathbb{E} \left[\frac{Y_{k,r}^{(n)}}{v_{k,r}^{(n)}} - \frac{Y_{k,r}^{(n)} + L_k^{(n)}}{v_{k,r}^{(n)} + K_k^{(n)}} \right] \right| \\ &= \left| \mathbb{E} \left[\frac{Y_{k,r}^{(n)}(v_{k,r}^{(n)} + K_k^{(n)}) - (Y_{k,r}^{(n)} + L_k^{(n)})v_{k,r}^{(n)}}{v_{k,r}^{(n)}(v_{k,r}^{(n)} + K_k^{(n)})} \right] \right| \\ &= \left| \mathbb{E} \left[\frac{Y_{k,r}^{(n)}K_k^{(n)}}{v_{k,r}^{(n)}(v_{k,r}^{(n)} + K_k^{(n)})} - \frac{L_k^{(n)}}{v_{k,r}^{(n)} + K_k^{(n)}} \right] \right| \\ &\leq \mathbb{E} \left[\frac{Y_{k,r}^{(n)}}{v_{k,r}^{(n)}} \cdot \frac{1}{v_{k,r}^{(n)} + K_k^{(n)}} \cdot K_k^{(n)} + \frac{1}{v_{k,r}^{(n)} + K_k^{(n)}} \cdot L_k^{(n)} \right]. \end{aligned}$$

Define

$$Z_n := \frac{Y_{k,r}^{(n)}}{v_{k,r}^{(n)}} \cdot \frac{1}{v_{k,r}^{(n)} + K_k^{(n)}} \cdot K_k^{(n)} + \frac{1}{v_{k,r}^{(n)} + K_k^{(n)}} \cdot L_k^{(n)}.$$

It remains to prove that $\mathbb{E}[Z_n] \xrightarrow{n \rightarrow \infty} 0$. Since $Y_{k,r}^{(n)} \leq v_{k,r}^{(n)}$, $k \leq v_{k,r}^{(n)}$, and $L_k^{(n)} \leq K_k^{(n)}$, we have

$$Z_n \leq \min\{2K_k^{(n)}, 2\}.$$

By Lemma 1 we have that $v_k^{(n)} \xrightarrow{n \rightarrow \infty} k$ and $v_{k,r}^{(n)} \xrightarrow{n \rightarrow \infty} k$ (notice that $\|P_r^{(n)}\|_\infty \leq \|P^{(n)}\|_\infty$). Therefore,

$$K_k^{(n)} = v_k^{(n)} - v_{k,r}^{(n)} \xrightarrow{n \rightarrow \infty} 0.$$

This implies that $Z_n \xrightarrow{n \rightarrow \infty} 0$. Since $Z_n \leq 2$ we have that $\lim_{n \rightarrow \infty} \mathbb{E}[Z_n] = 0$. □

Remark 6. The above proposition shows that $\|P^{(n)}\|_\infty \xrightarrow{n \rightarrow \infty} 0$ is a sufficient condition to ensure asymptotic fairness, regardless of the value of the parameter $k \in \mathbb{N}$. Applying this result to the sequence of probability distributions $(P^{(n)})_{n \in \mathbb{N}}$ defined by the Zipf law (see (2.6)), we see that for $s \leq 1$ the voting scheme $(id, 1)$ is asymptotically fair.

5. Simulations, conjectures, and questions

In this section, we present some numerical simulations to complement our theoretical results. We are interested in the rate of convergence in the case of asymptotic fairness, Theorem 1, and we also want to support some conjectures for the situation where our theoretical results do not apply.

In the simulations, we always consider the voting scheme $(id, 1)$ and a Zipf law for the nodes' weight distribution; see the relation (2.6). The reasons for this assumption are presented in Subsection 2.2.

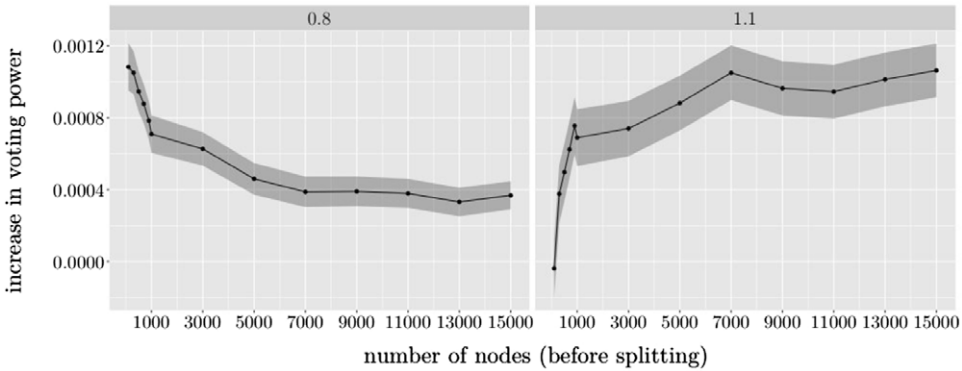


FIGURE 4. Increase in voting power for fixed $k = 20$, varying N , and two different values of s .

Figure 4 presents the results of a Monte Carlo simulation for a Zipf distribution with parameter $s \in \{0.8, 1.1\}$ and different network sizes on the x -axis. For real-world applications we expect values of k to be at least 20 (see also [12]) and therefore set the sample size to $k = 20$. The y -axis shows the gain in voting power when the heaviest node splits into two nodes of equal weight. For each choice of network size, we performed 1 000 000 simulations; we use the empirical average as an estimator for the gain in voting power. The gray zone corresponds to the confidence interval of level 0.95. Let us note that to decrease the variance of the estimation, we couple, as in the proof of Theorem 1, the sampling in the original network with the sampling in the network after splitting.

Theorem 1 and Remark 6 state that if the Zipf parameter $s \leq 1$, then the voting scheme is asymptotically fair, i.e., the difference in the voting power before and after the splitting of a node $i \in \mathbb{N}$ goes to zero as the number of nodes in the network increases. The left panel of Figure 4 indicates the speed of convergence for $s = 0.8$. The right panel of Figure 4 indicates that for $s = 1.1$ the voting scheme is not asymptotically fair. Corollary 2 states that for $k = 2$, if the sequence of weight distributions $(P^{(n)})_{n \in \mathbb{N}}$ converges to a non-trivial probability distribution on \mathbb{N} , then the voting scheme $(id, 1)$ is not asymptotically fair.

Conjecture 1. Let $(m^{(n)})_{n \in \mathbb{N}}$ be a sequence of weight distributions with corresponding probability distributions $(P^{(n)})_{n \in \mathbb{N}}$, $P^{(n)} = (p_i^{(n)})_{i \in \mathcal{N}}$, on \mathcal{N} . Let $m^{(\infty)}$ be a weight distribution such that for its corresponding probability distribution $P^{(\infty)} = (p_i^{(\infty)})_{i \in \mathcal{N}}$ we have that

$$\|P^{(n)} - P^{(\infty)}\|_{\infty} = \sup_{i \in \mathcal{N}} |p_i^{(n)} - p_i^{(\infty)}| \xrightarrow{n \rightarrow \infty} 0.$$

Furthermore, we consider a sequence of r -splittings $m_{i_1^{(r)}}^{(n)}, \dots, m_{i_r^{(r)}}^{(n)} > 0$ of a node i such that

$$m_{i_j^{(r)}}^{(n)} \xrightarrow{n \rightarrow \infty} m_{i_j}^{(\infty)}, \quad j \in \{1, 2, \dots, r\},$$

for some r -splitting $m^{(\infty)}$. Then, for any choice of $k \in \mathbb{N}$,

$$\begin{aligned} \lim_{n \rightarrow \infty} \left(\left(\sum_{j=1}^r V_k^{\widehat{P}_{r,i}^{(n)}}(i_j^{(r)}) \right) - V_k^{P^{(n)}}(i) \right) \\ = \left(\sum_{j=1}^r V_k^{\widehat{P}_{r,i}^{(\infty)}}(i_j^{(r)}) \right) - V_k^{P^{(\infty)}}(i) > 0. \end{aligned}$$

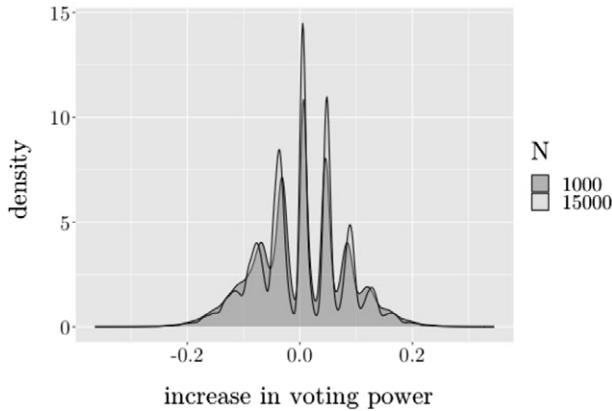


FIGURE 5. Density estimation for increase in voting power for two choices of network size ($k = 20$, $s = 1.1$).

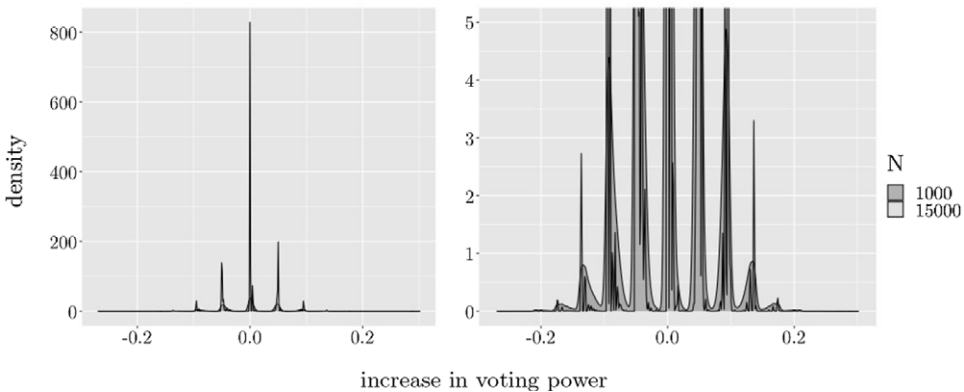


FIGURE 6. Density estimation for increase in voting power for two choices of network size ($k = 20$, $s = 0.8$). The right panel is a zoom of the left panel.

We take a closer look at the distribution of the increase in voting power in the above setting. Figures 5 and 6 present density estimations, with a Gaussian kernel, of the density of the increase in voting power. Again we simulated each data point 1 000 000 times. The density’s multimodality can be explained by the different possibilities for whether the sample includes the heaviest node before and after splitting. For instance, in the case of two nodes and $k = 2$, both nodes will be present in every sample before the splitting. After the heaviest node splits into two nodes, we have two cases: the sample contains either both splittings of the heaviest node or only one splitting. In the first case the heaviest node has gained voting influence, while in the second case it has lost influence.

Figure 6 clearly shows the asymptotic fairness; the probability of having only a small change in voting power converges to 0 as the number of participants grows to infinity. Figure 7 compares the densities for different choices of s in a network of 1000 nodes.

These figures also show that even in the case where a splitting leads to an increase on average of the voting power, the splitting can also lead to less influence in a single voting round.

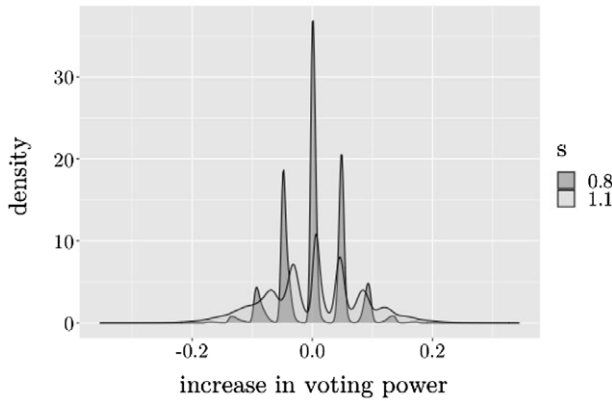


FIGURE 7. Density estimation for increase in voting power for two choices of the Zipf parameter s for a network size of 1000 nodes and $k = 20$.

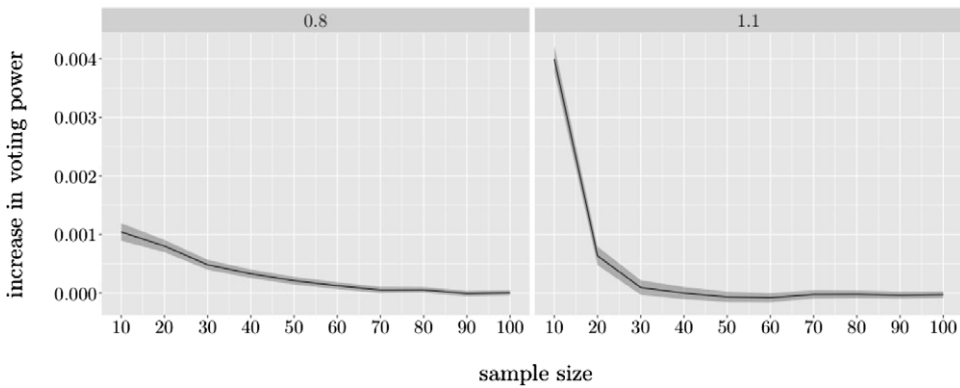


FIGURE 8. Mean increase in voting power for different values of the sample size k , two different values of s , and a network size of 1000 nodes.

In the previous simulations, we kept the sample size $k = 20$. Increasing the sample size increases the quality of the voting, but at the cost of a higher message complexity. Figure 8 compares the increase in voting power for different values of k and s . We can see that increasing k increases the fairness of the voting scheme and that for some values of k the increase in voting power may even be negative.

Figure 9 presents density estimations of the increase in voting power. We can see the different behaviors in the more decentralized setting, $s < 1$, and the centralized setting, $s > 1$. In the first case, it seems that the density converges to a point mass in 0, whereas in the second case the limit may be described by a Gaussian density. Figure 10 provides Q–Q plots supporting this first visual impression.

While the study of the actual distribution of the increase in voting power is outside the scope of this paper, we think that the following questions may be of interest.

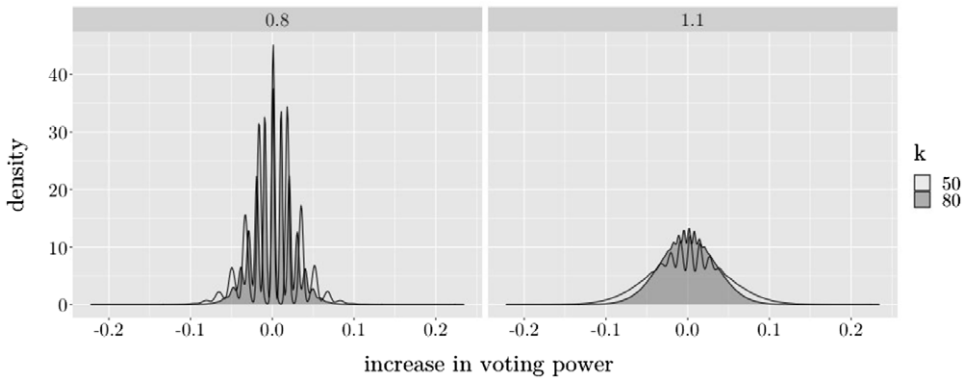


FIGURE 9. Density estimation for increase in voting power for two choices of the Zipf parameter s for a network size of 1000 nodes and two choices of k .

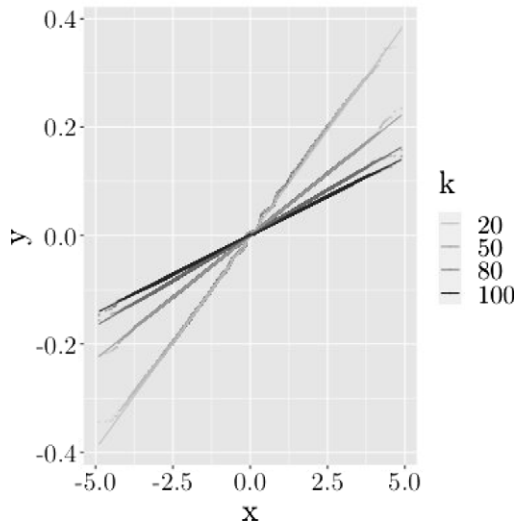


FIGURE 10. Q-Q plots of the increase in voting power against a Gaussian distribution for different choices of k and $s = 1.1$ in a network of size 1000.

Question 1. How can the distribution of the increase in voting power be described?

Question 2. What characteristics of the distribution of the increase in voting power are important for the voting scheme and its applications?

Recall that, so far, we have only considered the change in voting power of the heaviest node when it splits into two nodes of equal weight. The goal of the next two simulations (see Figures 11 and 12) is to inspect what happens to the voting power of a node when it splits into more than two nodes.

For the simulation shown in Figure 11, we fixed the value of the parameter k and varied the value of the parameter s . In Proposition 4 we showed that for $k = 2$, a node always gains voting power with splitting. This result holds without any additional assumptions on the weight distribution of the nodes in the network. We ran simulations with $k = 20$ in which we split the

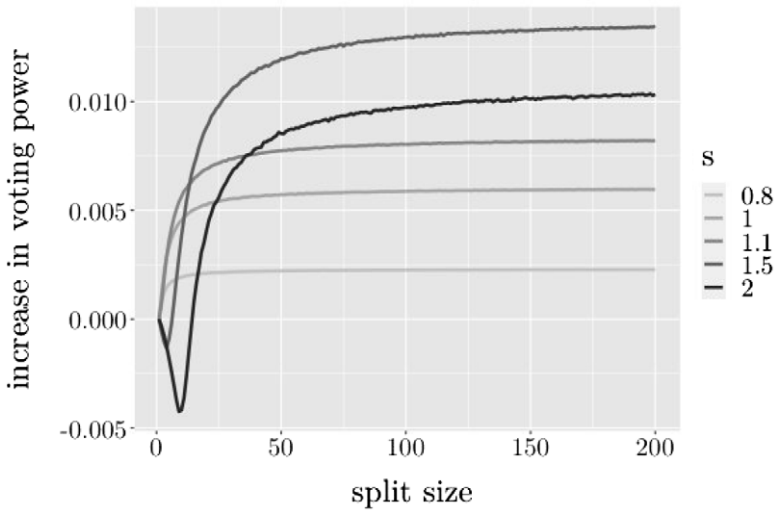


FIGURE 11. The effect of multiple splitting on the voting power of a node for $k = 20$ and different s in a network of size 1000.

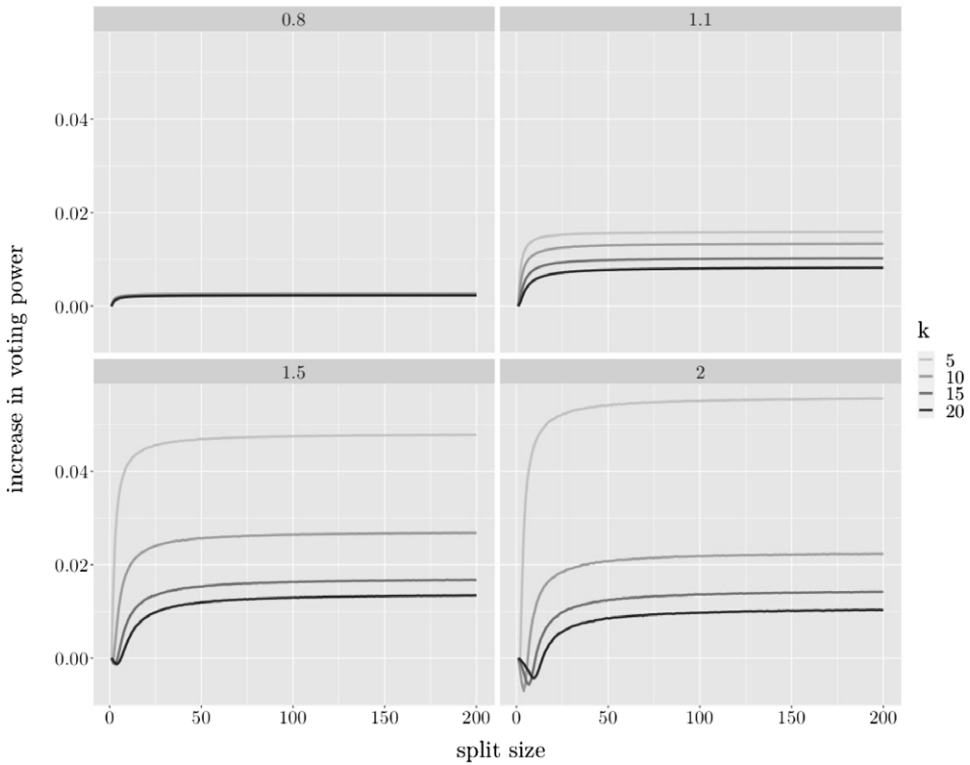


FIGURE 12. The effect of multiple splitting on the voting power of the heaviest node in a network of size 1000 ($s \in \{0.8, 1.1, 1.5, 2\}$, $k \in \{5, 10, 15, 20\}$).

heaviest node into r nodes, r ranging from 2 to 200. We kept the network size equal to 1000 and varied the parameter s in the set $\{0.8, 1, 1.1, 1.5, 2\}$. For each value of s we ran 100 000 simulations of the voting scheme $(id, 1)$. Several conjectures can be made from Figure 11. It seems that if the parameter k is equal to 20, the voting power may even drop for small values of the parameter r . This drop appears to be more significant the bigger the parameter s is. But if we split into more nodes (i.e., we set r to be sufficiently large), it seems that splitting gives us more voting power, and the gain is bigger for values of s larger than 1. This suggests that it is possible to have robustness to splitting into r nodes for r smaller than some threshold δ , and robustness to merging of r nodes for $r > \delta$.

The simulations presented in Figure 12 show the change in the voting power of a node after it splits into multiple nodes for different values of the parameters k and s . As in the previous simulation, we consider a network size of 1000 and assume that the first node splits into r different nodes (where r again ranges from 2 to 200). For each combination of values of parameters k and s , we ran 100 000 simulations. Our results suggest that for $s \leq 1$, we always gain voting power with additional splittings. On the other hand, if $s > 1$, then the behavior of the voting power depends even more on the precise value of k . It seems that for small k , we still cannot lose voting power by splitting, but that for k sufficiently large there is a region where the increase in voting power is negative.

Question 3. How does the increase in voting power of the heaviest node depend on k, s , and N ? For which choices of these parameters is the increase in voting power negative?

The above simulation study is far from complete, but we believe that our results already show the model’s richness.

Question 4. How does the increase in voting power of a node of rank M depends on M, k, s , and N ?

Note that in the simulations above, we only split the heaviest node. In a more realistic model, not only one but all nodes may simultaneously optimize their voting power. This is particularly interesting in situations that are not robust to splitting. We believe that in such a situation, it is reasonable to suppose that nodes may adapt their strategy from time to time to optimize their voting power. This simultaneous splitting or merging of nodes may lead to periodic behavior or to convergence to a stable situation, where none of the nodes has an incentive to split or merge.

Question 5. Construct a multi-player game where the aim is to maximize voting power. Do the corresponding weights always converge to a situation in which the voting scheme is fair?

As mentioned after the definition of a fair voting scheme, Definition 3, the existence of fair schemes is not well understood. While we focused on the particular scheme $(id, 1)$ in this paper, it is unclear whether other voting schemes could be fair or have other advantages over the $(id, 1)$ scheme.

Question 6. Identify all fair voting schemes.

Appendix A. Auxiliary results for Section 3

The first result is proved by induction.

Lemma 2. Let $w \in \mathbb{N}$, and let $a_1, a_2, \dots, a_w, b_1, b_2, \dots, b_w \in \mathbb{R}$; then

$$(a_1 a_2 \cdots a_w) - (b_1 b_2 \cdots b_w) = \sum_{j=1}^w a_1 a_2 \cdots a_{j-1} (a_j - b_j) b_{j+1} \cdots b_w.$$

Proposition 5. Let $(P^{(n)})_{n \in \mathbb{N}}$, $P^{(n)} = (p_i^{(n)})_{i \in \mathbb{N}}$, be a sequence of probability distributions on \mathbb{N} , and let $P^{(\infty)} = (p_i^{(\infty)})_{i \in \mathbb{N}}$ be a probability distribution on \mathbb{N} . Then the following statements are equivalent:

$$(a) \quad \|P^{(n)} - P^{(\infty)}\|_{\infty} = \sup_{i \in \mathbb{N}} |p_i^{(n)} - p_i^{(\infty)}| \xrightarrow{n \rightarrow \infty} 0.$$

$$(b) \quad \|P^{(n)} - P^{(\infty)}\|_1 = \sum_{i=1}^{\infty} |p_i^{(n)} - p_i^{(\infty)}| \xrightarrow{n \rightarrow \infty} 0.$$

Proof. (b) \Rightarrow (a): This follows immediately from the fact that

$$\sup_{i \in \mathbb{N}} |p_i^{(n)} - p_i^{(\infty)}| \leq \sum_{i=1}^{\infty} |p_i^{(n)} - p_i^{(\infty)}|.$$

(a) \Rightarrow (b): Let $\varepsilon > 0$. Choose $n_0 = n_0(\varepsilon)$ such that

$$\sum_{i=1}^{n_0} p_i^{(\infty)} > 1 - \varepsilon. \tag{A.1}$$

This can be done because $P^{(\infty)}$ is a probability distribution on \mathbb{N} . Furthermore, let $n_1 = n_1(\varepsilon, n_0) \in \mathbb{N}$ be such that for every $n \geq n_1$ we have

$$\|P^{(n)} - P^{(\infty)}\|_{\infty} < \frac{\varepsilon}{n_0}. \tag{A.2}$$

Using this, for all $n \geq n_1$ we get

$$\sum_{i=1}^{n_0} |p_i^{(n)} - p_i^{(\infty)}| \leq n_0 \cdot \|P^{(n)} - P^{(\infty)}\|_{\infty} < n_0 \cdot \frac{\varepsilon}{n_0} = \varepsilon. \tag{A.3}$$

On the other hand, we have that for all $n \geq n_1$,

$$\sum_{i=n_0+1}^{\infty} p_i^{(n)} < \varepsilon + \sum_{i=1}^{n_0} (p_i^{(\infty)} - p_i^{(n)}) \leq \varepsilon + \sum_{i=1}^{n_0} |p_i^{(\infty)} - p_i^{(n)}| < 2\varepsilon, \tag{A.4}$$

where in the first inequality we used (A.1) together with the fact that $\sum_{i=1}^{\infty} p_i^{(n)} = 1$, and in the last inequality we used (A.3). Combining (A.4) and (A.1) we get

$$\sum_{i=n_0+1}^{\infty} |p_i^{(n)} - p_i^{(\infty)}| \leq \sum_{i=n_0+1}^{\infty} p_i^{(n)} + \sum_{i=n_0+1}^{\infty} p_i^{(\infty)} < 2\varepsilon + \varepsilon = 3\varepsilon, \tag{A.5}$$

for all $n \geq n_1$. Finally, we have

$$\|P^{(n)} - P^{(\infty)}\|_1 = \sum_{i=1}^{n_0} |p_i^{(n)} - p_i^{(\infty)}| + \sum_{i=n_0+1}^{\infty} |p_i^{(n)} - p_i^{(\infty)}| < \varepsilon + 3\varepsilon = 4\varepsilon,$$

where we used (A.3) and (A.5). This proves that $\|P^{(n)} - P^{(\infty)}\|_1 \xrightarrow{n \rightarrow \infty} 0$ assuming that $\|P^{(n)} - P^{(\infty)}\|_{\infty} \xrightarrow{n \rightarrow \infty} 0$, which is exactly what we wanted to prove. □

Appendix B. Auxiliary results for Section 4

Proposition 6. *Let $x, y > 0$ be such that $x + y < 1$. Then*

$$1 + \frac{\log(1-x)}{x} + \frac{\log(1-y)}{y} > \frac{\log(1-(x+y))}{x+y}. \tag{B.1}$$

Proof. Define

$$g(x, y) := 1 + \frac{\log(1-x)}{x} + \frac{\log(1-y)}{y} - \frac{\log(1-(x+y))}{x+y}.$$

We now show that $g(x, y) > 0$ for all $x, y > 0$ such that $x + y < 1$. Let $y \in (0, 1)$ be arbitrary but fixed. Notice that

$$\lim_{x \rightarrow 0} g(x, y) = 0.$$

Hence, to prove that $g(x, y) > 0$ for all $x \in (0, 1 - y)$ (for fixed y), it is sufficient to show that $x \mapsto g(x, y)$ is strictly increasing on $(0, 1 - y)$. We have

$$\begin{aligned} \frac{\partial g}{\partial x}(x, y) &= \frac{\log(1-(x+y))}{(x+y)^2} + \frac{1}{(x+y)(1-(x+y))} - \frac{\log(1-x)}{x^2} - \frac{1}{x(1-x)} \\ &= h(x+y) - h(x) \end{aligned}$$

for

$$h(x) := \frac{\log(1-x)}{x^2} + \frac{1}{x(1-x)}.$$

Therefore, it is enough to show that $h(x)$ is a strictly increasing function on $(0, 1)$, since then (for $y \in (0, 1)$ and $x \in (0, 1 - y)$) we would have $\frac{\partial g}{\partial x}(x, y) = h(x+y) - h(x) > 0$. We verify that $h(x)$ is strictly increasing on $(0, 1)$ by showing that $h'(x) > 0$ on $(0, 1)$. We have that

$$h'(x) = \frac{1}{x^3} \left(\frac{x(3x-2)}{(1-x)^2} - 2 \log(1-x) \right).$$

Hence, it remains to prove that

$$\log(1-x) < \frac{x(3x-2)}{2(1-x)^2}. \tag{B.2}$$

One way to see this is to prove that

$$\log(1-x) < -x - \frac{x^2}{2} < \frac{x(3x-2)}{2(1-x)^2}. \tag{B.3}$$

As this is basic analysis we omit the details. □

Lemma 3. Let $p \in (0, 1)$ and let $D = \{x_1, x_2, \dots, x_r \in (0, 1) : \sum_{j=1}^r x_j = 1\}$. The function $g : D \rightarrow \mathbb{R}$ defined by

$$g(x_1, x_2, \dots, x_r) = \sum_{j=1}^r \frac{\log(1 - px_j)}{px_j}$$

has a unique maximum on the set D at the point $(x_1, x_2, \dots, x_r) = (\frac{1}{r}, \frac{1}{r}, \dots, \frac{1}{r})$.

The proof of Lemma 3 is a standard application of the method of Lagrange multipliers and is therefore omitted.

Proposition 7. Let $p \in (0, 1)$ and let

$$\tau_r(p) = (1 - p) \left[r + \frac{r^2 \log(1 - \frac{p}{r})}{p} - \frac{\log(1 - p)}{p} - 1 \right].$$

Then the sequence $(\tau_r(p))_{r \in \mathbb{N}}$ is an increasing sequence for all $p \in (0, 1)$, and it holds that

$$\tau(p) = \lim_{r \rightarrow \infty} \tau_r(p) = (1 - p) \left(-\frac{p}{2} - \frac{\log(1 - p)}{p} - 1 \right).$$

Proof. Let us first show that the sequence $(\tau_r(p))_{r \in \mathbb{N}}$ is strictly increasing. For this, it is sufficient to show that for $p \in (0, 1)$ the function

$$\varphi(x) := x + \frac{x^2 \log(1 - \frac{p}{x})}{p}$$

is strictly increasing on $[1, \infty)$, because the sequence $(\tau_r(p))_{r \in \mathbb{N}}$ satisfies $\tau_r(p) = (1 - p)(\varphi(r) - \varphi(1))$. We will show that $\varphi'(x) > 0$ for all $x \in [1, \infty)$. We have

$$\varphi'(x) = 1 + \frac{2x \log(1 - \frac{p}{x})}{p} + \frac{x}{x - p}.$$

Since $\lim_{x \rightarrow \infty} \varphi'(x) = 0$ and $\varphi'(x)$ is a continuous function on $[1, \infty)$, it is now enough to show that $\varphi'(x)$ is strictly decreasing on $[1, \infty)$ to be able to conclude that $\varphi'(x) > 0$ for all $x \in [1, \infty)$. Observe that

$$\varphi''(x) = \frac{2}{p} \log\left(1 - \frac{p}{x}\right) + \frac{2x - 3p}{(x - p)^2}.$$

We can now check that $\varphi''(x) < 0$ on $[1, \infty)$:

$$\begin{aligned} \varphi''(x) < 0 &\Leftrightarrow \frac{2}{p} \log\left(1 - \frac{p}{x}\right) < \frac{3p - 2x}{(x - p)^2} \\ &\Leftrightarrow \log\left(1 - \frac{p}{x}\right) < \frac{\frac{p}{x}(3\frac{p}{x} - 2)}{2(1 - \frac{p}{x})^2}. \end{aligned}$$

Since $p \in (0, 1)$ and $x \in [1, \infty)$, we have $p/x \in (0, 1)$, so the desired inequality follows from (B.2). Hence $\varphi'(x)$ is decreasing and we have that $\varphi'(x) > 0$ on $[1, \infty)$, which is exactly what we wanted to prove.

Let us now calculate the limit of the sequence $(\tau_r(p))_{r \in \mathbb{N}}$. Notice that

$$\tau_r(p) = (1-p) \left[\frac{1 + \frac{\log(1-\frac{p}{r})}{\frac{p}{r}}}{\frac{1}{r}} - \frac{\log(1-p)}{p} - 1 \right]. \quad (\text{B.4})$$

Applying L'Hospital's rule twice, we obtain

$$\lim_{x \rightarrow 0^+} \frac{1 + \frac{\log(1-px)}{px}}{x} = \lim_{x \rightarrow 0^+} \frac{\frac{-p^2x}{1-px} - p \log(1-px)}{p^2x^2} = \lim_{x \rightarrow 0^+} \frac{-\frac{p^3x}{(1-px)^2}}{2p^2x} = -\frac{p}{2}.$$

Plugging this into Equation (B.4), we obtain

$$\tau(p) = \lim_{r \rightarrow \infty} \tau_r(p) = (1-p) \left(-\frac{p}{2} - \frac{\log(1-p)}{p} - 1 \right),$$

which concludes the proof. \square

Acknowledgements

We wish to thank Serguei Popov for suggesting the name 'greedy sampling' and the whole IOTA research team for stimulating discussions. We also want to thank an anonymous referee for various helpful comments.

Funding information

A. Gutierrez was supported by the Austrian Science Fund (FWF) under Project P29355-N35. S. Šebek was supported by the Austrian Science Fund (FWF) under Project P31889-N35 and the Croatian Science Foundation under Project 4197. This financial support is gratefully acknowledged.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

- [1] ADAMIC, L. A. AND HUBERMAN, B. (2002). Zipf's law and the internet. *Glottometrics* **3**, 143–150.
- [2] CAPOSSELE, A., MUELLER, S. AND PENZKOFER, A. (2021). Robustness and efficiency of voting consensus protocols within Byzantine infrastructures. *Blockchain Res. Appl.* **2**, article no. 100007.
- [3] CHEN, X., PAPADIMITRIOU, C. AND ROUGHGARDEN, T. (2019). An axiomatic approach to block rewards. In *Proc. 1st ACM Conference on Advances in Financial Technologies*, Association for Computing Machinery, New York, pp. 124–131.
- [4] CONDORCET, J. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris.
- [5] DURRETT, R. (2010). *Probability: Theory and Examples*, 4th edn. Cambridge University Press.
- [6] GÁCS, P., KURDYUMOV, G. L. AND LEVIN, L. A. (1978). One-dimensional uniform arrays that wash out finite islands. *Problems Inf. Transmission* **14**, 223–226.
- [7] KAR, S. AND MOURA, J. M. F. (2007). Distributed average consensus in sensor networks with random link failures. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, Vol. 2, Institute of Electrical and Electronics Engineers, Piscataway, NJ, pp. II-1013–II-1016.
- [8] LEONARDOS, S., REIJSBERGEN, D. AND PILIOURAS, G. (2020). Weighted voting on the blockchain: improving consensus in proof of stake protocols. *Internat. J. Network Manag.* **30**, article no. e2093.

- [9] LI, W. (2002). Zipf's law everywhere. *Glottometrics* **5**, 14–21.
- [10] MOREIRA, A. A., MATHUR, A., DIERMEIER, D. AND AMARAL, L. (2004). Efficient system-wide coordination in noisy environments. *Proc. Nat. Acad. Sci. USA* **101**, 12085–12090.
- [11] MÜLLER, S., PENZKOFER, A., CAMARGO, D. AND SAA, O. (2021). On fairness in voting consensus protocols. In *Intelligent Computing*, Springer, Cham, pp. 927–939.
- [12] MÜLLER, S. *et al.* (2020). Fast probabilistic consensus with weighted votes. In *Proceedings of the Future Technologies Conference (FTC) 2020*, Vol. 2, Springer, Cham, pp. 360–378.
- [13] POPOV, S. (2016). A probabilistic analysis of the Nxt forging algorithm. *Ledger* **1**, 69–83.
- [14] POPOV, S. AND BUCHANAN, W. J. (2021). FPC-BI: Fast Probabilistic Consensus within Byzantine Infrastructures. *J. Parallel Distributed Comput.* **147**, 77–86.
- [15] POPOV, S. *et al.* (2020). The Coordicide. Available at https://files.iota.org/papers/20200120_Coordicide_WP.pdf.
- [16] RAJ, D. AND KHAMIS, S. H. (1958). Some remarks on sampling with replacement. *Ann. Math. Statist.* **29**, 550–557.
- [17] TAO, T. (2009). Benford's law, Zipf's law, and the Pareto distribution. Available at <https://terrytao.wordpress.com/2009/07/03/benfordslaw-zipfs-law-and-the-pareto-distribution>.