Optimal Proportion Computation with Population Protocols

Yves Mocquard
yves.mocquard@irisa.fr
IRISA / Université de Rennes 1, France

Emmanuelle Anceaume
emmanuelle.anceaume@irisa.fr
IRISA / CNRS, France

Bruno Sericola
bruno.sericola@inria.fr
Inria Rennes - Bretagne Atlantique, France

***Abstract*—The computational model of population protocols is a formalism that allows the analysis of properties emerging from simple and pairwise interactions among a very large number of anonymous finite-state agents. Significant work has been done so far to determine which problems are solvable in this model and at which cost in terms of states used by the agents and time needed to converge. The problem tackled in this paper is the *population proportion* problem: each agent starts independently from each other in one of two states, say A or B, and the objective is for each agent to determine the proportion of agents that initially started in state A, assuming that each agent only uses a finite set of states, and does not know the number n of agents. We propose a solution which guarantees that in presence of a uniform probabilistic scheduler every agent outputs the population proportion with any precision $\varepsilon \in (0, 1)$ with any high probability after having interacted $O(\log n)$ times. The number of states maintained by every agent is optimal and is equal to $2\lceil 3/(4\varepsilon)\rceil + 1$. Finally, we show that our solution is optimal in time and space to solve the counting problem, a generalization of the proportion problem. Finally, simulation results illustrate our theoretical analysis.**

Keywords Population protocols; Proportion; Majority; Counting; Performance evaluation.

This paper should be considered for the best student paper award. Yves Mocquard is PhD student at the University of Rennes 1. E-mail: yves.mocquard@irisa.fr

I. Introduction

In 2004, Angluin et al. [4] have proposed a model that allows the analysis of emergent global properties based on pairwise interactions. This model, named the population protocol, provides minimalist assumptions on the computational power of the agents: agents are finite-state automata, identically programmed, with no identity, unaware of the population size n, and they progress in their computation through random pairwise interactions. The objective of this model is to ultimately converge to a state from which the sought property can be derived from any agent [8]. Examples of systems whose behavior can be modeled by population protocols range from molecule interactions of a chemical process to sensor networks in which agents, which are small devices embedded on animals, interact each time two animals are in the same radio range. A considerable amount of work has been done so far to determine which properties can emerge from pairwise interactions between finite-state agents, together with the derivation of lower bounds on the time and space needed to reach such properties (e.g., [2], [6], [12], [14], [18]). Among them, is majority. Briefly, each agent starts independently from each other in one of two input states, say A and B, and the objective for each agent is to eventually output *yes* if a majority of agents started their execution in input state A and *no* otherwise. Section IV provides an overview of the results recently obtained for the majority task.

In this paper, we focus on a related but more general question. Namely, instead of having each agent answer *yes* if a majority of agents initially started their execution in input state A, one may ask the following question:

"Is it feasible for each agent to compute *quickly* and with any *high precision* the *proportion of agents* that started in the input state A?".

Answering such a question is very important in the context of, for example, infectious-disease surveillance of large-scale animal populations. In this context, different kinds of alerts could be triggered according to the infected population proportion (*e.g.*, Alert 1 is triggered if less than 0.05% of the population is infected, Alert 2 if this proportion lies in $[0.05\%, 3.0\%)$, Alert 3 if it lies in $[3.0\%, 10.0\%)$, and so on ...). Input state A would manifest an excessive temperature of an animal while input state B would indicate a safe temperature. By relying on the properties exhibited by our population protocol (convergence time logarithmic in the population size and memory space proportional to the sought precision), one can easily implement a regular and self-autonomous monitoring of large-scale populations.

We answer affirmatively to this question, and we propose a population protocol that allows each agent to converge to a state which, when queried, provides the proportion of agents that started in a given input state. Specifically, each agent is a $(2m + 1)$-finite state machine, $m \geq 1$, where m is the value associated to input state A and $-m$ is the one associated to input state B. Each agent starts its execution with m or $-m$, and each pair of agents that meet, adopt the average of their values (or as close as they can get when values are restricted to integers, as will be clarified in Section V). The rationale of this method [3], [16], [1] is to preserve the sum of the initial values, and after a small number of pairwise interactions, to ensure that every agent converges with high probability to a state from which it derives the proportion of agents that started in a given state. Technically, our protocol guarantees that each agent is capable of computing with any precision $\varepsilon \in (0, 1)$ the proportion of agents that initially started in a specific input state by using $2\lceil 3/(4\varepsilon)\rceil + 1$ states. This is achieved

This work was partially funded by the French ANR project SocioPlug (ANR-13-INFR-0003), and by the DeSceNt project granted by the Labex CominLabs excellence laboratory (ANR-10-LABX-07-01).

in no more than $(-2\ln\varepsilon + 8.47\ln n - 13.29\ln\delta - 2.88)$ interactions with probability at least $1-\delta$, for any $\delta \in (0,1)$.

Our second contribution relates to the counting problem. The counting problem generalizes the majority problem by requiring, for each agent, to converge to a state in which each agent is capable of, assuming the knowledge of n, computing n_A or n_B, where n_A and n_B represent respectively the number of agents that started in state A and B. In the present paper, we prove that the counting problem can be solved using $O(n)$ states per agent. This significantly improves upon a previous analysis [16] that shows that $O(n^{3/2}/\delta^{1/2})$ states allow each agent to converge to the exact solution in no more than a logarithmic number in n of interactions, with $\delta \in (0,1)$. What is very important to notice is that this drastic improvement is due to an original convergence analysis that allows us to refine previous results. Indeed, both [16] and the present paper rely on the same interaction rules, however by precisely characterizing the evolution of the interacting agents, our present analysis is highly tighter. We also demonstrate that any protocol that solves the counting problem requires $\Omega(\log n)$ parallel interactions to converge and $\Omega(n)$ local states. As will be detailed, this shows that our algorithm is an optimal solution both in space and time to solve the counting problem and optimal in space to solve the proportion one.

The remainder of this paper is organized as follows. Section II presents the population protocol model. Section III specifies the problem addressed in this work. Section IV provides an overview of the most recent population protocols. The protocol to compute the population proportion is presented in Section V. Analysis of the protocol is detailed in Section VI. We show in Section VII, that our protocol is optimal both in space and time. We have simulated our protocol to illustrate our theoretical analysis. Section VIII presents a summary of these simulation results. Finally, Section IX concludes.

II. Population protocols model

The population protocol model has been introduced by Angluin et al. [4]. This model describes the behavior of a collection of agents that interact pairwise. The following definition is from Angluin et al [7]. A population protocol is characterized by a 6-tuple $(Q, \Sigma, Y, \iota, \omega, f)$, over a complete interaction graph linking the set of n agents, where Q is a finite set of states, Σ is a finite set of input symbols, Y is a finite set of output symbols, $\iota : \Sigma \to Q$ is the input function that determines the initial state of an agent, $\omega : Q \to Y$ is the output function that determines the output symbol of an agent, and $f : Q \times Q \to Q \times Q$ is the transition function that describes how any two distinct agents interact and locally update their states. Initially all the agents start with a initial symbol from Σ, and upon interactions update their state according to the transition function f. Interactions between agents are orchestrated by a random scheduler: at each discrete time, any two agents are randomly chosen to interact with a given distribution. Note that the random scheduler is fair, meaning that any possible interaction cannot be avoided forever. The notion of time in population protocols refers to as the successive steps at which interactions occur, while the parallel time is equal to the total number of interactions averaged by n [8]. Agents do not maintain nor use identifiers (agents are anonymous and cannot determine whether any two interactions have occurred with the same agents or not). However, for ease of presentation, the agents are numbered $1, 2, \ldots, n$. We denote by $C_t^{(i)}$ the state of agent i at time t. The stochastic process $C = \{C_t,\ t \geq 0\}$, where $C_t = (C_t^{(1)}, \ldots, C_t^{(n)})$, represents the evolution of the population protocol. The state space of C is thus Q^n and a state of this process is also called a protocol configuration.

III. The Proportion Problem

We consider a set of n agents, interconnected by a complete graph, that start their execution in one of two input states of $\Sigma = \{A, B\}$. Let n_A be the number of agents whose input state is A and n_B be the number of agents that start in input state B. The quantity $\gamma_A = n_A/n$ (resp. $\gamma_B = n_B/n$) is the proportion of the agents that initially started in state A (resp. in state B). The output set Y is the set of all possible values of γ_A, that is a subset of $[0,1]$. In the following we introduce the notation $\gamma = \gamma_A - \gamma_B$. Let $\omega_A(C_t^{(i)})$ be the approximation of γ_A by agent i at time t.

A population protocol solves the proportion problem within τ steps (with preferably τ in $O(\log n)$) if for all $\delta \in (0,1)$, for all $\varepsilon \in (0,1)$ and for all $t \geq \tau$, we have

$$\mathbb{P}\{|\omega_A(C_t^{(i)}) - \gamma_A| < \varepsilon \text{ for all } i = 1, \ldots, n\} \geq 1 - \delta.$$

IV. Related Work

In 2004, Angluin et al. [4] have formalized the population protocol model, and have shown how to express and compute predicates in this model. Then in [5] the authors have completely characterized the computational power of the model by establishing the equivalence between predicates computable in the population model and those that can be defined in the Presburger arithmetic. Since then, there has been a lot of work on population protocols including the majority problem [12], [14], [6], [18], [2], the leader election problem [9], [15], in presence of faults [11], and on variants of the model [13], [10].

The closest problem to the one we address is the computation of the majority. In this problem, all the agents start in one of two distinguished states and they eventually converge to 1 if $\gamma > 0$ (*i.e.* $n_A > n_B$), and to 0 if $\gamma < 0$ (*i.e.* $n_A < n_B$). In [12], [14] the authors propose a four-state protocol that solves the majority problem with a convergence parallel time logarithmic in n but only in expectation. Moreover, the expected convergence time is infinite when n_A and n_B are close to each other (that is γ approaches 0). The authors in [6], [18] propose a three-state protocol that converges with high probability after a convergence parallel time logarithmic in n but only if γ is large enough, i.e when $|n_A - n_B| \geq \sqrt{n}\log n$. Alistarh et al. [2] propose a population protocol based on an average-and-conquer method to exactly solve the majority problem. Their algorithm uses two types of interactions, namely, averaging interactions and conquer ones. The first type of interaction is close to the one used in our protocol while the second one is used to diffuse the result of the computation to the zero state agents. Actually, to show their convergence time, they need to assume a rather large number of intermediate states (*i.e.* $2d$ states, with $d = 1,000$). This is essentially due to the fact that they need to prove that all the agents with maximum positive values and minimal negative

values will have sufficiently enough time to halve their values. Note that in practice, their algorithm does not require more than n state to converge to the majority, however their proof necessitates $m + 1,000 \log m \log n$ with $\log n \log m \leq m \leq n$ states, and at least $432 \log m \log n$ interactions per agent to converge to the majority, where m is the initial value associated to state A.

In [16], the authors have presented a solution to the counting problem. As previously said, the counting problem generalizes the majority problem by requiring, for each agent, to converge to a state in which each agent is capable of, assuming the knowledge of n, computing n_A or n_B, where n_A and n_B represent respectively the number of agents that started in state A and B. Both [16] and the present paper use the same interaction rules, but of course the output functions in both papers are different. The originality of [16], beyond tackling a new problem, was a proof of convergence based on tracking the euclidean distance between the random vector of all agents' values and the limiting distribution. In the present paper, we provide a highly tighter analysis which shows that the interaction rules together with the "counting" and "proportion" output functions are optimal solutions to solve both problems.

V. Computing the Proportion

Our protocol uses the average technique to compute the proportion of agents that started their execution in a given state A. The set of input of the protocol is $\Sigma = \{A, B\}$, and the input function ι is defined by $\iota(A) = m$ and $\iota(B) = -m$, with m a positive integer. This means that, for every $i = 1, \ldots, n$, we have $C_0^{(i)} \in \{-m, m\}$. At each discrete instant t, two distinct indices i and j are chosen among $1, \ldots, n$ with probability $p_{i,j}(t)$. Once chosen, the couple (i, j) interacts, and both agents update their respective local state $C_t^{(i)}$ and $C_t^{(j)}$ by applying the transition function f, leading to state C_{t+1}, given by $f(C_t^{(i)}, C_t^{(j)}) = (C_{t+1}^{(i)}, C_{t+1}^{(j)})$, with

$$\left(C_{t+1}^{(i)}, C_{t+1}^{(j)}\right) = \left(\left\lfloor \frac{C_t^{(i)} + C_t^{(j)}}{2} \right\rfloor, \left\lceil \frac{C_t^{(i)} + C_t^{(j)}}{2} \right\rceil\right) \text{ and}$$
$$C_{t+1}^{(m)} = C_t^{(m)} \text{ for } m \neq i, j. \quad (1)$$

The set Q of states is $\{-m, -m+1, \ldots, m-1, m\}$. The output function is given, for all $x \in Q$ by,

$$\omega_A(x) = (m + x)/2m.$$

Finally, the set of output Y is the set of all possible values of ω_A, i.e.

$$Y = \left\{0, \frac{1}{2m}, \frac{2}{2m}, \ldots, \frac{2m-2}{2m}, \frac{2m-1}{2m}, 1\right\}.$$

VI. Analysis of the Proportion Protocol

We denote by X_t the random variable representing the choice at time t of two distinct indices i and j among $1, \ldots, n$ with probability $p_{i,j}(t)$, that is $\mathbb{P}\{X_t = (i, j)\} = p_{i,j}(t)$. We suppose that the sequence $\{X_t,\ t \geq 0\}$ is a sequence of independent and identically distributed random variables. Since C_t is entirely determined by the values of $C_0, X_0, X_1, \ldots, X_{t-1}$, this means in particular that the random variables X_t and C_t are independent and that the stochastic process C is a discrete-time homogeneous Markov chain. As usual in population protocols, we suppose that X_t is uniformly distributed, i.e. that is

$$p_{i,j}(t) = \frac{1}{n(n-1)}.$$

We will use in the sequel the Euclidean norm denoted simply by $\|.\|$ and the infinite norm denoted by $\|.\|_\infty$ defined for all $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ by

$$\|x\| = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2} \text{ and } \|x\|_\infty = \max_{i=1,\ldots,n} |x_i|.$$

It is well-known that these norms satisfy $\|x\|_\infty \leq \|x\| \leq \sqrt{n}\|x\|_\infty$.

Lemma *1:* For every $t \geq 0$, we have

$$\sum_{i=1}^{n} C_t^{(i)} = \sum_{i=1}^{n} C_0^{(i)}.$$

Proof: The proof is immediate since the transformation from C_t to C_{t+1} described in Relation (1) does not change the sum of the entries of C_{t+1}. Indeed, from Relation (1), we have $C_{t+1}^{(i)} + C_{t+1}^{(j)} = C_t^{(i)} + C_t^{(j)}$ and the other entries do not change their values. ∎

We denote by ℓ the mean value of the sum of the entries of C_t and by L the row vector of $\mathbb{R}^n$ with all its entries equal to ℓ, that is

$$\ell = \frac{1}{n}\sum_{i=1}^{n} C_t^{(i)} \text{ and } L = (\ell, \ldots, \ell).$$

Our analysis is orchestrated as follows. By relying on the mathematical tool derived in Theorem 2, we show in Theorem 5 that the stochastic process C_t belongs to the ball of radius $\sqrt{n/2}$ and center L in the 2-norm, with any high probability, after no more than $O(\log n)$ parallel time. Then, assuming that the stochastic process C_t belongs to the ball of radius $\sqrt{n/2}$ and that $\ell - \lfloor \ell \rfloor \neq 1/2$, we demonstrate that the stochastic process C_t belongs to the open ball of radius $3/2$ and center L in the infinite norm, with any high probability after no more than $O(\log n)$ parallel time (Theorem 6). In practice this means that all the entries of the subsequent configurations will be among the three closest integer values of ℓ. Then by applying Theorem 5 and Theorem 6 (if $\ell - \lfloor \ell \rfloor \neq 1/2$) or Theorem 4 (otherwise), we derive our main theorem (see Theorem 7) which shows that in both cases the stochastic process C_t belongs to an open ball of radius $3/2$ and center L in the infinite norm, with any high probability in $O(\log n)$ parallel time. Finally, we have all the necessary tools to construct an output function which solves the proportion problem in $O\left(\log n - \log \varepsilon - \log \delta\right)$ parallel time, and with $O(1/\varepsilon)$ states, for any $\varepsilon, \delta \in (0, 1)$ (see Theorem 8). The detailed proofs of all the results are given in [17].

In order to simplify the writing we will use the notation $Y_t = \|C_t - L\|^2$ when needed and we denote by $1_{\{A\}}$ the indicator function which is equal to 1 if condition A is satisfied and 0 otherwise.

The following Theorem is a conditional version of Theorem 6 of [16].

Theorem *2:* For every $0 \leq s \leq t$ and $y \geq 0$, we have

$$\mathbb{E}\left(Y_t \mid Y_s \geq y\right) \leq \left(1 - \frac{1}{n-1}\right)^{t-s} \mathbb{E}\left(Y_s \mid Y_s \geq y\right) + \frac{n}{4}. \tag{2}$$

Proof: See [17]. ∎

Lemma *3:* The sequence $Y_t = \|C_t - L\|^2$ is decreasing with t.

Proof: See [16]. ∎

Theorem *4:* For all $\delta \in (0,1)$, if $\ell - \lfloor \ell \rfloor = 1/2$ and if there exists a constant K such that $\|C_0 - L\|_\infty \leq K$, then, for every $t \geq (n-1)\left(2\ln K + \ln n - \ln \delta\right)$, we have

$$\mathbb{P}\{\|C_t - L\|_\infty \neq 1/2\} \leq \delta.$$

Sketch of the Proof: If $\ell - \lfloor \ell \rfloor = 1/2$ then, since all the $C_t^{(i)}$ are integers, we have $\|C_t - L\|^2 \geq n/4$. From Theorem 2 in which we set $s = 0$ and $y = 0$, we obtain

$$\mathbb{E}(\|C_t - L\|^2 - n/4) \leq \left(1 - \frac{1}{n-1}\right)^t \mathbb{E}(\|C_0 - L\|^2).$$

Let $\tau = (n-1)\left(2\ln K + \ln n - \ln \delta\right)$. For $t \geq \tau$, the Markov inequality leads to

$$\mathbb{P}\{\|C_t - L\|^2 - n/4 \geq 1\} \leq \delta.$$

We then conclude by observing that

$$\mathbb{P}\{\|C_t - L\|_\infty \neq \frac{1}{2}\} = \mathbb{P}\{\|C_t - L\|^2 - n/4 \geq 1\} \leq \delta.$$

The reader is invited to read the detailed proof in [17]. ∎

Theorem *5:* For all $\delta \in (0, 4/5)$, if there exists a constant K such that $K \geq \sqrt{n/2}$ and $\|C_0 - L\| \leq K$ then, for all $t \geq n\theta$, we have

$$\mathbb{P}\{\|C_t - L\|^2 \geq n/2\} \leq \delta$$

where

$$\theta = 2\ln K - \ln n + 3\ln 2 - \frac{2\ln 2}{2\ln 2 - \ln 3}\ln \delta.$$

Sketch of the Proof: Let $(T_k)_{k \geq 0}$ be the sequence of instants defined by $T_0 = 0$ and

$$T_{k+1} = T_k + \left\lceil (n-1)\ln\left(\frac{8\mathbb{E}(Y_{T_k} \mid Y_{T_k} \geq n/2)}{n}\right)\right\rceil. \tag{3}$$

From Theorem 2 and Formula 3 we get

$$\mathbb{E}\left(Y_{T_{k+1}} \mid Y_{T_k} \geq n/2\right) \leq 3n/8. \tag{4}$$

Using the conditional Markov inequality, we deduce that

$$\mathbb{P}\{Y_{T_{k+1}} \geq n/2 \mid Y_{T_k} \geq n/2\} \leq 3/4.$$

We introduce the sequence α_k defined by $\alpha_0 = 3n/(8K^2)$ and, for $k \geq 1$,

$$\alpha_k = \max\left\{\mathbb{P}\{Y_{T_k} \geq n/2 \mid Y_{T_{k-1}} \geq n/2\}, 3n/(8K^2)\right\}.$$

We then obtain for every $k \geq 0$,

$$\mathbb{E}(Y_{T_k} \mid Y_{T_k} \geq n/2) \leq 3n/(8\alpha_k).$$

Summing the differences $T_{i+1} - T_i$ for i from 0 to $k-1$, we obtain, for $k \geq 1$,

$$T_k \leq (n-1)\left(k\ln(3) - \ln\left(\prod_{i=0}^{k-1}\alpha_i\right)\right) + k, \tag{5}$$

and we have

$$\mathbb{P}\{Y_{T_k} \geq n/2\} \leq \prod_{i=1}^{k}\alpha_i. \tag{6}$$

Now, for all $\delta \in (0, 4/5)$, there exists $k \geq 1$ such that

$$\prod_{i=1}^{k}\alpha_i < \delta \leq \prod_{i=1}^{k-1}\alpha_i \leq (3/4)^{k-1}.$$

From Relation (6) and using the fact that Y_t is decreasing (see Lemma 3) we get, for $t \geq n\theta$,

$$\begin{aligned}\mathbb{P}\{Y_t \geq n/2\} &\leq \mathbb{P}\{Y_{n\theta} \geq n/2\} \\ &\leq \mathbb{P}\{Y_{T_k} \geq n/2\} \\ &\leq \prod_{i=1}^{k}\alpha_i \leq \delta.\end{aligned}$$

The reader is invited to read the detailed proof in [17]. ∎

Theorem *6:* For all $\delta \in (0,1)$, if $\|C_0 - L\| \leq \sqrt{n/2}$ and $\ell - \lfloor \ell \rfloor \neq 1/2$ we have, for every $t \geq 1600(n-1)\left(\ln n - \ln \delta - 4\ln 2 + \ln 3\right)/189$,

$$\mathbb{P}\{\|C_t - L\|_\infty \geq 3/2\} \leq \delta.$$

Sketch of the Proof: Let λ be defined by

$$\lambda = \begin{cases} \ell - \lfloor \ell \rfloor & \text{if} \quad \ell - \lfloor \ell \rfloor < 1/2 \\ \ell - \lceil \ell \rceil & \text{if} \quad \ell - \lfloor \ell \rfloor > 1/2. \end{cases}$$

Note that λ is positive in the first case and negative in the second one. In both cases we have $|\lambda| < 1/2$ and $\ell - \lambda$ is the closest integer to ℓ.

We introduce the notation $B = \lceil 1/2 + \sqrt{n/2} \rceil$. B is the upper bound of $C_t^{(i)}$ for $i = 1, \ldots, n$, since $\|C_0 - L\| \leq \sqrt{n/2}$.

For $k \in \{-B, -B+1, \ldots, B\}$, we denote by $\alpha_{k,t}$ the number of agents with the value $\ell - \lambda + k$ at time t, that is

$$\alpha_{k,t} = \left|\left\{i \in \{1, \ldots, n\} \mid C_t^{(i)} = \ell - \lambda + k\right\}\right|.$$

It is easily checked that

$$\sum_{k=-B}^{B}\alpha_{k,t} = n. \tag{7}$$

Moreover we have

$$\sum_{k=-B}^{B} k\alpha_{k,t} = n\lambda. \tag{8}$$

Observing that $\|C_t - L\|^2 = \|C_t\|^2 - n\ell^2$ and using (7) and (8), we obtain

$$\sum_{k=-B}^{B} k^2 \alpha_{k,t} = \|C_t - L\|^2 + n\lambda^2. \tag{9}$$

Using the hypothesis $\|C_0 - L\|^2 \leq n/2$, we obtain

$$\sum_{k=-B}^{B} k^2 \alpha_{k,t} \leq n\lambda^2 + n/2. \tag{10}$$

Using these results, it can be shown that

$$\sum_{k=0}^{B} \alpha_{k,t} > 3n/8 \qquad \text{and} \qquad \sum_{k=-B}^{0} \alpha_{k,t} > 3n/8. \tag{11}$$

Let us now introduce the sequences $(N_t)_{t\geq 0}$ and $(\Phi_t)_{t\geq 0}$ defined by

$$N_t = \sum_{k=2}^{B} \alpha_{k,t} + \sum_{k=-B}^{-2} \alpha_{k,t}$$

and

$$\Phi_t = \sum_{k=2}^{B} k^2 \alpha_{k,t} + \sum_{k=-B}^{-2} k^2 \alpha_{k,t}.$$

Since $\alpha_{k,t}$ are non negative integers, we have, for every $t \geq 0$,

$$N_t = 0 \Longleftrightarrow \Phi_t = 0.$$

Note that our objective is to obtain $\Phi_t = 0$ because

$$\|C_t - L\|_\infty < 3/2 \Longleftrightarrow \Phi_t = 0.$$

We also introduce the sets H_t^+ and H_t^- defined by

$$H_t^+ = \{i \in \{1, \ldots, n\} \mid C_t^{(i)} - \ell + \lambda \geq 2\}$$

and

$$H_t^- = \{i \in \{1, \ldots, n\} \mid C_t^{(i)} - \ell + \lambda \leq -2\}$$

and we define $H_t = H_t^+ \cup H_t^-$. It can be shown that

$$N_t \leq 3n/16.$$

Let I_t^+ and I_t^- be the sets defined by

$$I_t^+ = \{i \in \{1, \ldots, n\} \mid C_t^{(i)} - \ell + \lambda \geq 0\}$$

and

$$I_t^- = \{i \in \{1, \ldots, n\} \mid C_t^{(i)} - \ell + \lambda \leq 0\}.$$

Relations (11) can be rewritten as

$$|I_t^+| \geq 3n/8 \text{ and } |I_t^-| \geq 3n/8. \tag{12}$$

Consider the probability that an agent of H_t^+ interacts with an agent of I_t^- or that an agent of H_t^- interacts with an agent of I_t^+, at time t. Let E denote the set of these interactions. It can be shown that

$$\mathbb{P}\{X_t \in E\} \geq \frac{21 N_t}{32(n-1)}. \tag{13}$$

We consider now the difference $\Phi_t - \Phi_{t+1}$ in function of the interactions occurring at time t. It can be shown that

$$\mathbb{E}(\Phi_t - \Phi_{t+1} \mid X_t \in E) \geq \frac{9\mathbb{E}(\Phi_t)}{50 N_t}.$$

Now, using (13), we have

$$\begin{aligned}
\mathbb{E}(\Phi_{t+1}) &= \mathbb{E}(\Phi_t) - \mathbb{E}(\Phi_t - \Phi_{t+1}) \\
&\leq \mathbb{E}(\Phi_t) - \mathbb{E}((\Phi_t - \Phi_{t+1}) \mid X_t \in E)\mathbb{P}\{X_t \in E\} \\
&\leq \mathbb{E}(\Phi_t) - \left(\frac{9\mathbb{E}(\Phi_t)}{50 N_t}\right)\left(\frac{21 N_t}{32(n-1)}\right) \\
&= \left(1 - \frac{189}{1600(n-1)}\right)\mathbb{E}(\Phi_t).
\end{aligned}$$

We easily get

$$\mathbb{E}(\Phi_t) \leq \left(1 - \frac{189}{1600(n-1)}\right)^t \mathbb{E}(\Phi_0).$$

Let τ be defined by

$$\tau = \frac{1600(n-1)}{189}\left(\ln n - \ln \delta - 4\ln 2 + \ln 3\right).$$

We then have for $t \geq \tau$

$$\mathbb{P}\{\|C_t^{(i)} - \ell\|_\infty \geq 3/2\} \leq \mathbb{P}\{\Phi_t \neq 0\} = \mathbb{P}\{\Phi_t \geq 4\} \leq \delta.$$

The reader is invited to read the detailed proof in [17]. ■

Theorem *7:* For all $\delta \in (0,1)$, if there exists a constant K such that $\|C_0 - L\| \leq K$ then, for every $t \geq n\left(2\ln K + 7.47 \ln n - 13.29 \ln \delta - 2.88\right)$, we have

$$\mathbb{P}\{\|C_t - L\|_\infty \geq 3/2\} \leq \delta.$$

Proof: We consider first the case where $\ell - \lfloor \ell \rfloor = 1/2$. Since $\|C_0 - L\|_\infty \leq \|C_0 - L\| \leq K$ and since

$$\begin{aligned}
&(n-1)(2\ln K + \ln n - \ln \delta) \\
&\qquad \leq n\left(2\ln K + 7.47 \ln n - 13.29 \ln \delta - 2.88\right),
\end{aligned}$$

Theorem 4 gives

$$\mathbb{P}\{\|C_t - L\|_\infty \neq 1/2\} \leq \delta,$$

for $t \geq n\left(2\ln K + 7.47 \ln n - 13.29 \ln \delta - 2.88\right)$.

Now since the $C_t^{(i)}$ are integers and since $\ell - \lfloor \ell \rfloor = 1/2$, we have

$$\mathbb{P}\{\|C_t - L\|_\infty \geq 3/2\} = \mathbb{P}\{\|C_t - L\|_\infty \neq 1/2\} \leq \delta.$$

Consider now the case where $\ell - \lfloor \ell \rfloor \neq 1/2$. We apply successively Theorem 5 and Theorem 6 replacing δ by $\delta/2$. We introduce the notation

$$\theta_1 = 2\ln K - \ln n + 3\ln 2 - \frac{2\ln 2}{2\ln 2 - \ln 3}\ln(\delta/2).$$

If $\|C_0 - L\| < \sqrt{n/2}$ then we have $\|C_0 - L\|^2 < n/2$ and since $\|C_t - L\|^2$ is decreasing (see Lemma 3), we get, for all $t \geq 0$,

$$\begin{aligned}
\mathbb{P}\{\|C_t - L\|^2 < n/2\} &\geq \mathbb{P}\{\|C_0 - L\|^2 < n/2\} \\
&= 1 \geq 1 - \delta/2.
\end{aligned}$$

If $\|C_0 - L\| \geq \sqrt{n/2}$ then from Theorem 5 we get, for all $t \geq n\theta_1$, $\mathbb{P}\{\|C_t - L\|^2 \geq n/2\} \leq \delta/2$, or equivalently

$$\mathbb{P}\{\|C_t - L\|^2 < n/2\} \geq 1 - \delta/2.$$

Let us introduce the instant τ defined by

$$\tau = n\theta_1 + \frac{1600(n-1)}{189}\left(\ln n - \ln(\delta/2) - 4\ln 2 + \ln 3\right).$$

We have, for all $t \geq \tau$,

$$\begin{aligned}\mathbb{P}\{\|C_t - L\|_\infty < 3/2\} &\geq \mathbb{P}\{\|C_t - L\|_\infty < 3/2, \|C_{n\theta_1} - L\|^2 < n/2\} \\ &= \mathbb{P}\{\|C_t - L\|_\infty < 3/2 \mid \|C_{n\theta_1} - L\|^2 < n/2\} \\ &\quad \times \mathbb{P}\{\|C_{n\theta_1} - L\|^2 < n/2\}.\end{aligned}$$

We have seen that $\mathbb{P}\{\|C_{n\theta_1} - L\|^2 < n/2\} \geq 1 - \delta/2$. Using the fact that the Markov chain $\{C_t\}$ is homogeneous and applying Theorem 6, we obtain

$$\begin{aligned}\mathbb{P}\{\|C_t - L\|_\infty < 3/2 \mid \|C_{n\theta_1} - L\|^2 < n/2\} &= \mathbb{P}\{\|C_{t-n\theta_1} - L\|_\infty < 3/2 \mid \|C_0 - L\|^2 < n/2\} \\ &= \mathbb{P}\{\|C_{t-n\theta_1} - L\|_\infty < 3/2 \mid \|C_0 - L\| < \sqrt{n/2}\} \\ &\geq 1 - \delta/2.\end{aligned}$$

Putting together these two results gives, for all $t \geq \tau$,

$$\mathbb{P}\{\|C_t - L\|_\infty < 3/2\} \geq (1 - \delta/2)^2 \geq 1 - \delta$$

or equivalently

$$\mathbb{P}\{\|C_t - L\|_\infty \geq 3/2\} \leq \delta.$$

The rest of the proof consists in simplifying the expression of τ. We have

$$\begin{aligned}\theta_1 &= 2\ln K - \ln n + 3\ln 2 - \frac{2\ln 2}{2\ln 2 - \ln 3}\ln(\delta/2) \\ &= 2\ln K - \ln n + \left(4 + \frac{\ln 3}{2\ln 2 - \ln 3}\right)\ln 2 \\ &\qquad - \frac{2\ln 2}{2\ln 2 - \ln 3}\ln\delta\end{aligned}$$

and

$$\begin{aligned}\tau &= n\theta_1 + \frac{1600(n-1)}{189}(\ln n - \ln(\delta/2) - 4\ln 2 + \ln 3) \\ &= n\theta_1 + \frac{1600(n-1)}{189}(\ln n - \ln\delta - 3\ln 2 + \ln 3) \\ &\leq n\left[2\ln K + \frac{1411}{189}\ln n - \left(\frac{1789}{189} + \frac{\ln 3}{2\ln 2 - \ln 3}\right)\ln\delta \right. \\ &\qquad \left. - \left(\frac{1348}{63} - \frac{\ln 3}{2\ln 2 - \ln 3}\right)\ln 2 + \frac{1600}{189}\ln 3\right] \\ &\leq n\left(2\ln K + 7.47\ln n - 13.29\ln\delta - 2.88\right),\end{aligned}$$

which completes the proof. ∎

We now apply these results to compute the proportion γ_A of agents whose initial input was A, with $\gamma_A = n_A/(n_A + n_B) = n_A/n$. Recall that the output function ω_A is given, for all $x \in Q$, by

$$\omega_A(x) = (m + x)/(2m).$$

Theorem *8:* For all $\delta \in (0,1)$ and for all $\varepsilon \in (0,1)$, by setting $m = \lceil 3/(4\varepsilon)\rceil$, we have, for all $t \geq n\left(8.47\ln n - 2\ln\varepsilon - 13.29\ln\delta - 2.88\right)$,

$$\mathbb{P}\{|\omega_A(C_t^{(i)}) - \gamma_A| < \varepsilon \text{ for all } i = 1, \ldots, n\} \geq 1 - \delta.$$

Proof: We have $\|C_0 - L\| \leq m\sqrt{n}$. Applying Theorem 7, with $K = \sqrt{n}/\varepsilon \geq \lceil 3/(4\varepsilon)\rceil\sqrt{n} = m\sqrt{n}$, we obtain for all $\delta \in (0,1)$ and $t \geq n\left(8.47\ln n - 2\ln\varepsilon - 13.29\ln\delta - 2.88\right)$,

$$\mathbb{P}\{\|C_t - L\|_\infty \geq 3/2\} \leq \delta$$

or equivalently

$$\mathbb{P}\{|C_t^{(i)} - (\gamma_A - \gamma_B)m| < 3/2, \text{ for all } i = 1, \ldots, n\} \geq 1 - \delta.$$

Since $\gamma_A + \gamma_B = 1$ we have

$$\begin{aligned}|C_t^{(i)} - (\gamma_A - \gamma_B)m| &= |C_t^{(i)} - (2\gamma_A - 1)m| \\ &= |m + C_t^{(i)} - 2m\gamma_A| \\ &= 2m|\omega_A(C_t^{(i)}) - \gamma_A|.\end{aligned}$$

Then

$$\mathbb{P}\{|\omega_A(C_t^{(i)}) - \gamma_A| < 3/(4m), \text{ for all } i = 1, \ldots, n\} \geq 1 - \delta.$$

So

$$\mathbb{P}\{|\omega_A(C_t^{(i)}) - \gamma_A| < \varepsilon, \text{ for all } i = 1, \ldots, n\} \geq 1 - \delta,$$

which completes the proof. ∎

From Theorem 8, the convergence time to get the proportion γ_A of agents that were in the initial state A, with any precision ε and with any high probability $1 - \delta$ is $O\left(n(\log n - \log\varepsilon - \log\delta)\right)$ and thus the corresponding parallel convergence time is $O\left(\log n - \log\varepsilon - \log\delta\right)$. Still from Theorem 8, the size of the set of states to compute γ_A is equal to $2\lceil 3/(4\varepsilon)\rceil + 1$. It is important to note that the number of states does not depend, even logarithmically, in n.

VII. Lower Bounds

The second contribution of our paper is the derivation of lower bounds on a more general problem, namely the counting problem, introduced in [16]. This problem aims, for each agent, at computing the exact number of agents that started in the initial state A. Using the interaction rules given in Relation (1) and the output function

$$\omega'_A(x) = \lfloor n(m + x)/(2m) + 1/2\rfloor,$$

we can exploit the results derived in the present paper to show that the counting problem can be solved with $O(n)$ states, improving upon [16] in which the number of states is in $O(n^{3/2})$. We show that $O(n)$ states and $O(\log n)$ parallel time are lower bounds to solve the counting problem.

Finally, we prove that any algorithm solving the proportion problem with a precision $\varepsilon \in (0,1)$, requires $\Omega(1/\varepsilon)$ states. This demonstrates that our proportion protocol is optimal in the number of states.

Theorem *9:* By setting $m = \lceil 3n/2\rceil$, for all $\delta \in (0,1)$ and for all $t \geq n\left(10.47\ln n - 13.29\ln\delta - 1.49\right)$, we have

$$\mathbb{P}\{\omega'_A(C_t^{(i)}) = n_A, \text{ for all } i = 1, \ldots, n\} \geq 1 - \delta.$$

Proof: Observe that we have

$$\omega'_A(x) = \lfloor n\omega_A(x) + 1/2\rfloor.$$

Applying Theorem 8 with $\varepsilon = 1/(2n)$ and for $t \geq n\left(10.47\ln n - 13.29\ln\delta - 1.49\right)$, we obtain

$$\mathbb{P}\{|n\omega_A(C_t^{(i)}) - n\gamma_A| < 1/2 \text{ for all } i = 1, \ldots, n\} \geq 1 - \delta.$$

Since $n\gamma_A = n_A$ is an integer, we get

$$\mathbb{P}\{\omega'_A(C_t^{(i)}) = n_A, \text{ for all } i = 1, \ldots, n\} \geq 1 - \delta,$$

which completes the proof. ■

Thus each agent can solve the counting problem in $O(\log n)$ parallel time and with $O(n)$ states.

Theorem *10:* Any algorithm solving the counting problem takes an expected $\Omega(\log n)$ parallel time to convergence.

Proof: Solving the counting problem bounds to solving the exact majority problem. By applying Theorem C.1 of [2], this algorithm takes an expected $\Omega(\log n)$ parallel time to convergence under a worst-case input. ■

Theorem *11:* Any algorithm solving the counting problem requires $\Omega(n)$ states.

Proof: To solve the counting problem, the size of the output set Y must be $n+1$. So, the number of states (*i.e.* $|Q|$) is at least $n+1$. The lower bound of the number of states is thus $\Omega(n)$. ■

Theorem *12:* Any algorithm solving the proportion problem with a precision $\varepsilon \in (0,1)$, requires $\Omega(1/\varepsilon)$ states.

Proof: The value of γ_A could be any rational value between 0 and 1, the difference between two output values cannot exceed 2ε, thus the lower bound for the size of the output Y is $\lceil 1/(2\varepsilon) \rceil + 1$. Hence, the number of states (*i.e.*, $|Q|$) is at least $\lceil 1/(2\varepsilon) \rceil + 1$. Thus the lower bound of the number of states is $\Omega(1/\varepsilon)$. ■

VIII. Simulation results

We have conducted simulations to illustrate our theoretical analysis. Figure 1 provides a summary of these simulations. In this figure, each point of the curves represents the mean of 100 simulations (with the maximum and the minimum of the 100 simulations), a simulation consisting in computing the total number of interactions, divided by n, needed for all the agents to converge to γ_A with precision ε. The number n of agents varies from 2^5 to 2^{22}, and the precision ε of the result is set to 10^{-1}, 10^{-3}, and 10^{-5}. Note that as shown theoretically, Figure 1(a) and Figure 1(b) illustrate the fact that the number of interactions per agent to converge is independent of the value of γ, that is independent from the difference between both proportions. From the generated data, for instance when $\delta = 1/2$, one can deduce for each curve an empirical approximation of the convergence parallel time given by $-2\ln\varepsilon + 0.62\ln n - 0.6$.

IX. Conclusion

This paper has shown that in a large-scale system, any agent can compute quickly and with a high precision specified in advance the proportion of agents that initially started in some given input state. This problem is a generalization of the majority problem. Specifically, our protocol guarantees that by using $2\lceil 3/(4\varepsilon) \rceil + 1$ states, any agent is capable of computing the population proportion with precision $\varepsilon \in (0,1)$, in no more than $(-2\ln\varepsilon + 8.47\ln n - 13.29\ln\delta - 2.88)$ interactions with probability at least $1-\delta$, for any $\delta \in (0,1)$. We have also shown that our solution is optimal both in time and space. As future work, we aim at using the same detailed analysis to obtain new results for the majority problem.

References

[1] Dan Alistarh, James Aspnes, David Eisenstat, Rati Gelashvili, and Ronald Rivest. Time-space trade-offs for population protocols. In *arXiv:1602.08032v1*, 2016.

[2] Dan Alistarh, Rati Gelashvili, and Milan Vojnovíc. Fast and exact majority in population protocols. In *Proceedings of the 34th annual ACM symposium on Principles of Distributed Computing (PODC)*, pages 47–56, 2015.

[3] Dan Alistarh, Rati Gelashvili, and Milan Vojnovic. Fast and exact majority in population protocols. In *Proceedings of the 34th annual ACM Symposium on Principles of Distributed Computing (PODC)*, 2015.

[4] Dana Angluin, James Aspnes, Zoë Diamadi, Michael J. Fischer, and René Peralta. Computation in networks of passively mobile finite-state sensors. *Distributed Computing*, 18(4):235–253, 2006.

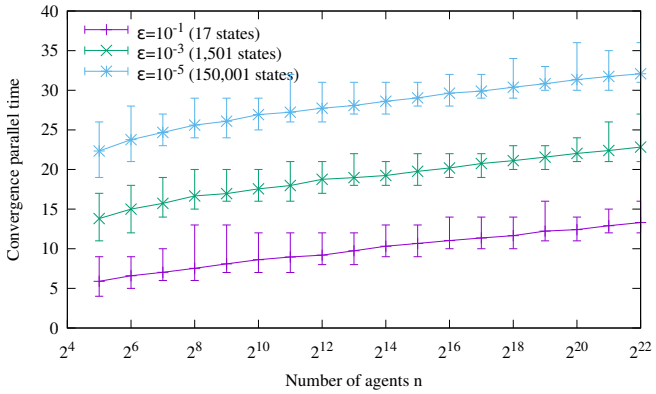
[5] Dana Angluin, James Aspnes, and David Eisenstat. Stably computable predicates are semilinear. In *Proceedings of the 25th annual ACM Symposium on Principles of Distributed Computing (PODC)*, pages 292–299, 2006.

[6] Dana Angluin, James Aspnes, and David Eisenstat. A simple population protocol for fast robust approximate majority. *Distributed Computing*, 20(4):279–304, 2008.

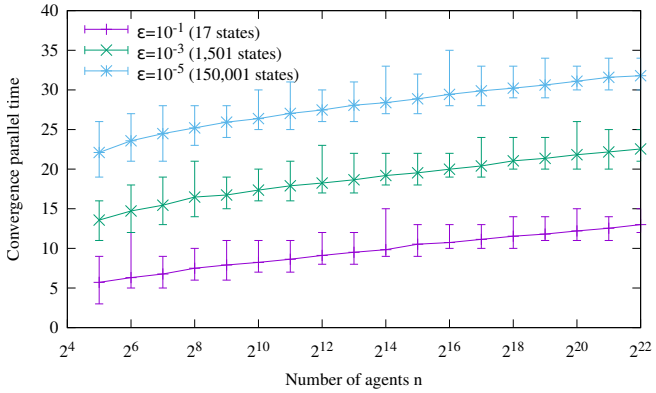
[7] Dana Angluin, James Aspnes, David Eisenstat, and Eric Ruppert. The computational power of population protocols. *Distributed Computing*, 20(4):279–304, 2007.

[8] James Aspnes and Eric Ruppert. An introduction to population protocols. *Bulletin of the European Association for Theoretical Computer Science, Distributed Computing Column*, 93:98–117, 2007.

[9] Joffroy Beauquier, Peva Blanchard, and Janna Burman. Self-stabilizing leader election in population protocols over arbitrary communication graphs. In *Proceedings of the 17th International Conference on Principles of Distributed Systems (OPODIS)*, 2013.



(a) $\gamma = 0$ (*i.e* $\gamma_A = \gamma_B$=1/2)



(b) $\gamma = 1/2$ (*i.e* $\gamma_A = 3/4$ and $\gamma_B = 1/4$)

Figure 1. Number of interactions per agent as a function of the size of the system.

[10] Olivier Bournez, Cohen Johanne, and Mikaël Rabie. Homonym population protocols. In *Proceedings of the 3rd International Conference on Networked Systems (NETYS)*, 2015.

[11] Carole Delporte-Gallet, Hugues Fauconnier, Rachid Guerraoui, and Eric Ruppert. When birds die: Making population protocols fault-tolerant. In *Proceedings of the 2nd IEEE Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 51–66, 2006.

[12] Moez Draief and Milan Vojnovic. Convergence speed of binary interval consensus. *SIAM Journal on Control and Optimization*, 50(3):1087–11097, 2012.

[13] Rachid Guerraoui and Eric Ruppert. Names trump malice: Tiny mobile agents can tolerate byzantine failures. In *Proceedings of the 36th International Colloquium on Automata, Languages and Programming: Part II (ICALP)*, pages 484–495, 2009.

[14] George B. Mertzios, Sotiris E. Nikoletseas, Christoforos Raptopoulos, and Paul G. Spirakis. Determining majority in networks with local interactions and very small local memory. In *Proceedings of the 41st International Colloquium (ICALP)*, pages 871–882, 2014.

[15] Ryu Mizoguchi, Hirotaka Ono, Shuji Kijima, and Masafumi Yamashita. On space complexity of self-stabilizing leader election in mediated population protocol. *Distributed Computing*, 25(6):451–460, 2012.

[16] Yves Mocquard, Emmanuelle Anceaume, James Aspnes, Yann Busnel, and Bruno Sericola. Counting with population protocols. In *Proceedings of the 14th IEEE International Symposium on Network Computing and Applications*, pages 35–42, 2015.

[17] Yves Mocquard, Emmanuelle Anceaume, and Bruno Sericola. Optimal Proportion Computation with Population Protocols. Technical report, August 2016. `https://hal.archives-ouvertes.fr/hal-01354352`

[18] Etienne Perron, Dinkar Vasudevan, and Milan Vojnovic. Using three states for binary consensus on complete graphs. In *Proceedings of the INFOCOM Conference*, pages 2527–2435, 2009.