Noname manuscript No.
(will be inserted by the editor)

Approximate Majority Analyses using Tri-molecular Chemical Reaction Networks

Anne Condon, Monir Hajiaghayi, David Kirkpatrick and Ján Maňuch

the date of receipt and acceptance should be inserted later

Abstract Approximate Majority is a well-studied problem in the context of chemical reaction networks (CRNs) and their close relatives, population protocols: Given a mixture of two types of species with an initial gap between their counts, a CRN computation must reach consensus on the majority species. Angluin, Aspnes, and Eisenstat proposed a simple population protocol for Approximate Majority and proved correctness and $O(\log n)$ time efficiency with high probability, given an initial gap of size $\omega(\sqrt{n}\log n)$ when the total molecular count in the mixture is n. Motivated by their intriguing but complex proof, we provide a new analysis of several CRNs for Approximate Majority, starting with a very simple tri-molecular protocol with just two reactions and two species. We obtain simple analyses of three bi-molecular protocols, including that of Angluin et al., by showing how they emulate the tri-molecular protocol. Our results improve on those of Angluin et al. in that they hold even with an initial gap of $\Omega(\sqrt{n\log n})$.

We prove that our tri-molecular CRN is robust even when there is some uncertainty in the reaction rates, when some molecules are Byzantine (i.e., adversarial), or when activation of molecules is triggered by epidemic. We also analyse a natural variant of our tri-molecular protocol for the more general problem of multi-valued consensus. Our analysis approach, which leverages the simplicity of a tri-molecular CRN to ultimately reason about these many variants, may be useful in analysing other CRNs too.

Keywords Approximate Majority, Chemical Reaction Networks, Population Protocols

The Department of Computer Science, University of British Columbia
E-mail: {condon,monirh,kirk,jmanuch}@cs.ubc.ca

1 Introduction

Stochastic chemical reaction networks (CRNs) and population protocols (PPs) model the dynamics of interacting molecules in a well-mixed solution [1] or of resource-limited agents that interact in distributed sensor networks [2]. CRNs are also a popular molecular programming language for computing in a test tube [3,4]. A central problem in these contexts is Approximate Majority [2,5]: in a mixture of two types of species where the gap between the counts of the majority and minority species is above some threshold, which species is in the majority? Angluin et al. [6] proposed and analysed a PP for Approximate Majority, noting that "Unfortunately, while the protocol itself is simple, proving that it converges quickly appears to be very difficult". Here we provide new, simpler analyses of CRNs for Approximate Majority and several variants. Before describing our contributions, we provide background on the CRN and PP models and the Approximate Majority problem.

1.1 CRNs and Population Protocols

A CRN is specified as a finite set of chemical reactions, such as those in Figures 1 and 2. The underlying model describes how counts of molecular species evolve when molecules interact in a well-mixed solution. Any change in the molecular composition of the system is attributable to a sequence of one or more interaction events that trigger reactions from the specified set. The model is probabilistic at two levels. First, which interaction occurs next, as well as the time between interaction events, is stochastically determined, reflecting the dynamics of collisions in a well-mixed solution [7]. Sec-

ond, an interaction can trigger more than one possible reaction, and rate constants associated with reactions determine the relative likelihood of each outcome. For example, reactions (0'x) and (0'y) of Figure 2(c) are equally likely reactions triggered by an interaction involving one molecule of species X and one of species Y. The method of Soloveichik et al. [8] for simulating CRNs with DNA strand displacement cascades can support such probabilistic reactions.

Angluin et al. [2] introduced the closely related population protocol (PP) model, in which agents interact in a pairwise fashion and may change state upon interacting. Agents and states of a PP naturally correspond to molecules and species of a CRN. A scheduler specifies the order in which agents interact, e.g., by choosing two agents randomly and uniformly, somewhat analogous to stochastic collision kinetics of a CRN. The models differ in other ways. For example, PP interactions always involve two agents, and as such correspond to bi-molecular interactions, while the CRN model allows for interactions of other orders, including unimolecular and tri-molecular interactions. Unlike CRNs, PP interactions may be asymmetric: one agent is the designated initiator and the other is the responder, and their new states may depend not only on their current states but also on their designation. Also, while CRN reaction outcomes may be probabilistic, PP state transition function outcomes are deterministic. Nevertheless, probabilistic transitions can be implemented in PPs by leveraging both asymmetry and the randomness of interaction scheduling [6,9].

1.2 The Approximate Majority Problem

Consider a mixture with n molecules, some of species X and the rest of species Y. Here and throughout, we denote the number of copies of X and Y during a CRN computation by random variables x and y respectively. The *Approximate Majority* problem [6] is to reach *consensus* — a configuration in which all molecules are X ($x = n$) or all are Y ($y = n$), from an initial configuration in which $x + y = n$ and the gap $|x - y|$ is above some threshold. If initially $x > y$, the consensus should be X-majority ($x = n$), and if initially $y > x$ the consensus should be Y-majority. We focus on the case when initially $x > y$ since the CRNs that we analyse are symmetric with respect to X and Y.

The reactions of the simple tri-molecular CRN of Figure 1 involve just the two species X and Y that are present initially. Intuitively, the tri-molecular reactions sample triples of molecules and amplify the majority species by exploiting the facts that (i) every triple must have a majority of either X or Y, and (ii) for healthy populations of X and Y, the ratio of the number of triples with two X-molecules and one Y-molecule to the number of triples with two Y-molecules and one X-molecule, is essentially the ratio of X-molecules to Y-molecules.

It is also natural to consider protocols and CRNs in which interactions may involve just two agents or molecules. Angluin et al. [10] proposed and analysed the Single-B CRN of Figure 2(c). Informally, reactions (0'x) and (0'y) are equally likely to produce B's (blanks) from X's or Y's respectively, while reactions (1') and (2') recruit B's to become X's and Y's respectively. (Angluin et al. described this as a population protocol, using asymmetry, that provides 1/2 rates, and the randomness of the scheduler to implement the random reactions (0'x) and (0'y).) When X is initially in the majority ($x > y$ initially), a productive reaction event (i.e. one that results in some change in the chemical composition) is more likely to be (1') than (2'), with the bias towards (1') increasing as x gets larger. Angluin et al. showed *correctness*: if initially $x - y = \omega(\sqrt{n} \log n)$, then with "high" probability $1 - n^{-\Omega(1)}$, Single-B reaches X-majority consensus. They also showed *efficiency*: with high probability for any initial gap value $x - y$, Single-B reaches consensus within $O(n \log n)$ interaction events. They also proved (i) correctness and efficiency in "initiation by infection" protocols, in which agents are initially inactive and only participate in the protocol upon receipt of a "wake-up" signal, and (ii) correctness with respect to a (slightly but unavoidably relaxed) form of consensus, as well as efficiency, in protocols with $o(\sqrt{n})$ Byzantine agents. Several others have subsequently and independently studied the problem; we'll return to related work after describing our own contributions.

1.3 Our Contributions

After describing the CRN model and our analysis tools in Section 2, in Section 3 we analyse Tri, the tri-molecular CRN of Figure 1. Our primary motivation is to provide the simplest and most intuitive proof of correctness and efficiency that we can, with the goal of adapting our techniques to reason about other variants, as well as CRNs for other problems. We show that consensus is reached within $O(\gamma n \lg n)$ interaction events, with probability $1 - \exp(-\Omega(\gamma \lg n))$ and, if initially the gap between the majority and minority species is at least $\sqrt{\gamma n \lg n}$ then, with the same high probability, the consensus is on the majority species.

When the initial gap is at least $\sqrt{\gamma n \lg n}$, we analyse Tri in three phases. In the first phase we model the evolution of the gap $x - y$ as a sequence of random walks (referred to as *stages*) with increasing bias of success

$$X+X+Y \to X+X+X \quad (1)$$
$$X+Y+Y \to Y+Y+Y \quad (2)$$

Fig. 1 Tri, a tri-molecular chemical reaction network (CRN) for Approximate Majority. Reactions have rate constant 1.

		$X+Y \overset{1/2}{\to} X+B$ (0'x)
$X+Y \to \mathbb{B}$ (0')	$X+Y \to B+B$ (0')	$X+Y \overset{1/2}{\to} Y+B$ (0'y)
$\mathbb{B}+X \to X+X+X$ (1')	$X+B \to X+X$ (1')	$X+B \to X+X$ (1')
$\mathbb{B}+Y \to Y+Y+Y$ (2')	$Y+B \to Y+Y$ (2')	$Y+B \to Y+Y$ (2')
(a) Heavy-B	(b) Double-B	(c) Single-B

Fig. 2 Three bi-molecular chemical reaction networks (CRNs) for Approximate Majority. Reactions (0'x) and (1'y) of Single-B have rate constant 1/2 while all other reactions have rate constant 1.

(i.e., increase in $x - y$). Similarly, in the second phase we model the evolution of the count of y as a sequence of random walks with increasing bias of success (decrease in y). We use a simple biased random walk analysis to show that these walks make forward progress with high probability, thereby ensuring correctness. To show efficiency, we model each random walk as a sequence of independent trials, observe a natural lower bound on the probability of progress, and apply Chernoff bounds. In the third and last phase we model the "end game" as y decreases from $\Theta(\log n)$ to 0, and apply the random walk analysis and Chernoff bounds a final time to show correctness and efficiency, respectively. We then extend our analysis to small initial gaps, showing that consensus is reached efficiently with high probability, although the consensus may not be on the majority species.

In Section 4 we analyse the three bi-molecular CRNs of Figure 2 by relating them to the tri-molecular CRN. In each of these protocols, blanks are in a natural sense a proxy for $X + Y$ (an interaction between X and Y), and so reactions (1') and (2') behave exactly like the corresponding reactions of our tri-molecular CRN. The Heavy-B CRN in part (a) of the figure simply uses a blank species as a proxy for $X + Y$. The total species count can vary as a result, though can never be less than half of the initial species count. The Double-B CRN of part (b) uses two blanks to represent $X + Y$, thereby keeping the total molecular count constant. Double-B is implementable and symmetric even in the PP setting, and was among the earliest CRN algorithms constructed with strand displacement chemistry, by Chen et al. [11]. The Single-B CRN of part (c) is essentially the same as that of Angluin et al. [6].

In subsequent sections, we build on our analysis of Tri to show correctness and efficiency of other CRNs for the Approximate Majority problem or for more general variants of the problem. The table in Figure 3 provides a summary. In Section 5 we consider the case where reaction rate constants of the tri-molecular CRN may fluctuate in unpredictable ways over the course of a computation. One motivation for analysing this scenario arises when a CRN is "compiled" to a DNA strand displacement system [8]. It may be that the DNA strand displacement reaction rate constants closely approximate, but are not exactly equal to, the CRN reaction rates. For a given lower bound on the relative rates of reactions (1) and (2), we determine an initial gap that is sufficient to ensure that our analysis of the tri-molecular CRN goes through. In turn, we show in Section 6 how the uncertain rates analysis naturally leads to an analysis of the Byzantine case (with relaxed consensus), when the initial population of X is $(n + \Delta_0)/2 \geq (n + \Omega(\sqrt{n \lg n}))/2$ and $z \leq \Delta_0/16$ molecules behave adversarially. In Section 7 we analyse the natural generalization of our CRN when there are three or more species and the goal is to reach consensus on the plurality (most populous) species. Once again we can leverage our analysis of the tri-molecular protocol for two species, by first analysing a period in which the initial plurality species grows to be greater than the sum of all the others, and then showing that the computation proceeds in essentially the same way as the original protocol. In Section 8 we analyse a variant of Tri for a scenario also considered by Angluin et al. in which molecules become active only when a single distinguished molecule initiates an infection.

We present empirical results in Section 9, and in Section 10 we conclude with directions for future work.

The results presented in this paper extend our earlier conference paper [12]. We have simplified our analyses of the Double-B and Single-B bi-molecular protocols, and added the Heavy-B protocol as an additional AM bimolecular CRN. Our sections on uncertain rate constants, Byzantine behaviour, initiation by infection, and multi-valued consensus are all new and we have expanded our empirical results accordingly.

Problem	**CRNs**	**Initial gap Δ_0**	**# Interaction events**
Approximate Majority	Tri Heavy-B Double-B Single-B	$\sqrt{\gamma n \lg n}$	$\Theta(\gamma n \lg n)$
Approximate Majority with Uncertain Rate Constants (1, $\alpha < 1$)	Tri(1,α)	$\Theta(\frac{1-\alpha}{1+\alpha} n)$	$\Theta(\gamma n \lg n)$
Relaxed Approximate Majority with $\leq \Delta_0/16$ Byzantine Molecules	Byzantine-Tri	$\sqrt{\gamma n \lg n}$	$\Theta(\gamma n \lg n)$
m-Species Plurality Consensus	m-species-Tri	$\sqrt{\gamma n \lg n}$	$\Theta(\gamma m n \lg n)$
Approximate Majority with Initiation by Infection	Tri-with-Infection	$\Theta(\sqrt{\gamma n \lg n})$	$\Omega(\gamma n \lg n)$

Fig. 3 Our results for CRNs for the Approximate Majority (AM) problem and variants. An initial gap lower bound Δ_0, between the count of the majority (or plurality) species and the count of any other species, is specified as a function of n, the initial number of molecules, and a positive constant γ. For all CRNs the error probability is $\exp(-\Omega(\gamma \lg n))$. See Section 3.2 for our analysis of the Tri CRN with a small initial gap, which is not included in this table.

1.4 Comparison with Related Work

Angluin et al.'s work on the Single-B protocol [6] stimulated much interest in CRNs and PPs for approximate majority. Our analyses improve slightly on theirs in some cases. Their majority-consensus analysis of Single-B assumes an initial gap of $\sqrt{n} \lg n$, while ours is $\sqrt{n \lg n}$. For the Byzantine case their result, assuming the same initial excess of the majority species, holds for $o(\sqrt{n})$ Byzantine agents while ours holds for a population of Byzantine molecules that is proportional to the initial excess of the majority species, which is $\Omega(\sqrt{n \lg n})$ and could be $\Theta(n)$. Their result on protocols that are initiated by infection requires an initial gap of $\Omega(n^{3/4+\epsilon})$, compared with our initial gap of $\Omega(\sqrt{n \lg n})$. Their protocol for multi-valued consensus applies Approximate Majority in a bit-wise fashion; they do not analyse an m-species protocol such as ours. Angluin et al. also analyse Single-B in the case when the initial gap is $o(\sqrt{n \log n})$, showing that Single-B reaches consensus with high probability, though not necessarily on the Majority species. We also provide a simpler proof of this result.

In contrast with our asynchronous model, other related work focuses on a synchronous model in which n participating agents (corresponding to molecules) update their states in rounds, with each of the n participating agents initiating one interaction that involves a constant number of additional processes chosen uniformly at random. For this synchronous model, Doerr et al.'s [13] "median rule" consensus protocol involves rules that are identical to the interactions of the Tri CRN of Figure 1 in the case of two types of agents (species). Their analysis shows that the protocol efficiently reaches stable consensus with high probability in the m-species case, from an arbitrary initial configuration. They also show that a weaker form of consensus is reached when at most $\sqrt{n}$ Byzantine faults occur per round. The result of each synchronous round is very similar to what is accomplished in one time unit of the CRN or PP models, in which a sequence of n random interactions occur. Accordingly there are strong similarities between our analysis and that of Doerr et al. For example, our analysis is staged in a way that allows us to assume that interactions within each stage are driven by essentially the same population sizes. Note however that in our CRN model, unlike the Doerr et al. model, there may be molecules that participate in no interaction within a given unit of time. This difference becomes evident in our end game analysis, which requires $\Theta(n \log n)$ time units to ensure that, with high probability, the few remaining minority species interact and thus are converted to the majority species. In contrast, the end game is completed in $O(1)$ rounds with high probability in the synchronous model.

The median rule of Doerr et al. may not converge on the plurality species, even if there is a large initial gap between the count of this species and the count of any other species. For this reason, Beschetti et al. [14] analyse a 3-majority protocol that is quite similar to the m-species Tri CRN when the number m of species is at least three. For the synchronous model they show that for any $m \leq n$, for some constant $c > 0$, consensus on the plurality species is reached with probability at least $1 - n^{-c}$ in $O(\min\{m, (n/\lg n)^{1/3}\} \lg n)$ rounds when the initial gap between the plurality species and the re-

maining species is $\Omega(\sqrt{\min\{m,(n/\lg n)^{1/3}\}n\lg n})$. In subsequent work, Becchetti et al [15] show, again with respect to the synchronous model, that if $m \leq n^{\alpha}$ where α is a positive constant, their protocol reaches consensus with high probability, even from an arbitrary initial configuration, in time polynomial in m and $\log n$. Their results hold for relaxed consensus in the presence of $o(\sqrt{n})$ Byzantine faults per round. Apart from the fact that Becchetti et al.'s analysis is for a different model than ours (synchronous vs asynchronous), there are several differences in the details of their results and ours. For example, they require an initial gap that depends on m while ours is independent of m. Neither Doerr et al. nor Becchetti et al. consider protocols in which interactions involve just two agents, uncertain reaction rates, or initiation by infection.

Perron et al. [16] analyse Single-B when $x + y = n$ and $y \leq \epsilon n$. They use a biased random walk argument to show that Single-B reaches consensus on X-majority with exponentially small error probability $1 - e^{-\Theta(n)}$. The results of Perron et al. do not apply to smaller initial gaps. Mertzios et al. [17] showed somewhat weaker results for Single-B when initially $x - y \geq \epsilon n$ (the main focus of their paper is when interactions are governed by a more general interaction network). Cruise and Ganesh [18] devise a family of protocols in network models where agents (nodes) can poll other agents in order to update their state. Their family of protocols provides a natural generalization of our tri-molecular CRN and their analysis uses interesting connections between random walks and electrical networks.

Yet other work on Approximate Majority pertains to settings with different assumptions about the number of states per agent, the types of interaction scheduling rules, and possibly adversarial behaviour, or analyses more general multivalued consensus problems [9, 10, 19].

2 Preliminaries

2.1 Chemical Reaction Networks

Let $\mathcal{X} = \{X_1, X_2, \ldots, X_m\}$ be a finite set of *species*. A solution *configuration* $c = (x_1, x_2, \ldots, x_m)$, where the x_i's are non-negative integers, specifies the number of molecules of each species in the mixture. Molecules in close proximity are assumed to interact. We denote an *interaction* that simultaneously involves $s_i \geq 0$ copies of X_i, for $1 \leq i \leq m$, by a vector $s = (s_1, s_2, \ldots, s_m)$, and define the *order* of the interaction to be $s_1 + s_2 + \ldots + s_m$.

We model interacting molecules in a well-mixed solution, under fixed environmental conditions such as temperature. The well-mixed assumption has two important implications that allow us to draw on aspects of both CRN models [1,3,20] and also PP models [2], aiming to serve as a bridge between the two. The first, that all molecules are equally likely to reside in any location, supports a stochastic model of chemical kinetics, in which the time between molecular interactions of fixed order is a continuous random variable that depends only on the number of molecules and the volume of the solution. The second, that any fixed interaction is equally likely to involve any of the constituent molecules, and is therefore sensitive to the counts of different species, supports a discrete, essentially combinatorial, view of interactions reminiscent of, but more general than, those in standard PP models. In Section 2.3 we compare our model with that of Cook et al. [3].

In this paper we will only be interested in interactions of a single order (either two or three), in a fixed volume. According to a stochastic model of chemical kinetics [1], at any moment, the *time* until the next interaction of order o, what we refer to as an *interaction event*, occurs is exponentially distributed with parameter $\binom{n'}{o}/v^{o-1}$, where n' denotes the total number of molecules in the system and v denotes the total volume of the solution. Consequently, the expected time between interaction events of order o is $v^{o-1}/\binom{n'}{o}$ and the variance is $(v^{o-1}/\binom{n'}{o})^2$. It follows that, if n is the total number of molecules initially in the system, and at all times $n' = \Theta(n)$ and $v = \Theta(n)$ (as will be the case for all CRNs studied in this paper), the time T_n for n interaction events has expected value $E[T_n] = \Theta(n^o/\binom{n}{o}) = \Theta(1)$ and variance $\mathrm{Var}[T_n] = \Theta((n^o/\binom{n}{o})^2/n) = \Theta(1/n)$. By Chebyshev's inequality, we have that:
$\mathbb{P}[|T_n - \mathrm{E}[T_n]| \geq h\sqrt{\mathrm{Var}[T_n]}] = \mathbb{P}[|T_n - n^o/\binom{n}{o}| \geq h(n^o/\binom{n}{o})/\sqrt{n}] \leq 1/h^2$. By setting $h = \sqrt{n}$ we see that the time for n interaction events is $\mathrm{O}(1)$ with probability at least $1 - 1/n$. Thus we use the number of interaction events, divided by n, as a proxy for time.

When the solution is in configuration $c = (x_1, x_2, \ldots, x_m)$ where $\sum_i x_i = n$, the well-mixed property dictates that the probability that a given interaction event of order o is the particular interaction $s = (s_1, s_2, \ldots, s_m)$ is $\lambda(c, s) = \left[\prod_{i=1}^{m} \binom{x_i}{s_i}\right] / \binom{n}{o}$.

Some interaction events lead to an immediate change in the configuration of the solution, while others do not. The change (possibly null) arising from an interaction can be described as a (possibly unproductive) reaction event. Formally, a *reaction* $r = (s, t) = ((s_1, s_2, \ldots, s_m), (t_1, t_2, \ldots, t_m))$ is a pair of non-negative integer vectors describing reactants and products, where, for *productive* reactions, some $s_i > 0$ and for at least one

i, $s_i \neq t_i$. Reaction r is *applicable* in configuration $c = (x_1, x_2, \ldots, x_m)$ if $s_i \leq x_i$, for $1 \leq i \leq m$. If reaction r occurs in configuration c, the new configuration of the mixture is $c' = (x_1 - s_1 + t_1, x_2 - s_2 + t_2, \ldots, x_m - s_m + t_m)$. In this case we say that the transition from configuration c to configuration c' is *realized* by reaction r and we write $c \Rightarrow^r c'$. Each reaction r has an associated rate constant $0 < k_r \leq 1$, specifying the probability that the reaction is consummated, given the interaction specified by the reactant vector is satisfied, so the probability that reaction $r = (s, t)$ occurs as the result of an interaction event in a configuration c is just $k_r \lambda(c, s)$.

A *chemical reaction network (CRN)* is a pair $(\mathcal{X}, \mathcal{R})$, where $\mathcal{X}$ is a finite set of species and $\mathcal{R}$ is a finite set of productive reactions, such that, for all reactant vectors s, if $\mathcal{R}(s)$ is the subset of $\mathcal{R}$ with reactant vector s, then $\sum_{r \in \mathcal{R}(s)} k_r \leq 1$. To ensure that all interactions have a fully specified outcome, we take as implicit in this formulation the existence, for every reactant vector s, including all possible interactions of order o, of a non-productive reaction with rate constant $1 - \sum_{r \in \mathcal{R}(s)} k_r$.

2.2 CRN Computations

Next we describe how the mixture of molecules evolves when reactions of a CRN $(\mathcal{X}, \mathcal{R})$ occur. For most of the CRNs that we analyse, there is some order o such that for every reaction (s, t) of $\mathcal{R}$, $s_1 + s_2 + \ldots s_m = t_1 + t_2 + \ldots t_m = o$. (As we shall see, this assumption is relaxed in Section 4, where the necessary modification to the model is discussed separately.) Thus the number n of molecules in the system does not change over time. We furthermore assume that the volume v of the solution is fixed and proportional to n.

A random sequence of interaction events triggers a sequence of (not necessarily productive) reaction events, reflected in a sequence of configurations that we interpret as a computation. More formally, a *computation* of the CRN $(\mathcal{X}, \mathcal{R})$, with respect to an initial configuration c_0, is a discrete Markov process whose states are configurations. The probability, denoted $\pi(c, c')$, of a transition, via a reaction event, from configuration c to configuration c' is just the sum of the probabilities of all reactions r such that $c \Rightarrow^r c'$.

Furthermore, the probability, denoted $\pi^*(c_0, c_{\text{final}})$, that a sequence of transitions from configuration c_0 reaches configuration c_{final}, is just the sum, over all configuration sequences
$c_0, c_1, \ldots, c_t = c_{\text{final}}$ where $c_i \neq c_{\text{final}}$ for $0 \leq i \leq t-1$, of $\Pi_{i=1}^{t} \pi(c_{i-1}, c_i)$.

2.3 Relationship between our CRN model and that of Cook et al.

Other CRN models define reaction probabilities and computation time somewhat differently than we do, but these differences can easily be reconciled. For example, in the model of Cook et al. [3], if k'_r is the rate constant associated with reaction $r = (s, t)$ of order o and the system is in configuration $c = (x_1, x_2, \ldots, x_m)$, then the propensity, or rate, of r is

$$\rho_r(c) = k'_r[\prod_{i=1}^{m}(x_i!/(x_i - s_i)!)]/v^{o-1}.$$

If $\rho^{tot}(c) = \sum_r \rho_r(c)$ for all reactions r of order o, then the probability that a reaction event is reaction r is $\rho_r(c)/\rho^{tot}(c)$, and the expected time until a reaction event occurs is $1/\rho^{tot}(c)$. (In this model, reaction rate constants can be greater than 1, and may depend not only on the number of reactants of each species, but also on other properties of a species such as its shape, capturing the fact that the likelihood of different types of interactions may not all be the same.)

If in our model we set $k_r = k'_r \prod_{i=1}^{m} s_i!$ for each productive reaction, and normalize by $\sum_r k_r$ if necessary to ensure that $\sum_{r \in \mathcal{R}(s)} k_r \leq 1$ (adjusting the underlying time unit accordingly), a straightforward calculation shows that, when in a given configuration c, the probability that a reaction event is a given reaction r is the same in our model and that of Cook et al. [1] See

[1] Here is the calculation for the probability conversion.

$$\begin{aligned}
\rho_r(c) &= k'_r.[\prod_{i=1}^{m}(x_i!/(x_i - s_i)!)]/v^{o-1} \\
&= k'_r.[\prod_{i=1}^{m} s_i!].[\prod_{i=1}^{m}\binom{x_i}{s_i}]/v^{o-1} \\
&= [\binom{n}{o}/v^{o-1}]k'_r[\prod_{i=1}^{m} s_i!].[\prod_{i=1}^{m}\binom{x_i}{s_i}]/\binom{n}{o} \\
&= [\binom{n}{o}/v^{o-1}]k_r[\prod_{i=1}^{m}\binom{x_i}{s_i}]/\binom{n}{o},
\end{aligned}$$

where

$$k_r = [k'_r[\prod_{i=1}^{m} s_i!]. \tag{1}$$

We can interpret the last of these expressions for $\rho_r(c)$ as the product of three terms. The first term, namely $\binom{n}{o}/v^{o-1}$, corresponds to the (normalized) average rate of an interaction of order o. The last term, namely $[\prod_{i=1}^{m}\binom{x_i}{s_i}]/\binom{n}{o}$, is the probability that the reaction of order o has exactly the reactants of r. The middle term k_r depends on the s_i's, but could also model situations where different types of interactions have different rates, e.g., if some molecular species are larger than others. Normalizing the k_r's by $\sum k_r$ yields rate constants for our model.

the example of Figure 4. Also, the expected time until the next reaction event differs between the models by a constant factor that is independent of c. Conversely, to convert from our model to that of Cook et al., divide our rate constant k_r by $[\prod_{i=1}^{m} s_i!]$ and multiply all rate constants by the same constant factor in order to adjust time units as needed.

2.4 Analysis Tools

We will use the following well-known property of random walks, as well as familiar Chernoff tail bounds on the sum of independent random variables.

Lemma 1 (Asymmetric one-dimensional random walk [21](XIV.2)) *If we run an arbitrarily long sequence of independent trials, each with success probability at least p, then the probability that the number of failures ever exceeds the number of successes by b is at most $(\frac{1-p}{p})^b$.*

Lemma 2 (Chernoff tail bounds [22]) *If we run N independent trials, with success probability p, then S_N, the number of successes, has expected value $\mu = Np$ and, for $0 < \delta < 1$,*
(a) $\mathbb{P}[S_N \leq (1-\delta)\mu] \leq \exp(-\frac{\delta^2\mu}{2})$, and
(b) $\mathbb{P}[S_N \geq (1+\delta)\mu] \leq \exp(-\frac{\delta^2\mu}{3})$.

3 Approximate Majority Using Tri-molecular Reactions

In this section we analyse the behaviour of the tri-molecular CRN of Figure 1. We prove the following:

Theorem 1 *(a) For any constant $\gamma \geq 1$, a computation of the tri-molecular CRN reaches a consensus in $O(\gamma n \lg n)$ interaction events, with probability $1-\exp(-\Omega(\gamma \lg n))$. (b) Furthermore, provided the initial molecular count of X exceeds that of Y by at least $\sqrt{\gamma n \lg n}$ this consensus is X-majority, with probability $1-\exp(-\Omega(\gamma \lg n))$.*

We first address our primary concern (part (b)), the situation where the initial molecular count of X exceeds that of Y by at least $\sqrt{\gamma n \lg n}$ and we want to reach a consensus of X-majority.

3.1 Initial gap at least $\sqrt{\gamma n \lg n}$

Recall that we denote by x and y the random variables corresponding to the molecular count of X and Y respectively. We note that the probability that an interaction event triggers reaction (1) (respectively, reaction (2)) is just $\binom{x}{2}y/\binom{n}{3}$ (respectively, $\binom{y}{2}x/\binom{n}{3}$). Hence, the probability that an interaction triggers one of these (a productive reaction event) is $xy(x+y-2)/(2\binom{n}{3})$, and the probability that such a reaction event is reaction (1) is $(x-1)/(x+y-2) \geq x/n$, provided $x \geq y$.

We assume that n is sufficiently large (in particular $\gamma \lg n \leq n/6$), and divide the computation into a sequence of three, possibly degenerate, phases:

phase 1: It starts with $x - y \geq \sqrt{\gamma n \lg n}$ and continues while $\sqrt{\gamma n \lg n}/2 \leq x - y < 2n/3$. It *completes properly* if $x - y \geq 2n/3$ (equivalently, $y \leq n/6$).
phase 2: It starts with $y \leq n/6$ and continues while $n/3 \geq y > \gamma \lg n$. It *completes properly* if $y \leq \gamma \lg n$.
phase 3: It starts when $y \leq \gamma \lg n$ and continues while $2\gamma \lg n \geq y > 0$. It *completes properly* if $y = 0$.

Of course the assertion that a computation can be partitioned in such a way that these phases occur in sequence (i.e., they all complete properly) holds only with sufficiently high probability. To prove this assertion and to analyse the efficiency of the phases, we divide both phase 1 and phase 2 into $\Theta(\lg n)$ stages, defined by integral values of t and s, as follows:

- A typical stage t in phase 1 starts with $x - y \geq 2^t\sqrt{\gamma n \lg n}$ and continues while $2^{t-1}\sqrt{\gamma n \lg n} \leq x - y < 2^{t+1}\sqrt{\gamma n \lg n}$. It *completes properly* if $x - y \geq 2^{t+1}\sqrt{\gamma n \lg n}$.
- A typical stage s in phase 2 starts with $y \leq n/2^s$ and continues while $n/2^{s-1} \geq y > n/2^{s+1}$. It *completes properly* if $y \leq n/2^{s+1}$.

Our proof of correctness (all phases/stages complete properly) and our efficiency analysis (how many interaction events does it take to realize the required number of productive reaction events for proper phase/stage completions) exploit the simple and familiar tools set out in the previous section, taking advantage of bounds on the probability of reactions (1) and (2) that hold throughout a given phase/stage:

(a) [High probability of proper phase/stage completion within a small number of productive reaction events] Within a fixed phase/stage the computation can be viewed as a sequence of independent trials (choice of reaction (1) or (2)) with a fixed lower bound on the probability of success (choice of reaction (1)). This allows us to establish, by a direct application of Chernoff's upper tail bound Lemma 2(a), an upper bound, for each phase/stage, on the probability that the phase/stage completes improperly within a specified number of productive reaction events.
(b) [High probability that the productive reaction events occur within a small number of molecular interactions] Within a fixed phase/stage the choice of productive reaction events, among interaction events,

$$X + X + Y \xrightarrow{1} X + X + X \ (r_1) \qquad Y + Y + Y \xrightarrow{2} X + X + Y \ (r_2)$$
(a)

$$X + X + Y \xrightarrow{2/14} X + X + X \ (r_1) \qquad Y + Y + Y \xrightarrow{12/14} X + X + Y \ (r_2)$$
(b)

Fig. 4 (a) A CRN specified with respect to the Cook et al. model. The reaction rates when the system is in configuration (3,3) are $k'_{r_1} = 18/v^2$ and $k'_{r_2} = 12/v^2$. The reaction probabilities are $\rho_{r_1}(3,3) = 3/5$ and $\rho_{r_2}(3,3) = 2/5$. (b) The mapping of the CRN of part (a) to our model by changing the rate constants (using Equation 1 of footnote 1) and normalizing by $\sum k_r$. The probability that a reaction event is r_1 is $(18/14)/(30/14) = 18/30$, and the probability of r_2 is $12/30$. Thus, reaction probabilities are preserved exactly.

can be viewed as a sequence of independent trials with a fixed lower bound on the probability of success (the interaction corresponds to a productive reaction event). Thus our timing analysis (proof of efficiency) is another direct application of Chernoff's upper tail bound (Lemma 2(a)).

Lemma 3 *At any point in the computation, if* $0 < x - y = \Delta < n$*, then, with probability* $1 - \exp(-\Omega(\Delta^2/n))$*,* $x - y$ *increases to* $\min\{2\Delta, n\}$ *within* $2n$ *productive reaction events.*

Proof Since $x - y \geq \Delta/2$ up to the point (if ever) when we first have $x - y < \Delta/2$, it follows that the probability that a productive reaction is reaction (1) is at least $\frac{\binom{x}{2}y}{\binom{x}{2}y + x\binom{y}{2}} > \frac{x}{n} \geq \frac{1}{2} + \frac{\Delta}{4n}$. Thus, we can view the change in $x - y$ resulting from productive reaction events as a random walk, starting at Δ, with success (an increase in $x - y$ by two) probability $p > \frac{1}{2} + \frac{\Delta}{4n}$.

Since $\frac{1-p}{p} < \frac{2n-\Delta}{2n+\Delta} = 1 - \frac{2\Delta}{2n+\Delta}$, it follows from Lemma 1 that reaching a configuration where $x - y < \Delta/2$ (which entails an excess of $\Delta/4$ failures to successes) is less than $(1 - \frac{2\Delta}{2n+\Delta})^{\Delta/4}$ which is at most $\exp(-\Delta^2/(4n + 2\Delta))$.

As long as $x - y$ is at least $\Delta/2$, the probability that it fails to be increased to $\min\{2\Delta, n\}$ within $2n$ productive reactions is just the probability that a sequence of $2n$ independent trials with success probability $p > 1/2 + \Delta/(4n)$ results in fewer than $n + \Delta/4$ successes, i.e., at least $\Delta/4$ fewer than expected. By Lemma 2(a), this probability is at most $\exp(-\Delta^2/(32n + 16\Delta))$.

Corollary 1 *Since* $x - y \geq \sqrt{\gamma n \lg n}$ *at the start of each stage of phase 1, every such stage completes properly within at most* $2n$ *productive reaction events, with probability at least* $1 - \exp(-\Theta(\gamma \lg n))$*.*

Lemma 3 shows that progress, in the form of a doubling of the gap between the majority and minority species, happens with high probability within $\Theta(n)$ productive reaction events. The next lemma shows that progress, in the form of a reduction of the population of the minority species, occurs at a rate proportional to the population of that species, provided that population is bounded away from $n/2$ (the extent to which is captured by the parameter β).

Lemma 4 *Let* $1 < \beta \leq 2$ *be any constant. At any point in the computation, if* $y = n/k$*, where* $1 + \beta \leq k \leq n/(\gamma \lg n)$*, then, with probability* $1 - \exp(-\Omega((\beta - 1)\gamma \lg n))$*,* y *decreases to* 0 *within* $2n/(k-2)$ *productive reaction events.*

Proof Let $d = \gamma \lg n$. Since $y \leq n/k + d$ up to the point (if ever) when we first have $y > n/k + d$, it follows that the probability that a productive reaction is reaction (1) is at least $\frac{x}{n} \geq 1 - \frac{n-kd}{kn} > 1 - 1/k$. Thus we can view the change in y resulting from productive reaction events as a random walk, starting at n/k, with success (a decrease in y by one) probability p satisfying $p > 1 - 1/k$. It follows from Lemma 1 that reaching a configuration where $y > n/k + d$ (which entails an excess of d failures to successes) is less than $(\frac{1-p}{p})^d \leq (\frac{1}{k-1})^d$, which is $\exp(-\Omega((\beta - 1)\gamma \lg n))$, since $\lg \beta \geq (\beta - 1)$, for $1 < \beta \leq 2$.

Let $\lambda = 2k/(k-2)$. As long as y remains no larger than $n/k + d$ then the probability that it fails to decrease to 0 within $\lambda n/k$ productive reactions is just the probability that a sequence of $\lambda n/k$ independent trials with success probability $p > 1 - 1/k = 1/2 + 1/\lambda$ results in fewer than $(\lambda + 1)n/(2k)$ successes, i.e., at least $n/(2k)$ fewer than expected. It follows, by Lemma 2(a), that this probability is at most $\exp(-\Theta(n/(\lambda k)))$, which is $\exp(-\Omega((\beta - 1)\gamma \lg n))$ since $\lambda = \Theta(1/(\beta - 1))$.

Corollary 2 *Since stage* s *of phase 2 starts with* $y \leq n/2^s \leq n/6$ *and ends with* $y = \min\{n/2^{s+1}, \gamma \lg n\}$*, it completes properly, within at most* $3n/2^s$ *productive reaction events, with probability at least* $1 - \exp(-\Theta(\gamma \lg n))$*.*

Corollary 3 *Since phase 3 starts with* $y \leq \gamma \lg n \leq n/6$ *it completes properly, within at most* $3\gamma \lg n$ *productive reaction events, with probability at least* $1 - \exp(-\Theta(\gamma \lg n))$*.*

Lemma 5
(i) The at most $2n$ *productive reaction events of each stage of phase 1 occur within* $\Theta(n)$ *interaction events,*

with probability $1-\exp(-\Omega(n))$.
(ii) The at most $3n/2^s$ *productive reaction events of stage* s *of phase 2 occur within* $\Theta(n)$ *interaction events, with probability* $1-\exp(-\Omega(n/2^s))$.
(iii) The at most $3\gamma \lg n$ *productive reaction events of phase 3 occur within* $\Theta(\gamma n \lg n)$ *interaction events, with probability* $1-\exp(-\Omega(\gamma \lg n))$.

Proof It is an immediate consequence of Lemma 2(a) that, if during some sequence of r interaction events the probability that interaction events are productive is at least p, then the probability that the sequence gives rise to fewer than $rp/2$ productive reaction events is no more than $\exp(-rp/8)$. Since at any one time the probability that an interaction event results in a productive reaction is $\frac{\binom{x}{2}y+x\binom{y}{2}}{\binom{n}{3}} > \frac{3xy}{n^2}$, it suffices to observe the following lower bounds on the product xy in individual phases/stages:
(i) in phase 1, $n/2 < x < 7n/8$, so $xy > 7n^2/64$;
(ii) in stage s of phase 2, $y > n/2^{s+1}$, so $xy > \frac{2^{s+1}-1}{2^{2s+2}}n^2$;
(iii) in phase 3, $y \geq 1$ and so $xy \geq n-1$.

Finally, we prove Theorem 1(b), for initial population gaps of size at least $\sqrt{\gamma n \lg n}$, using the pieces proved until now.

Proof (of Theorem 1(b))
(i) [Correctness] It follows directly from Corollaries 1, 2 and 3 that all phases, including their $\Theta(\lg n)$ sub-stages, complete properly, with probability $1-\exp(-\Omega(\gamma \lg n))$.
(ii) [Efficiency] It is immediate from Lemma 5 that a total of $\Theta(\gamma n \lg n)$ interaction events suffice to ensure the proper completion of all phases/stages, with probability $1-\exp(-\Omega(\gamma \lg n))$.

3.2 Small initial gap

To prove Theorem 1(a) it remains to argue that even with an arbitrarily small initial gap our tri-molecular CRN reaches a consensus in $O(\gamma n \lg n)$ interaction events, with probability $1-\exp(-\Omega(\gamma \lg n))$. Of course, given the results of the preceding subsection, it suffices to show that a gap of at least $\sqrt{\gamma n \lg n}$ is reached, within $O(\gamma n \lg n)$ interaction events, with probability $1-\exp(-\Omega(\gamma \lg n))$.

In previous analyses, we viewed the evolution of the gap between the populations of X and Y, effected by single productive reactions, as a random walk on the state set $\{2i \mid i \in \mathbb{N}\}$. For our current purposes it is helpful to distinguish a subset of *checkpoint* states, namely $\{g_j \mid j \in \mathbb{N}\}$, where $g_0 = 0$ and $g_j = 2^{j+3}\sqrt{n}$, for $j \geq 1$ and to consider transitions between adjacent such states, effected by a sequence of productive reactions. [2]

It is straightforward to confirm that: (i) with probability at least 3/4, starting from a configuration with population gap g_0, the population gap is increased to g_1 within $2^{10}n$ productive reactions; and (ii) for $j \geq 1$, with probability at least $1-\exp(-2^{2j+6}/48)$ $(> 1-1/(2^{j+1}+2))$, starting from a configuration with population gap g_j, the population gap is increased to g_{j+1}, before it is reduced to g_{j-1}, within $2^{10}n$ productive reactions. (Assertion (i) follows directly from the fact that the expected number of steps for an unbiased random walk to move distance d from its initial position is d^2, and (ii) follows from Lemma 3).

Thus, (i) the transition between checkpoint states g_0 and g_1 has associated probability at least 3/4, and (ii) for $j \geq 1$, the transition between checkpoint states g_j and g_{j+1} has associated probability at least $1-1/(2^{j+1}+2)$.

Observe that if this random walk on checkpoint states is augmented with transitions from every checkpoint state g_j to state g_0 with some associated probability p_j, then any increase in p_j above 0 (reducing the other transition probabilities from state g_j accordingly) can only decrease the probability of reaching some checkpoint state s from some initial state $g_{\text{init}} < s$. This follows immediately from the fact that any walk from state 0 to state s must pass through all states j, $0 \leq j \leq s$.

Given this, it suffices to analyse an augmented random walk on the checkpoint states in which
(i) the initial state is g_0;
(ii) the transition from checkpoint state g_0 to g_0 has associated probability $p_0 = 1/4$;
(iii) for $j \geq 1$, the transition from checkpoint state g_j to g_0 has associated probability $p_j = 1/(2^{j+1}+2)$; and
(iv) for $j \geq 1$, the transition from checkpoint state g_j to g_{j+1} (resp., g_{j-1}) has associated probability $1-p_j$ (resp., 0).

Note that, with probability exactly $1/2^{t+1}$, $t \geq 1$ transitions of this augmented random walk reach state g_{t-1} before returning to state 0, something we call a *length* t *foray*. Accordingly, with probability $1/2+1/2^{t+1}$, it reaches state g_t after $t \geq 1$ transitions, and so finite-length forays have a total probability of 1/2. This reveals a simple probability-preserving bijection between length-t forays and sequences of unbiased coin flips of the form 1^t0, for $t \geq 1$.

This extends to a natural bijection between a sequence of f finite-length forays with a total of T transitions and a random binary sequence of length $f+T$ that

[2] Here, we tacitly assume that $n = (2k)^2$, for some integer k. Extending the argument to general n, though notationally cumbersome, is straightforward.

(i) starts with a 1 and (ii) contains no two consecutive 0's.

Since a random binary sequence of length $2\gamma \lg n$ fails to contain two consecutive 0's with probability at most $\exp(-\Omega(\gamma \lg n))$, it follows that with probability at most $\exp(-\Omega(\gamma \lg n))$ a sequence of forays with a total of $\gamma \lg n$ transitions fails to reach state $g_{(\lg n)/2-3} = n$.

By construction, our augmented random walk on checkpoint states makes a transition with every $2^{10}n$ productive reactions. Thus, with probability $1 - \exp(-\Omega(\gamma \lg n))$, our tri-molecular CRN must produce a population gap of at least $\sqrt{\gamma \lg n}$ within $2^{10}\gamma n \lg n$ productive reactions. But, as detailed in Lemma 5, since the populations of X and Y are both $\Theta(n)$ throughout this process, these productive reactions occur within $O(\gamma n \lg n)$ interaction events.

Remark. The results of this section are based on the assumption that reactions (1) and (2) of the tri-molecular CRN have rate constants 1. If the rates are reduced but remain equal, the analyses of correctness as well as stage and phase completion within a small number of productive reaction events (i.e., Corollaries 1, 2, and 3) are identical. Analysis of the number of interaction events needed for phase completion changes very slightly to account for the constant factor change in the probability that an interaction results in a productive reaction event, although the statement of Lemma 5 also remains unchanged.

4 Approximate Majority Using Bi-molecular Reactions

Here we provide some details of correctness and efficiency of the bi-molecular CRNs Heavy-B, Double-B and Single-B of Figure 2.

Theorem 2 *For any constant $\gamma \geq 1$, a computation of the Heavy-B, Double-B or Single-B bi-molecular CRNs reaches a consensus of X-majority, with probability $1 - \exp(-\Omega(\gamma \lg n))$, in $O(\gamma n \lg n)$ interaction events, provided (i) the initial molecular count of X and Y together is at least $n/2$, and (ii) the initial count of X exceeds that of Y by at least $\sqrt{\gamma n \lg n}$.*

The bi-molecular CRNs emulate each of the reactions of our tri-molecular system as a sequence of two bi-molecular reactions. While their success in doing so might seem obvious at first, arguing that the introduction of the new composite molecule (denoted as $\mathbb{B}$ in the Heavy-B CRN and as B in the Double-B and Single-B CRNs) preserves both the correctness and the efficiency of the underlying tri-molecular system is not entirely trivial. As evidence of this, note that an emulation that uses composite molecules consisting of two X's or two Y's (instead of $\mathbb{B}$) is conceptually similar to Heavy-B but may never reach a stable consensus configuration.

4.1 Correctness of the Heavy-B emulation

Recall that x (resp., y) denotes the population of molecule X (resp., Y). Let b denote the population of the composite molecule $\mathbb{B}$. It is easy to confirm that $x+y+2b$ is invariant over time and is equal to n, the initial number of molecules. Note that the total number of molecules $m = x + y + b$ satisfies $n/2 \leq m \leq n$ at any point in a computation. Our analysis uses exactly the same three phases (and the same sub-phase stages) that we used in our tri-molecular analysis.

We measure progress throughout in terms of the change in the molecular counts $\hat{x}$, defined as $x + b$, and $\hat{y}$, defined as $y + b$. Note that reaction (0') leaves these counts unchanged and reactions (1') and (2') change $\hat{x}$ and $\hat{y}$ in exactly the same way that the corresponding tri-molecular reactions (1) and (2) change x and y. In each phase, we note that the relative probability of reaction (1') to that of (2'), equals the relative probability of reaction (1) to that of (2) in the tri-molecular CRN. This allows us to conclude that the evolution of $\hat{x}$ and $\hat{y}$ in the bi-molecular system exactly matches that of x and y in the tri-molecular CRN.

4.2 Efficiency of the Heavy-B emulation

To show efficiency, we first observe that if $y \leq y_0$ at the start of a sequence of k productive reaction events, then the number of these reaction events that are of type (1') or (2') is at least $(k-y_0)/3$. Otherwise, at most $2(k-y_0)/3-2$ Y's could be produced by such reactions, and these, together with the at most y_0 Y's available initially, are insufficient to "fuel" the more than $k - (k - y_0)/3$ remaining reactions of type (0'). Therefore, with high probabilities (the probabilities exactly as for the the corresponding statements of the tri-molecular protocol):

- $7n$ productive reaction events are sufficient to complete any stage of phase 1 (since in any stage of phase 1, initially $y \leq n$);
- $16n/2^s$ productive reaction events are sufficient to complete any stage s of phase 2 (since in stage s of phase 2, initially $y \leq n/2^s$); and
- $16\gamma \lg n$ productive reaction events are sufficient to complete phase 3 (since in phase 3, initially $y \leq \gamma \lg n$).

It remains to show that, with high probability, the number of interaction events needed for any given stage or phase to complete is, to within a constant factor, the same as the number of interaction events needed in corresponding phases/stages of the tri-molecular protocol.

To this end we first observe that at any time the propensities of reactions (0') and (2') sum to $(xy + by)/\binom{n}{2}$, which is at least $y_{\min}/(4n)$, where $y_{\min}$ denotes the minimum value of y during a particular phase/stage. It follows immediately that $\Theta(n)$ interactions suffice, with probability $1 - \exp(-\Theta(\gamma \lg n))$, to produce the $\Theta(y_{\min})$ productive reactions needed in any properly-completing stage within phases 1 and 2. Furthermore, since $y_{\min} = 1$ for stage 3, it follows that $\Theta(\gamma n \lg n)$ interactions suffice, with probability $1 - \exp(-\Theta(\gamma \lg n))$, to produce the $\Theta(\gamma \lg n)$ productive reactions needed in phase 3.

4.3 Double-B

The analysis of the Double-B bi-molecular CRN of Figure 2(b) is very similar to that of Heavy-B. Molecular counts in Double-B computations satisfy the invariant $x + y + b = n$. Correctness follows by very similar reasoning. The only slight differences are that (i) we define $\hat{x}$ and $\hat{y}$ to be $x + b/2$ and $y + b/2$, respectively, so that reaction (0') leaves these counts unchanged, and (ii) the counts $\hat{x}$ and $\hat{y}$ increase or decrease by $1/2$ due to reactions (1') and (2') when the counts x and y increase or decrease by one due to reactions (1) and (2) of the tri-molecular CRN. With regard to efficiency, if $y \leq y_0$ at the start of a sequence of k productive reaction events, at least $(k - y_0)/2$ of them will be of type (1') or (2'). The result of these changes means that with high probability, the number of productive reactions that are sufficient for stages 1, 2 and 3 to complete are $9n$, $11n/2^s$, and $11\gamma \lg n$, respectively. The rest of the analysis of Double-B is identical to that of Heavy-B.

4.4 Single-B

We analyse phase 1 of Single-B; changes to the analysis of other phases are similar. The statement of Lemma 3 for the tri-molecular protocol changes very slightly for Single-B; we simply replace x and y by $\hat{x} = x + b/2$ and $\hat{y} = y + b/2$ and the number of productive reactions needed increases from $2n$ to $15n$. To restate: At any point in the computation, if $\hat{x} - \hat{y} = \Delta > 0$, where $\sqrt{\gamma n \lg n}/2 < \Delta < n$, then with probability at least $1 - \exp(-\Theta(\gamma \lg n))$, $x - y$ increases to $\min\{2\Delta, n\}$, without decreasing below $\Delta - \frac{\gamma n \lg n}{2\Delta}$, within $15n$ productive reaction events.

The proof must account for the fact that reaction (0'x) increases $\hat{x}$ by $1/2$ and decreases $\hat{y}$ by $1/2$, while (0'y) does the opposite. As in the proof of Lemma 3, we let $d = \frac{\gamma n \lg n}{8\Delta} \leq \Delta/2$. Among kn reactions of type (0'x) or (0'y), the expected number of type (0'x) is $kn/2$, since both are equally likely. Thus by Lemma 2, the number of type (0'x) is at least $kn/2 - d/4 \geq kn/2 - \Delta/8$ with probability at least

$$1 - \exp(-(k/4)n(\frac{\gamma \lg n}{4k\Delta})^2) \geq 1 - \exp(-\Theta(\gamma \lg n)).$$

Assume in what follows that, among kn reactions of type (0'x) or (0'y), the number of type (0'x) is at least $kn/2 - d/4 \geq (k-1)n/2$ for sufficiently large n. Then there is an excess of at most $d/2$ failures (reactions of type (0'x)) to successes (reactions of type (0'y)) among the reactions of type (0'x) or (0'y). Also, in a sequence of $15n$ productive reaction events, the number of events that are of type (1') or (2') is at least $4n$. Otherwise, at most $4n - 1$ Y's could be produced by such reactions, and these, together with the at most n Y's available initially are insufficient for the $(15n - 4n - n)/2 = 5n$ or more reactions of type (0'x).

Following the same reasoning as in Lemma 3, we can show that the probabiliy of an excess of $d/2$ failures to successes among the at least $4n$ reactions of type (1') or (2') is less than $(1 - \frac{2\Delta}{2n+\Delta})^{d/2}$, which is at most $\exp(-\Theta(\gamma \lg n))$. Since also, among the reactions of type (0'x) or (0'y), there is an excess of at most $d/2$ failures to successes, the overall probability that $\hat{x} - \hat{y}$ becomes less than $\Delta - d$ is $\exp(-\Omega(\gamma \lg n))$.

Assuming that $\hat{x} - \hat{y}$ remains at least $\Delta - d$, and taking into account the excess of at most $d/2 \leq \Delta/4$ failures to successes among the reactions of type (0'x) or (0'y), the probability that $\hat{x} - \hat{y}$ reaches $\min\{2\Delta, n\}$ is at least the probability of an excess of $\Delta + \Delta/4$ successes to failures among the at least $4n$ reactions of type (1') and (2'). This is just the probability that a sequence of $4n$ independent trials with success probability $p > 1/2 + \Delta/(4n)$ results in at least $2n + \Delta/2 + \Delta/8 = 2n + 5\Delta/8$ successes. By Lemma 2, this probability is at most $\exp(-9\Delta^2/(64(2n + \Delta)))$, which is $\exp(-\Omega(\gamma \lg n))$.

5 Uncertainty or Variability in the Rate Constants

There are several reasons to consider the sensitivity of the results of our tri-molecular analysis (in Section 3) to uncertainty or variability in the reaction rate constants. For example, it is known that, when a CRN is "compiled" to a DNA strand displacement system, the DNA strand displacement (DSD) reaction rate constants can

only be made to closely approximate, but not necessarily equal, the desired CRN reaction rates. In fact, Soloveichik et al. [8] provides a DSD construction that can control the reaction rate constants over 6 orders of magnitude that is still considered as a constant error. Furthermore, since the presence of Byzantine molecules (cf. Section 6 can be modelled in terms of their distortion of reaction rates, understanding the tradeoff between relative reaction rates and population gaps allows us to quantify precisely the changes needed to accommodate a fixed population of Byzantine molecules.

Our goal in this section is to determine the size of the gap between X and Y populations that is sufficient to guarantee correct and efficient computations for Approximate Majority with high probability, when there is uncertainty or variability in the relative reaction rates. To this end we suppose that the rate of reaction (2) remains a constant 1 while the rate of reaction (1) is some not-necessarily-constant value $\alpha \leq 1$ (what we call the *relative reaction rate*). For a given α, we revisit our analysis of Section 3, asking at each point what bound on the population gap is sufficient to ensure that the computation will advance quickly, with high probability.

Our results, summarized in Lemmas 6 and 7, capture the robustness of our tri-molecular CRN in the face of uncertain or changing reaction rates. In the next section, this is applied directly to the analysis of our protocol in the presence of Byzantine molecules. Lemma 7 by itself allows us to establish a tight bound on the size of the initial gap required for correct and efficient consensus, with high probability, in the special case where the relative reaction rate is constant, but different from 1.

Remark. Note that, assuming $x > y$, at any point in time the probability p that a productive reaction is reaction (1) is at least $\frac{\alpha\binom{x}{2}y}{\alpha\binom{x}{2}y+x\binom{y}{2}} > \frac{\alpha x}{\alpha x+y} = 1 - \frac{y}{\alpha x+y} = 1/2 + \frac{\alpha x-y}{2(\alpha x+y)}$. If for some β, $\alpha \geq \beta y/x$, then $p \geq 1/2 + \frac{\beta-1}{2(\beta+1)}$ and $\frac{1-p}{p} \leq \frac{1}{\beta}$.

Lemma 6 *At any point in the computation, if $0 < x - y = \Delta < n$, then with probability at least $1 - \exp(-\Theta(\Delta^2/n))$, $x - y$ increases to $\min\{2\Delta, n\}$ within $5n$ productive reaction events, provided the relative reaction rate satisfies $\alpha \geq 1 - \frac{\Delta}{2n}$.*

Proof We argue along the same lines as the proof of Lemma 3. Suppose that $\alpha \geq 1 - \frac{\Delta}{2n} = \beta\frac{n-\Delta/2}{n+\Delta/2}$, where $\beta = 1 + \frac{\Delta}{2n}$. Since $x - y \geq \Delta/2$ up to the point (if ever) when we first have $x - y < \Delta/2$, it follows that $x \geq n/2 + \Delta/4$ and $y \leq n/2 - \Delta/4$, and hence $\alpha \geq \beta y/x$. Thus, by the remark above, the probability p that a productive reaction is reaction (1) is at least $1/2 + \frac{\beta-1}{2(\beta+1)}$ and $\frac{1-p}{p} \leq \frac{1}{\beta}$ up to this point.

As before, we can view the change in $x - y$ resulting from productive reaction events as a random walk, starting at Δ, with success (an increase in $x-y$ by two) probability p. It follows from Lemma 1 that the probability of reaching a configuration where $x - y < \Delta/2$ (which entails an excess of $\Delta/4$ failures to successes) is less than $(\frac{1-p}{p})^{\Delta/4} \leq (\frac{1}{\beta})^{\Delta/4} = (\frac{2n}{2n+\Delta})^{\Delta/4} < (1 - \frac{\Delta}{2n})^{\Delta/4}$, which is at most $\exp(-\Theta(\Delta^2/n))$.

As long as $x - y$ remains at least $\Delta/2$, the probability that it fails to be increased to $\min\{2\Delta, n\}$ within $5n$ productive reactions is just the probability that a sequence of $5n$ independent trials with success probability $p \geq 1/2 + \frac{\beta-1}{2(\beta+1)} = 1/2 + \frac{\Delta}{2(4n+\Delta)} \geq 1/2 + \frac{\Delta}{10n}$ results in fewer than $5n/2 + \Delta/4$ successes, i.e. at least $\Delta/4$ fewer than expected. By Lemma 2, this probability is at most $\exp(-\Theta(\frac{\Delta^2}{n}))$.

Lemma 7 *Let $1 < \beta \leq 2$ be any constant. At any point in the computation, if $y = n/k$, where $1 + \beta \leq k \leq n/(\gamma \lg n)$, then, with probability $1 - \exp(-\Omega((\beta - 1)\gamma \lg n))$, y decreases to 0 within $\Theta(\frac{n}{(\beta-1)k})$ productive reaction events, provided $\alpha \geq \beta\frac{n/k+\gamma \lg n}{n(k-1)/k-\gamma \lg n}$.*

Proof We argue along the same lines as the proof of Lemma 4. Suppose that, while $y \leq n/k+\gamma \lg n$, reaction (1) has associated reaction rate $\alpha \geq \beta\frac{n/k+\gamma \lg n}{n(k-1)/k-\gamma \lg n}$. Since $y \leq n/k + \gamma \lg n$ (and $x \geq n(k-1)/k - \gamma \lg n$) up to the point (if ever) when we first have $y > n/k+\gamma \lg n$, it follows that $\alpha \geq \beta y/x$. By the remark preceding Lemma 6, the probability p that a productive reaction is reaction (1) is at least $1/2 + \frac{\beta-1}{2(\beta+1)}$, and $\frac{1-p}{p} \leq \frac{1}{\beta}$ up to this point.

Let $\lambda = \frac{2(\beta+1)}{\beta-1}$. As before, we can view the change in y resulting from productive reaction events as a random walk, starting at n/k, with success (a decrease in y) probability $p \geq 1/2 + 1/\lambda$. It follows from Lemma 1 that reaching a configuration where $y > n/k + \gamma \lg n$ (which entails an excess of $\gamma \lg n$ failures to successes) is less than $(\frac{1-p}{p})^{\gamma \lg n} \leq (\frac{1}{\beta})^{\gamma \lg n}$. This is $\exp(-\Omega(\gamma(\beta - 1) \lg n))$, since $\lg \beta \geq (\beta - 1)$, for $1 < \beta \leq 2$.

As long as y remains no larger than $n/k + \gamma \lg n$, the probability that it fails to be decreased to 0 within $\lambda n/k = \Theta(n/((\beta-1)k)$ productive reactions is just the probability that a sequence of $\lambda n/k$ independent trials with success probability $p \geq 1/2 + 1/\lambda$ results in fewer than $(\lambda+1)n/(2k)$ successes, i.e., at least $n/(2k)$ fewer than expected. By Lemma 2, this probability is at most $\exp(-\Theta(n/(\lambda k)))$, which is $\exp(-\Omega((\beta-1)\gamma \lg n))$, since $\lambda = \Theta(1/(\beta - 1))$.

Let $\alpha < 1$ be any fixed constant. We refer to the tri-molecular CRN with reaction rate α (respectively, 1) for reaction (1) (respectively, (2)), as α-relaxed tri-molecular CRN. It is interesting to determine the size of the smallest initial gap for which the α-relaxed tri-molecular CRN can reach the correct consensus with high probability. The following theorem confirms our empirical results, detailed in Section 9, that this threshold is arbitrarily close to $\frac{1-\alpha}{1+\alpha}n$.

Theorem 3 *For any constant $\gamma \geq 1$, a computation of the α-relaxed tri-molecular CRN reaches a consensus of X-majority, with probability $1-\exp(-\Omega(\frac{\xi(1-\alpha^2)}{2\alpha-(1-\alpha)\xi}\gamma \lg n))$ in $O(\frac{\alpha(2\alpha-(1-\alpha)\xi)}{\xi(1-\alpha^2)}\gamma n \lg n)$ interaction events, provided the initial molecular count of X exceeds that of Y by at least $(1+\xi)\frac{1-\alpha}{1+\alpha}n$, where $0 < \xi < \alpha$ is any constant.*

Proof Suppose that $0 < \xi < \alpha < 1$. Let $c = (1+\xi)\frac{1-\alpha}{1+\alpha} < \alpha < 1$. If the initial molecular count of X exceeds that of Y by $\Delta \geq cn$ then the initial count of Y is at most n/k, where $k = \frac{2}{1-c}$.

Let $\beta = \alpha\frac{(k-1)n-k\gamma \lg n}{n+k\gamma \lg n}$. Choose $0 < \epsilon \leq \xi(1-\alpha^2)/4 < \frac{\xi(1-\alpha^2)/2}{2\alpha-\xi(1-\alpha)}$. For n sufficiently large, $\alpha(k-1) - \epsilon < \beta < \alpha(k-1)$. But, substituting for k and c, and simplifying, we see that $\alpha(k-1) = 1+\frac{\xi(1-\alpha^2)}{2\alpha-\xi(1-\alpha)}$. Hence $\beta < 2$ (since $\xi < \alpha$) and $\beta - 1 = \Theta(\frac{\xi(1-\alpha^2)}{2\alpha-\xi(1-\alpha)})$.

Hence we can apply Lemma 7 to conclude that rate α suffices to sustain an efficient and correct computation with probability at least $1-\exp(-\Omega((\beta-1)\gamma \lg n))$, within $\Theta(\frac{n}{(\beta-1)k})$ productive reaction events. Since productive reaction events in an α-relaxed tri-molecular CRN, occur at least α times as often as they do in our standard CRN, it follows from Lemma 5, we have that, with high probability, these productive reaction events occur within $\Theta(\frac{\alpha}{(\beta-1)}\gamma n \lg n)$ interaction events.

6 Byzantine Molecules

Here, we study the impact on our tri-molecular protocol if some subset Z of the molecules can exhibit unreliable behaviour. So-called Byzantine molecules, can, at any time, either be neutral or play the role of either X or Y in a reaction event and, in any event they remain Byzantine throughout the computation. Note that, if $z = |Z| \geq 2$, a true consensus, where all the non-Byzantine molecules are either X or Y, cannot be sustained, because even when all the non-Byzantine molecules are X, any interaction between an X molecule and two Byzantine molecules can trigger a productive reaction event which results in consuming one X molecule and producing one Y molecule. We define a *relaxed X-consensus* to be a configuration in which the population of X is $n - 8z$. [3] Even a relaxed consensus is impossible to sustain indefinitely: assuming that Byzantine molecules consistently behave like the minority species Y, with probability 1 any relaxed consensus will eventually be followed by consensus on the minority species. However, we will show:

Theorem 4 *Consider computations of the tri-molecular CRN with initial populations x, y and z of X, Y and Byzantine molecules, respectively. If $\gamma \geq 1$ is any constant then, with probability $1-\exp(-\Omega(\gamma \lg n))$, such a computation reaches a relaxed X-consensus within $O(\gamma n \lg n)$ interaction events, provided $\Delta_0 = 2x - n \geq \sqrt{\gamma n \lg n}$ and $z < \Delta_0/16$. Furthermore, with the same probability, this relaxed X-consensus is preserved over the subsequent n^γ interaction events.*

In our tri-molecular protocol, it straightforward to show that impact is most severe if Byzantine molecules consistently behave like the minority species Y. Specifically, in the worst case we can assume that Byzantine reactions:

$$X+Y+Z \to X+X+Z \text{ and } Y+Z+Z \to X+Z+Z$$

have associated rate constants 0, and the reactions:

$$X+Y+Z \to Y+Y+Z \text{ and } X+Z+Z \to Y+Z+Z$$

have associated rate constants 1.

Lemma 8 *From any configuration (x,y,z) and any $0 < \ell \leq n-x-z$, the probability $\pi^*((x,y,z),(x+\ell,y-\ell,z))$ is minimized when the above assumptions hold.*

Proof It suffices to observe that $\forall \ell > 0, \pi^*((x,y,z),(x+\ell,y-\ell,z)) \leq \pi^*((x+1,y-1,z),(x+\ell,y-\ell,z))$. This follows immediately from the fact that, by the nature of our protocol, the populations of X and Y change incrementally, and hence any computation that takes configuration (x,y,z) to configuration $(x+\ell,y-\ell,z)$, must pass through configuration $(x+1,y-1,z)$.

Proceeding with this worst-case assumption concerning the behaviour of Byzantine molecules, and preserving the invariant that the number of Byzantine molecules remains unchanged, we can combine the reactions involving one X and two elements of $\hat{Y} = Y \cup Z$ into $X+\hat{Y}+\hat{Y} \to \hat{Y}+\hat{Y}+\hat{Y}$. Similarly, we can combine the reactions involving two X's and one element of $\hat{Y}$ into $X+X+\hat{Y} \xrightarrow{r} X+X+X$, where the reaction rate $r = y/(y+z)$ is just the probability that a randomly chosen element of $\hat{Y}$ is a Y molecule.

[3] A similar, though more involved argument, can be used to show the same result with relaxed X-consensus defined as an X-population of size $n-(1+\epsilon)z$, for any $\epsilon > 0$.

The resulting system:

$$X+X+\hat{Y} \xrightarrow{r} X+X+X \text{ and } X+\hat{Y}+\hat{Y} \xrightarrow{1} \hat{Y}+\hat{Y}+\hat{Y}$$

is just our original tri-molecular system with Y replaced by $\hat{Y}$, and the rate of the first reaction reduced to r. We refer to $y/(y+z)$ as the *effective rate* of the first reaction and, taking advantage of the results of the previous section, it suffices to show that, in any configuration, this effective reaction rate is sufficiently large to sustain the Byzantine tri-molecular system to a relaxed consensus on the majority species.

Lemma 9 *Suppose that $z \leq \Delta_0/16$. If $\Delta = 2x - n$ satisfies $\Delta_0/2 \leq \Delta \leq n/2$, then the effective reaction rate is at least $1 - \frac{\Delta}{2n}$.*

Proof Since $z \leq \Delta_0/16 \leq \Delta/8$ and $\Delta \leq n/2$ the effective rate satisfies $\frac{y}{y+z} = 1 - \frac{2z}{n-\Delta} \geq 1 - \frac{\Delta/4}{n/2} = 1 - \frac{\Delta}{2n}$.

Lemma 10 *Suppose that $x < n - 2z$. If $y + z = n/k$, where $k \geq 4$, then the effective reaction rate is greater than $\frac{6}{5}\frac{n/k+\gamma \lg n}{n(k-1)/k-\gamma \lg n}$.*

Proof Since $y = n - (x + z) > z$, $y/(y + z) > 1/2 > \frac{6}{5}\frac{n/k+\gamma \lg n}{n(k-1)/k-\gamma \lg n}$, for n sufficiently large.

It follows from Lemmas 6 and 9 that the tri-molecular protocol with Byzantine molecules, proceeds until $\Delta > n/2$, or equivalently until $y + z < n/4$, and it follows from Lemmas 7 and 10 that the protocol proceeds from this point to a configuration with $x \geq n - 2z$, or equivalently until $y < z$, all with probability $1 - \exp(-\Omega(\gamma \lg n))$. To complete the proof of Theorem, we simply observe that once such a configuration has been reached, a subsequent doubling of $y + z$ happens with probability at most $\exp(-\Theta(\gamma \lg n))$. Thus, two consecutive doublings of $y + z$ (necessary to escape a relaxed consensus) happen within n^γ interactions, with probability at most $\exp(-\Theta(\gamma \lg n))$.

7 Multi-species Consensus

Multi-species consensus involves a total of n molecules drawn from a set $\{Z_0, Z_1, \ldots, Z_{m-1}\}$ of m distinct species, where $3 \leq m < n$. The goal is to reach consensus on the plurality (most populous) species. Here we analyse a CRN, which we call m-species-Tri, for m-species consensus. The CRN has $m(m-1)$ reactions, two for each distinct pair (Z_i, Z_j) of species, referred to as (Z_iZ_j) reactions:

$Z_i+Z_i+Z_j \to Z_i+Z_i+Z_i$
$Z_i+Z_j+Z_j \to Z_j+Z_j+Z_j$

and is the natural generalization of Tri, our tri-molecular CRN for approximate majority. We will show the following:

Theorem 5 *For any constant $\gamma \geq 1$, if $m \leq n/(\gamma \lg n)$, a computation of m-species-Tri reaches consensus on the plurality species with probability $1-\exp(-\Omega(\gamma \lg n))$, in $O(mn \lg n)$ interaction events, provided the initial count of the plurality species exceeds that of any other species by at least $\sqrt{\gamma n \lg n}$.*

7.1 Analysis Overview

We denote the initial plurality species by X and the remaining species by $\mathcal{Z} = \{Z_1, Z_2, \ldots, Z_{m-1}\}$. At any time the count of X (resp. Z_i) is denoted by x (resp., z_i). We abbreviate $\max_{Z\in\mathcal{Z}}\{z\}$ as z_{max} and $\sum_{Z\in\mathcal{Z}}\{z\}$ as z_{sum}.

Much of our analysis leverages our earlier two-species analysis. We will show that, if initially $x - z_{max} \geq \sqrt{\gamma n \lg n}$, then with high probability, in what we will refer to as *phase 0*, the computation will reach a point where $x - z_{sum} \geq \sqrt{\gamma n \lg n}$. Thereafter the analysis of correctness and efficiency follows directly from the analysis of the two-species computation, with z_{sum} substituted for y. This is because (i) all reactions that change z_{sum} involve X, and (ii) the probability of failure (an increase in z_{sum}) is maximized when the entire population z_{sum} is concentrated in one species, which is exactly the two-species case.

Our phase 0 analysis parallels that of phase 1 of the two-species analysis, using a similar composition into short stages, with $x - z_{max}$ doubling in each stage. With high probability, these stages occur in sequence, i.e., there is no "backsliding". (No backsliding also ensures that X remains the plurality species.) We reason about changes in $x - z_{max}$ by considering an arbitrary species $Y \in \mathcal{Z}$, with count y, and analysing changes in $x-y$. Compared with the two-species analysis, the only additional complication that arises in analysing $x - y$ is the fact that the populations, both absolute and relative, of X and Y are impacted by reactions involving other minority species (what we refer to as *third-party* reactions).

7.2 No backsliding

We consider two types of "backsliding" events that could arise in the computation, starting at a point where $x - z_{max} \geq \Delta$.

Two-party backsliding: For some $Y \neq X$, (XY) reactions result in a cumulative decrease of at least $\Delta/4$

to $x-y$ at some subsequent point in the computation.

Third-party backsliding: Third-party reaction events, i.e., of types (XZ) or (YZ) for some $Z \notin \{X, Y\}$, result in a cumulative decrease of at least $\Delta/4$ to $x-y$ at some subsequent point in the computation.

Lemma 11 (No Backsliding) *Suppose that $x - z_{max} \geq \Delta$ at some point in the computation. Then the probability of two-party backsliding is $m\exp(-\Omega(\Delta^2/n))$, and so also is the probability of third-party backsliding. Consequently, $x - z_{max} \geq \Delta/2$ for the remainder of the computation with probability $1 - m\exp(-\Omega(\Delta^2/n))$.*

Proof Consider the computation from the given point up to a point (if ever) when we first have either two-party or third-party backsliding. During this period, $x - z_{max} \geq \Delta/2$. We show that for any given species Y, the probabilities of two-party and third-party backsliding are $\exp(-\Omega(\Delta^2/n))$. Then since Y is arbitrary among the $m-1$ species other than X, the overall probability of backsliding is $m\exp(-\Omega(\Delta^2/n))$.

Two-party backsliding: Upon reaction events of type (XY), $x-y$ either increases by two (success) or decreases by two (failure). The probability of success is at least

$$\begin{aligned} x/(x+y) &= 1/2 + (x-y)/(2(x+y)) \\ &> 1/2 + \Delta/(4x) \\ &> 1/2 + \Delta/(4n). \end{aligned}$$

From Lemma 1, with $p = 1/2 + \Delta/(4n)$, the probability that such reactions ever result in a cumulative decrease of $\Delta/4$, which requires an excess of $\Delta/8$ failures to successes, is at most

$$\left(\frac{1-p}{p}\right)^{\Delta/8} < (1-\Delta/n)^{\Delta/8} = \exp(-\Omega(\Delta^2/n)).$$

Third-party backsliding: Upon reaction events (XZ) or (YZ), success and failure increase and decrease $x-y$ by 1, respectively.

- If $x > z > y$ then both $x/(x+z)$ and $z/(z+y)$ are greater than $1/2$, ensuring that success is more likely than failure. Moreover, since $x - z > \Delta/2$, the probability of success when z interacts with x is at least $1/2 + \Delta/(4n)$. Since z is more likely to interact with x than y, the overall probability of success in this case is at least $1/2 + \Delta/(8n)$.
- If $x > y > z$, then the probability of success is at least

$$\begin{aligned} &(xx+yz)/(xx+yz+xz+yy) \\ &\geq 1/2 + (x(x-z) - y(y-z))/(2(xx+yz)) \\ &\geq 1/2 + \Delta/(4x) \geq 1/2 + \Delta/(8n). \end{aligned}$$

Applying Lemma 1 again, with $p = 1/2 + \Delta/(8n)$, the probability that the number of successes ever exceeds the number of failures by at least $\Delta/4$ is at most

$$\left(\frac{1-p}{p}\right)^{\Delta/4} < (1-\Delta/(2n))^{\Delta/4} = \exp(-\Omega(\Delta^2/n)).$$

7.3 Multi-species Phase 0 Analysis

Within phase 0, we track the evolution of $x - z_{max}$ in stages. A typical stage, with integer index t, starts with $x - z_{max} \geq 2^t\sqrt{\gamma n \lg n}$ and continues while $2^{t-1}\sqrt{\gamma n \lg n} \leq x - z_{max} < \min\{2^{t+1}\sqrt{\gamma n \lg n}, n\}$. A stage completes *properly* when $x - z_{max} \geq \min\{2^{t+1}\sqrt{\gamma n \lg n}, n\}$. Initially $x - z_{max} \geq \sqrt{\gamma n \lg n}$, and phase 0 ends if either some stage fails to complete properly or if all stages complete properly and $z_{sum} \leq x - \sqrt{\gamma n \lg n}$. Note that since $x - z_{max}$ doubles in each stage, phase 0 has $O(\lg n)$ stages.

We next show that proper stage completion is likely and happens efficiently, under suitable conditions on m and the initial gap $x - z_{max}$.

Lemma 12 (Proper stage completion) *Consider a point in the computation where $x - z_{max} = \Delta \geq \sqrt{\gamma n \lg n}$. Let $m \leq n/(\gamma \lg n)$. Then $x - z_{max} \geq \min\{2\Delta, n\}$ within $\Theta(n)$ productive reaction events, with probability $1 - \exp(-\Omega(\gamma \lg n))$.*

Proof Consider a computation that starts with $x = x_0$ and $x - z_{max} = \Delta_0 \geq \sqrt{\gamma n \lg n}$, and ends with either (a) $x - z_{max} < \Delta_0/2$ (i.e., backsliding occurs), (b) the completion of $2n$ productive reaction events, (c) the completion of $x_0/2$ productive reaction events involving X, or (d) $x - z_{max} \geq \min\{\frac{48+1}{48}\Delta_0, n\}$, whichever occurs first. We will argue that completions (a), (b) and (c) are unlikely, and use the fact that if none of (a), (b) or (c) happen then (d) happens with high probability.

By Lemma 11, completion of type (a) occurs with probability $m\exp(-\Omega(\gamma \lg n)) = \exp(-\Omega(\gamma \lg n))$ since $m < n$.

Suppose that we have completion of type (b). Then, since (b) precedes (c), $x \in [x_0/2, 3x_0/2]$ throughout the computation, and so the probability that a given productive reaction involves X is at least

$$\frac{\binom{x}{2}(n-x) + \sum_{Z\in\mathcal{Z}}\binom{z}{2}x}{\binom{x}{2}(n-x) + \sum_{z\in\mathcal{Z}}\binom{z}{2}(n-z)} \geq \frac{\sum_{Z\in\mathcal{Z}}\binom{z}{2}x}{\sum_{Z\in\mathcal{Z}}\binom{z}{2}n} \geq \frac{x_0/2}{n}.$$

Thus we expect the $2n$ productive reactions to include at least x_0 productive reactions involving X. Since fewer than $x_0/2$ actually occur, this completion occurs with probability at most $\exp(-x_0/8) = \exp(-\Omega(\gamma \lg n))$ since $x_0 \geq n/m$ and $m \leq n/(\gamma \lg n)$.

Next, consider completions of type (c). Let Y, with count y, be an arbitrary species in the set $\mathcal{Z}$ of minority species such that, at the time of completion, $x-y < \frac{48+1}{48}\Delta_0$. Since (c) precedes (d), there is at least one such species Y. Also, since (c) precedes (a), there is no backsliding throughout the computation that we are considering. Recall from the proof of Lemma 11 that since no backsliding occurs, the probability of success (increase of $x-y$) among two-party productive reactions is at least $1/2 + \Delta_0/(4x)$ and the probability of success among third-party productive reactions is at least $1/2 + \Delta_0/(8x)$.

Thus, since $x \leq 3x_0/2$ throughout the computation that we are considering, the probability of success among productive reactions involving X or Y is at least $1/2 + \Delta_0/(24x_0)$, and hence we expect the at least $x_0/2$ productive reaction events involving X or Y to produce at least $x_0/4 + \Delta/48$ successes. Since $x - y \leq \frac{48+1}{48}\Delta_0$, it must be that fewer than $x_0/4 + \Delta_0/96$ actually occur, which is an event that has probability $\exp(-\Theta(\Delta_0^2/x_0))$, which is $\exp(-\Omega(\gamma \lg n))$ since $\Delta_0 \geq \sqrt{\gamma n \lg n}$ and $x_0 \leq n$.

Repeating the argument above at most $48 \times 3 = 144$ times with $\Delta_0 = \Delta$ the first time and with $\frac{48+i}{48}\Delta \leq \Delta_0 < n$ on the ith repetition, we conclude that within at most $288n$ productive reaction events we have $x - y \geq \min\{4\Delta, n\}$, with probability $1 - \exp(-\Omega(\gamma \lg n))$. Therefore, by Lemma 11, with the same high probability, $x - y \geq \min\{2\Delta, n\}$ from that point onward. Summing these probabilities over all Y, we have that $x - z_{max} \geq 2\Delta$ with probability $1 - \exp(-\Omega(\gamma \lg n))$.

Finally, we analyse the efficiency of phase 0 in terms of interaction events:

Lemma 13 *Let $m \leq n/(\gamma \lg n)$. For each of the $O(\lg n)$ stages of phase 0, the at most $288n$ productive reaction events that suffice for high-probability successful completion occur within $\Theta(mn)$ interaction events, with probability $1 - \exp(-\Omega(\gamma \lg n))$.*

Proof The probability that an interaction is a productive reaction event is greater than $\frac{\sum_{Z \in \mathcal{Z}} (n-z)\binom{z}{2}}{\binom{n}{3}}$. Recall that during phase 0, we have $z_{sum} \geq x - \sqrt{\gamma n \lg n}$. Since $n - z > n - x = z_{sum} \geq (n - \sqrt{\gamma n \lg n})/2$, it follows that this probability is at least $\frac{n-\sqrt{\gamma n \lg n}}{2\binom{n}{3}} \sum_{Z \in \mathcal{Z}} \binom{z}{2}$. The latter is minimized when $z = z_{sum}/(m-1)$ for all $z \in \mathcal{Z}$, in which case, its value is at least

$$\frac{n - \sqrt{\gamma n \lg n}}{2\binom{n}{3}} (m-1) \binom{\frac{n-\sqrt{\gamma n \lg n}}{2(m-1)}}{2}$$

which is $\Theta(1/m)$.

Thus for sufficiently large λ, the probability that a sequence of λmn interactions gives rise to fewer than $288n$ productive reaction events is $\exp(-\Theta(n/m)) = \exp(-\Omega(\gamma \lg n))$.

We can now prove our main result, Theorem 5, on multi-species consensus.

Proof Lemmas 12 and 13 show that, with probability $1 - \exp(-\Omega(\lambda \lg n))$, $x - z_{max}$ increases in phase 0 of the computation, within $\Theta(mn \lg n)$ interactions, until the total population z_{sum} of the non-plurality species is at most $x - \sqrt{\gamma n \lg n}$. The correctness and efficiency of the remainder of the computation is immediate from Theorem 1.

8 Initiation by Infection

8.1 Overview

Here we prove that our tri-molecular Approximate Majority results still hold if we assume that the protocol is initiated by an epidemic that is triggered by a single distinguished molecule. This provides a counterpart to the results of Section 6 in Angluin et al. [10], where it is shown that, under the same initiation assumptions, with high probability the single-B protocol reaches consensus on the initial majority value in $\Theta(n \lg n)$ interactions, provided the initial gap is $\Omega(n^{3/4+\epsilon})$. Our somewhat more involved analysis shows that a gap of $\Theta(\sqrt{\gamma n \lg n})$ suffices:

Theorem 6 *For any constant $\gamma \geq 1$, a computation of the tri-molecular CRN that is initiated by infection reaches a consensus of X-majority, with probability $1 - \exp(-\Omega(\gamma \lg n))$, in $O(\gamma n \lg n)$ interaction events, provided the initial count of X exceeds that of Y by $\Omega(\sqrt{\gamma n \lg n})$.*

It seems straightforward to extend the result, with the same gap bound, to the bi-molecular protocols.

To be precise, we first recall from earlier work on epidemics in essentially the same model (cf. [6], Lemma 2) that the number of interactions that suffice to infect all n molecules, with probability $1 - \exp(-\Omega(\gamma \lg n))$, starting from a single infected molecule, is at most $c^* \gamma n \lg n$, for some suitably large constant c^*, which we will assume is at least 6. Similarly, while the number of infected molecules is $n^{\Theta(1)}$ but not yet $n/2$, the number of interactions necessary and sufficient to double the number of infected molecules, with high probability, is at most $c^* n$.

With this in mind (and following Angluin et al.), we assume that in the infection model there is initially

just one active molecule (either X or Y) and among the remaining $n-1$ inactive molecules, the population of X exceeds that of Y by at least $120d^*\sqrt{\gamma n \lg n}$, where $d^* = (c^*)^2$. Interactions involving only inactive molecules result in no change; interactions involving only active molecules proceed as in the standard trimolecular protocol, and mixed interactions result in activation of previously inactive molecules, but otherwise no change.

By these assumptions, the initially inactive molecules are activated in a random sequence. We will first show (Lemma 15) that the set of molecules that are activated by contiguous subsequences within the first quarter of this random activation sequence have bounded X-*excess*. That is, the difference between the population of X and the population of Y that has been activated is bounded (in terms of the length of the subsequence and the initial gap), with high probability.

Building on our bound on the X-excess among newly activated molecules, we analyse the X-excess among active molecules. It is certainly possible that the X-excess among active molecules is negative for some period of time. Thus, we start by bounding the size and growth rate of the Y-excess (the negative X-excess) among active molecules in stages that lead to the point where n/d^* molecules have been activated. We show (Lemma 16) that at the end of the first stage, involving $a = n/(\gamma n \lg n)^{1/5}$ activations, the Y-excess among active molecules is $O(c^*\sqrt{\gamma a \lg n})$ with high probability. After each subsequent doubling of a, the number of active molecules, which we refer to as a *(doubling) stage*, the Y-excess among active molecules increases geometrically, and is thus $O(c^*\sqrt{\gamma a \lg n})$ after a total of $a \leq n/d^*$ activations (Lemma 17).

We then show (Lemma 18) that, with high probability, one more doubling stage brings an X-excess of newly activated molecules sufficient to overwhelm this tentative Y-excess among active molecules. Thus when $2n/d^*$ of the molecules have been activated, the X-excess among active molecules is at least $2\sqrt{\gamma n \lg n}$.

Finally, we conclude (Lemma 19) from the results of Section 3 that by the time all molecules have been activated the X-excess among active molecules is at least $\sqrt{\gamma n \lg n}$. Consensus on the initial majority follows after $O(n \lg n)$ further interactions, with high probability and Theorem 6 follows.

8.2 The details

The main challenge in bounding the Y-excess among active molecules is to show that non-uniformity in the infection process cannot somehow conspire to undo the population advantage of the unactivated majority species. We use the following lemma to establish this; the proof follows directly from a result of Hoeffding [23] on sampling without replacement . (The reader may find it helpful to picture red balls as X-molecules and blue balls as Y-molecules.)

Lemma 14 *Let $\mathcal{C}$ be a collection of ν red and blue balls of which at least $\nu/2 + \sqrt{\gamma\nu \lg \nu}$ are red, and consider a sequence of s random selections from $\mathcal{C}$, without replacement. Then the number of red balls selected, S_R, satisfies $|S_R - E[S_R]| \leq \sqrt{\gamma s \lg \nu}$, with probability $1 - \exp(-\Theta(\gamma \lg \nu))$.*

We use Lemma 14 to bound the X-excess of activation subsequences.

Lemma 15 *Let S be a contiguous activation subsequence of length s that is contained within the first $n/4$ activations of the computation. Then with probability $1 - \exp(-\Omega(\gamma \lg n))$, the Y-excess within any prefix of S never exceeds $\sqrt{\gamma s \lg n}$, in which case the X-excess is never lower than $-\sqrt{\gamma s \lg n}$.*

Proof Let S' be the activation subsequence from the start of the computation up to but not including the start of S. Note that S' has length at most $n/4$. It follows directly from Lemma 14 that, after the sequence of activations in S', the X-excess among inactive molecules is at least $60d^*\sqrt{\gamma n \lg n}$ (half of our lower bound on the initial X-excess among inactive molecules), with probability $1 - \exp(-\Omega(\gamma \lg n))$. Assuming that the X-excess among inactive molecules remaining at the start of S is at least $60d^*\sqrt{\gamma n \lg n}$, we can apply Lemma 14 again to this set of inactive molecules to conclude that the lemma is true.

We now analyse the composition of the set of active molecules after the first $n/(\gamma n \lg n)^{1/5}$ activations.

Lemma 16 *After the first $a = n/(\gamma n \lg n)^{1/5}$ activations the Y-excess among active molecules is at most $3c^*\sqrt{\gamma a \lg n}$, with probability $1 - \exp(-\Omega(\gamma \lg n))$.*

Proof By Lemma 15, with probability $1-\exp(-\Omega(\gamma \lg n))$, the Y-excess among activated molecules remains less than $\sqrt{\gamma a \lg n}$ during the first $a = n/(\gamma n \lg n)^{1/5}$ activations. Since, with high probability, this initial batch of activations takes place within $c^*\gamma n \lg n$ interactions, and the probability that an interaction gives rise to an active reaction is at most $(a/n)^3 = (\gamma n \lg n)^{-3/5}$, it follows from Lemma 2 that the a active molecules participate in at most $4/3(a/n)^3 c^* \gamma n \lg n = 4/3c^*\sqrt{\gamma a \lg n}$ active reactions. So with high probability, even if all of these reactions are of type (2) (converting one X-molecule to a Y-molecule and thus increasing the Y-excess among active molecules by 2), the total actual

Y-excess is less than $(8/3c^* + 1)\sqrt{\gamma a \lg n}$ after the first $a = n/(\gamma n \lg n)^{1/5}$ activations. This quantity is at most $3c^*\sqrt{\gamma a \lg n}$, since we assume that c^* is at least 6.

We now consider how the Y-excess among active molecules changes with each subsequent activation doubling stage, up to the point where n/d^* molecules have been activated.

Lemma 17 *Suppose that a given doubling stage starts with $n/(\gamma n \lg n)^{1/5} \le a < n/d^*$ active molecules and a Y-excess among them of at most $14c^*\sqrt{\gamma a \lg n}$. Then the Y-excess among active molecules at the end of the stage is at most $14c^*\sqrt{\gamma(2a)\lg n}$.*

Proof There are two sources that could contribute to an increase in the Y-excess among active molecules during the stage:
(i) a (positive) Y-excess among the molecules activated within the stage; and
(ii) a (positive) excess of reactions of type (2) (converting an X-molecule to a Y molecule) over reactions of type (1) (converting a Y-molecule to an X molecule), among active (including newly activated) molecules within the stage.

From Lemma 15, the contribution from the first of these sources is bounded by $\sqrt{\gamma a \lg n}$. Thus, to establish the desired result it suffices to show that, with high probability, the excess of active reactions of type (2), over the stage, is at most $2c^*\sqrt{\gamma a \lg n}$. (Sufficiency follows since such a type (2) reaction excess contributes at most $4c^*\sqrt{\gamma a \lg n}$ to the Y-excess among active molecules, which combined with the less than $c^*\sqrt{\gamma a \lg n}$ from the first source increases the initial Y-excess of at most $14c^*\sqrt{\gamma a \lg n}$ to at most $19c^*\sqrt{\gamma a \lg n} < 14c^*\sqrt{\gamma(2a)\lg n}$.)

But we know that the stage consists of at most c^*n interactions with probability $1 - \exp(-\Omega(\gamma \lg n))$, and hence (by Lemma 2) no more than $2c^*(2a)^3/n^2 = 16c^*a^3/n^2$ active reactions. Furthermore, up to the point (if ever) that the Y-excess among active molecules exceeds $19c^*\sqrt{\gamma a \lg n}$, the probability that an active reaction is of type (2) is at most $1/2 + 19/2c^*\sqrt{\gamma a \lg n}/a$. So, a type (2) reaction excess of $2c^*\sqrt{\gamma a \lg n}$ is realized before the end of the stage only if $16c^*a^3/n^2$ trials (active reactions) with success (type (2)) probability at most $1/2 + 19/2c^*\sqrt{\gamma a \lg n}/a$, give rise to at least $8c^*a^3/n^2 + c^*\sqrt{\gamma a \lg n}$ successes, which is at least $c^*\sqrt{\gamma a \lg n} - (19/2c^*\sqrt{\gamma a \lg n}/a) \cdot (16c^*a^3/n^2)$
$\ge c^*\sqrt{\gamma a \lg n}(1 - 152c^*a^2/n^2) \ge c^*\sqrt{\gamma a \lg n}(1 - 152/(c^*)^3)$ more than expected. By Lemma 2, if $c^* \ge 6$, this has probability at most $\exp(-\Theta(\gamma \lg n))$.

It follows from Lemma 17 that when n/d^* molecules have been activated, the actual Y-excess is at most $14c^*\sqrt{\gamma(n/d^*)\lg n} = 14\sqrt{\gamma n \lg n}$ with high probability.

To continue with our overall proof, we now take advantage of the fact that the X-excess, among molecules activated in the activation doubling stage starting with n/d^* active molecules, is more than enough to overwhelm any previously accumulated actual Y excess.

Lemma 18 *After the doubling stage that starts with $a = n/d^*$ active molecules, the X-excess among active molecules is at least $2\sqrt{\gamma n \lg n}$.*

Proof From Lemma 15, the Y-excess among activated molecules prior to the doubling stage with a active molecules never exceeds $\sqrt{\gamma a \lg n} < \sqrt{\gamma n \lg n}$ with high probability. Moreover, the X-excess among the n/d^* molecules activated in this stage is at least $48\sqrt{\gamma n \lg n}$. (This follows since the expected number of X-molecules activated is at least $n/(2d^*) + 30\sqrt{\gamma n \lg n}$, and so with high probability the actual number is at least $n/(2d^*) + 24\sqrt{\gamma n \lg n}$.)

Thus the only way that the X-excess among active molecules at the end of the stage can be less than $2\sqrt{\gamma n \lg n}$ is if there is an excess of more than $16\sqrt{\gamma n \lg n}$ in the number of active reactions of type (2) within the stage. Such an excess would increase the number of active Y-molecules by more than $16\sqrt{\gamma n \lg n}$ and decrease the number of active X-molecules by the same amount, causing the net X-excess among active molecules to be at most $48\sqrt{\gamma n \lg n} - 14\sqrt{\gamma n \lg n} - 32\sqrt{\gamma n \lg n}$). Note that without such an excess in the number of active reactions of type (2) within the stage, the stage would complete without the Y-excess among active molecules ever exceeding $30\sqrt{\gamma n \lg n}$.

But such a type (2) reaction excess has low probability since (i) the expected number of active reactions in the stage is at most $8c^*a^3/n^2$, and hence, with high probability, the actual number A is at most $10c^*a^3/n^2$, (ii) the probability, up to the point (if ever) that the Y-excess among active molecules exceeds $30\sqrt{\gamma n \lg n}$, that an active reaction has type (2), is at most $1/2 + 15\sqrt{\gamma n \lg n}/a$, and so (iii) the expected number of active reactions of type (2) is at most $A/2 + (A/a)15\sqrt{\gamma n \lg n}$ $\le A/2 + (150/(c^*)^3)\sqrt{\gamma n \lg n}$. Hence, with high probability, the excess of active reactions of type (2) is less than $16\sqrt{\gamma n \lg n}$.

Finally, we show that, with high probability, the X-excess among active molecules remains $\Omega(\sqrt{\gamma n \lg n})$ when all molecules have been activated.

Lemma 19 *If the X-excess among active molecules is at least $2\sqrt{\gamma n \lg n}$ when $2n/d^*$ molecules have been activated then it remains at least $\sqrt{\gamma n \lg n}$ when all molecules have been activated.*

Proof It suffices to observe that, with high probability, (i) by Lemma 14, the Y-excess in the last $n - 2n/d^*$

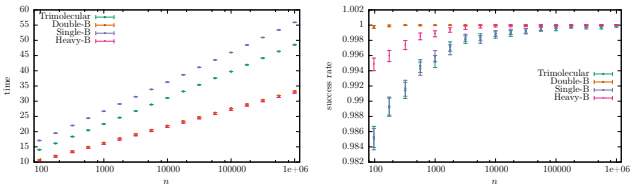

Fig. 5 Comparison of the time (left) and success rate, i.e., probability of correctness (right) of tri-molecular, Double-B, Single-B and Heavy-B CRNs for Approximate Majority. Each point in the plot is an average over 50,000 trials. The initial configuration has no B's and the imbalance between X's and Y's is $\sqrt{n \ln n}$. Plots show confidence intervals at 99% confidence level.

molecules to be activated is never more than $\sqrt{\gamma n \lg n}$, and hence is never enough to offset the initial X-excess among active molecules by more than a constant factor, and (ii) following the analysis of Lemma 3, the X-excess continues to increase by a constant factor with every $\Theta(n)$ successive batch of active reactions.

9 Empirical Results

Figure 5 compares time (efficiency) and success rates (probability of correctness) of the tri-molecular and the three bimolecular CRNs to reach consensus, as a function of the log of the initial count n of molecules.The plots show that time grows linearly with the log of the molecular count, and the success rate is close to 1 for large n. A fit to the data of that figure shows that the expected times of the tri-molecular, Double-B, Single-B and Heavy-B CRNs grow as $3.7 \ln n$, $2.4 \ln n$, $4.2 \ln n$, and $2.4 \ln n$ respectively. For $n \geq 100$, the tri-molecular CRN has at least 99% probability of correctness and the bi-molecular CRNs have at least 97% percent probability of correctness. These probabilities all tend to 1 as n gets larger. On the other hand, if the initial imbalance is reduced to $\sqrt{n}$, the success rate appears to be constant as n grows, cf. Figure 6.

Figure 7 looks at convergence of success rate of the tri-molecular CRN with initial imbalance between X and Y populations being $\Delta_0 = \sqrt{n \ln n}$. In simulations, Byzantine molecules consistently behave like the minority species Y. It shows clear convergence when the count of Byzantine molecules is 1/8 and 1/4 of the initial imbalance. While the convergence is less clear if the count is only 1/2 of the initial imbalance, it appears that the success rate converges to 1/2 as n grows. [This could be explained by looking at the initial success rate, $x/(x + \hat{y})$, where $\hat{y} = (y + z)^2/y$. This rate equals to $(2n + \Delta)/(4n + 2\Delta + \Delta^2/y)$ which tends to 1/2 as $n \to \infty$.] If the number of Byzantine molecules is 3/4 of the initial imbalance between x and y, the success rate stays close to 0.

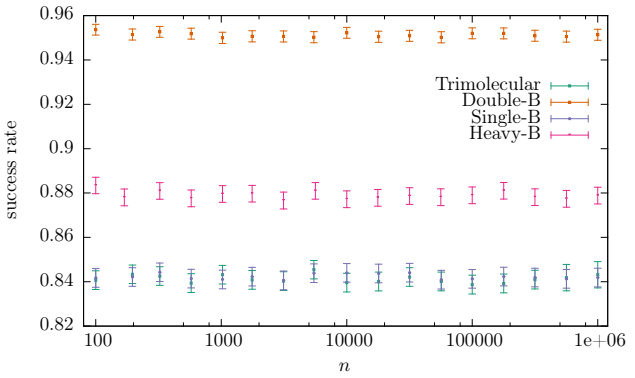

Fig. 6 Success rate of tri-molecular, Double-B, Single-B and Heavy-B CRN for Approximate Majority with initial imbalance $\sqrt{n}$. Each point in the plot is an average over 50,000 trials. Plots show confidence intervals at 99% confidence level.

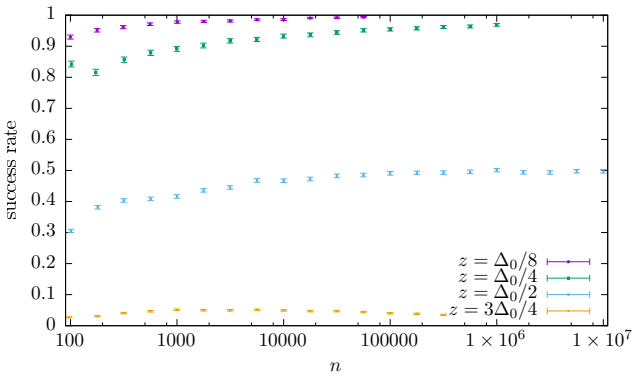

Fig. 7 Success rate of the tri-molecular CRN for Approximate Majority with initial imbalance between X and Y molecules $\Delta_0 = \sqrt{n \ln n}$ in a presence of Byzantine molecules. The number of Byzantine molecules is set to 1/8, 1/4, 1/2, and 3/4 of this initial imbalance, respectively. Each point in the plot is an average over 50,000 trials. Plots show confidence intervals at 99% confidence level.

Figure 8 provides some evidence that the expected execution time of m-species-Tri, the multivalued tri-molecular CRN, with initial gap $\sqrt{n \log n}$, is $\Theta(mn \lg n)$, which matches the upper bound proved in Theorem 5. The top plot shows the progression of the count of the majority species (X) until it reaches a complete majority, for CRNs with $m = 3, \ldots, 32$ species. The time does seem to increase linearly in the number of species m. This is more clear from the middle plot, which is based on 50,000 experiments. It shows that the length of phase 1 is increasing linearly in m, while the durations of phases 2 and 3 remain constant (here, constants d_γ and e_γ used to define when the phases end are set to 10 and 1). The graph also shows that productive reactions involving non-majority species neither slow down nor speed up the other phases. This is not surprising: during interactions between non-majority species, in the first case, these species keep randomly changing identi-

ties, while in the second case nothing happens. Since the counts of non-majority species are more or less equal, the productive reactions between non-majority species have little effect on overall computation. The bottom figure shows counts of all species in one experiment. Non-majority species are packed together. The green dots show the average number of interaction events until a productive reaction event occurs and purple dots the average number of productive reactions per x_1-productive reaction. (Each dot corresponds to 1000 productive reactions). For a big portion of simulation this number fluctuates around $m^2/3(m-1)$ — this is the expected rate at the start of simulation when all species counts are around n/m, but then it drops suddenly before it shoots up in the last phase of the computation, as expected.

Figure 9 compares average time needed to achieve X-majority for CRNs with and without initiation by infection. A fit to the data of that figure suggests that the expected times of the Tri, Double-B, Single-B and Heavy-B CRNs with initiation by infection grow as $4.7\ln n$, $3.4\ln n$, $5.2\ln n$, and $3.2\ln n$ respectively. This is an increase of approximately $\ln n$ in each case compared with the time when all molecules are initially active. The $\ln n$ additional time probably accounts for the time needed to activate each molecule.

Figure 10 provides experimental confirmation that the lower bound $\frac{1-\alpha}{1+\alpha}$ on the initial imbalance for trimolecular CRN with unequal reaction rates proved in Theorem 3 is tight. This figure suggests the following generalization of Theorem 3 for values of $\alpha > 1$. Assume that (unlike Approximate Majority), we are only interested in species X achieving the majority. In this case, if the initial imbalance $\Delta_0 = x_0 - y_0$ is negative, consensus of X can still happen with high probability. In this case, Y is initially the majority species, and we want to determine the relative reaction rate that will ensure that Y fails with high probability. If we slow down both reactions by α, it will not affect the success/failure rates of the system (although the expected overall time until consensus is reached will be multiplied by α). Then the rate of the reaction that is producing "majority" Y is $1/\alpha$. The experiments in Figure 10 suggest that when the initial imbalance $y_0 - x_0$ is less than $\frac{1-1/\alpha}{1+1/\alpha} = -\frac{\alpha-1}{\alpha+1}$ then with high probability consensus on X-majority is achieved. This suggests that the result of Theorem 3 extends to values $\alpha > 1$ if we redefine success as "species X wins".

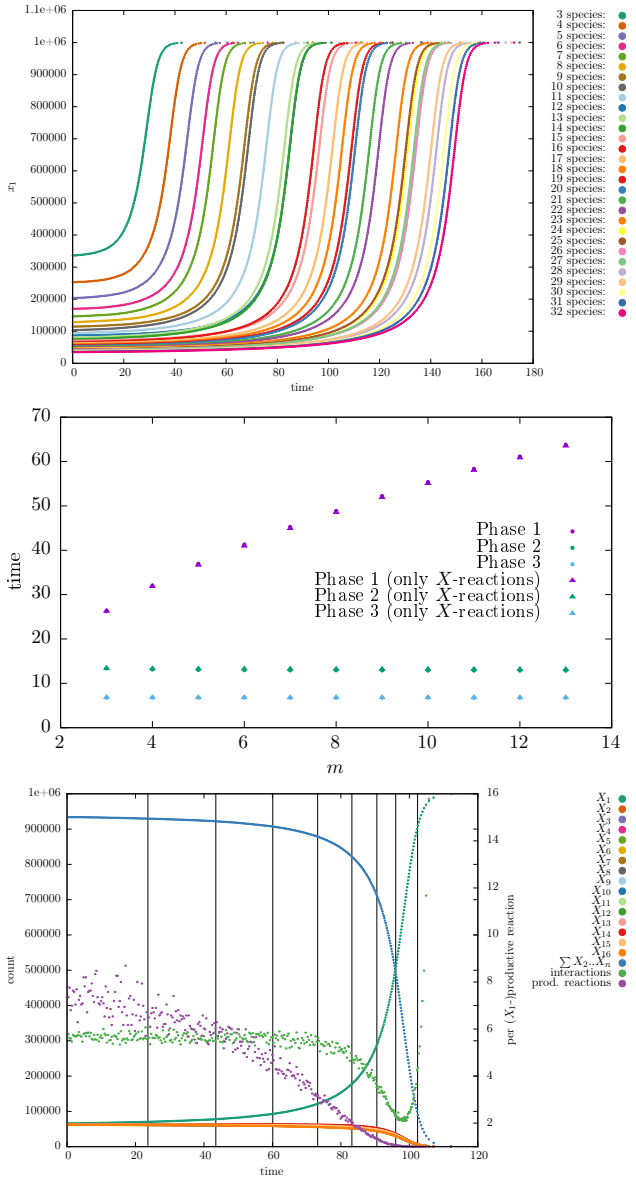

Fig. 8 Top: Comparison of executions of the m-species-Tri CRN for various numbers of species ($m = 3, \ldots, 32$), with the initial gap between the majority species X and any other species equal to $\sqrt{n \ln n}$ with $n = 1,000,000$. Middle: comparison of running time of different phases of m-species-Tri with $n = 100,000$ based on 50,000 experiments. "Only X-reactions"refers to a multi-valued trimolecular CRN without reactions that do not involve majority species. The comparison shows that reactions between non-majority species do not have much influence in the running time of the system. Bottom: More detailed view of the execution of the protocol with 16 species. The figure also shows the average number of interactions per productive reaction and average number of productive reactions per X-productive reaction. The vertical lines show the ends of subphases of phase 1.

10 Conclusions and Directions for Future Work

In this work we have provided a straightforward analysis of a tri-molecular CRN for Approximate Majority, as

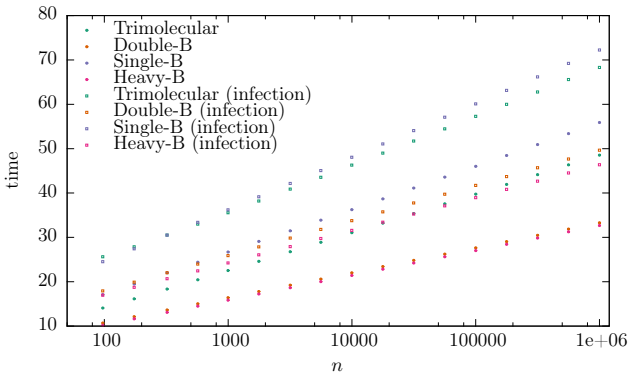

Fig. 9 Average time of the Tri, Double-B , Single-B and Heavy-B CRNs for Approximate Majority, with initial imbalance $\sqrt{n \ln n}$, and with and without initiation by infection. Each point in the plot is an average over 50,000 trials.

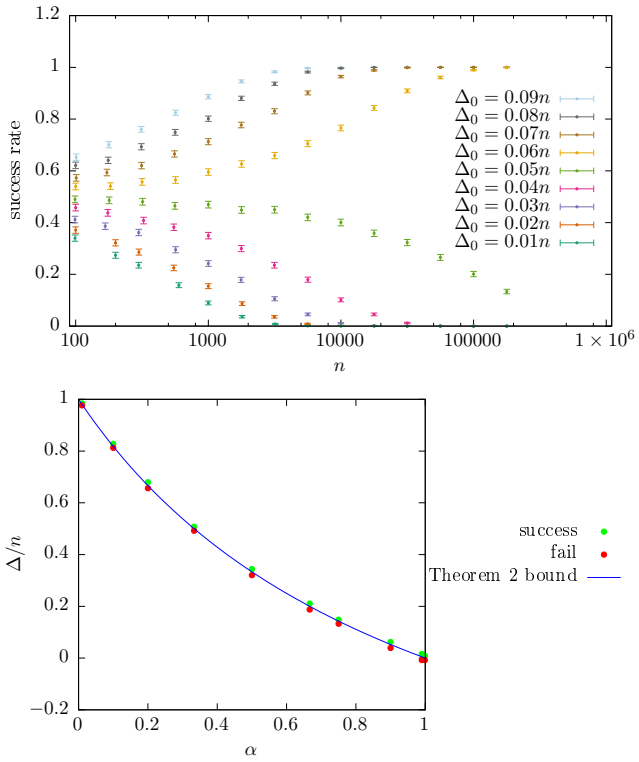

Fig. 10 Tri-molecular CRN with unequal rates of the two reactions: reaction (1) has rate α and reaction (2) has rate 1. The top plot shows convergence of the success rate when $\alpha = 0.9$ and the initial imbalance ($\Delta_0 = x_0 - y_0$) is one of the following linear functions: $0.01n, 0.02n, \ldots, 0.09n$. The plot shows that with the initial imbalance at least $0.06n$, the success rate tends to 1 as n grows, and with the initial imbalance at most $0.05n$, it tends to 0. The bottom plot is constructed from a series of such experiments for various values of α and linear functions used for the initial imbalance. The green dots show values of the initial imbalance for which success rate tends to 1 (with success rate at least 0.99 for $n = 100,000$) and red dots show such values for which success rate tends to 0 (with success rate at most 0.01 for $n = 100,000$). The blue line shows the bound obtained in Theorem 3.

well as CRNs for several variants of the problem: when reaction rates are uncertain, when molecules become active by infection, when there are multiple species, and when molecules may react in a Byzantine fashion. We have also provided simpler analyses of bi-molecular CRNs that have been well-studied in previous work, obtaining slightly improved results in the process. Perhaps the biggest contribution of the paper is to demonstrate how analysis simplicity is achieved by first focusing on the simplest of all the CRNs, namely the tri-molecular CRN, and then leveraging this in the remaining analyses.

Of course, there are many combinations of the problem variants that we have studied (for example, multiple species some fraction of which react in a Byzantine fashion) that would be interesting to analyse. We are confident that each such combination could be analysed using minor variations of the techniques that we have used here; establishing this rigorously could be useful future work.

Our techniques may also be useful for proving correctness of yet other variants, such as the Chen et al. strand displacement implementation of Double-B [11], which involves so-called fuel species and waste products in addition to molecules that represent the species of the CRN, or CRNs in which some or all of the reactions are reversible. For example, if the blank-producing reaction (0') of Double-B is made reversible, the modified CRN appears to still be both correct and efficient, while having the additional nice property that a stable state with neither X-consensus nor Y-consensus cannot be reached, even with very low probability. On the other hand, some caution needs to be applied when reversing reactions. For instance, making reactions (0'x) and (0'y) of Single-B reversible can lead to a system that fluctuates around a state with an equal number of Xs and Ys, and some ratio of Bs. This would happen when the rate of reversed reactions (0'x) and (0'y) is greater or equal to the rate of reactions (1') and (2'). We believe that our analyses can easily generalize to these scenarios.

Acknowledgement

We thank Frederik Mallman-Trenn for helpful discussions about the literature on synchronous models and their relationship to our asychronous models.

References

1. D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Physical Chemistry*, 81:2340–2361, 1977.
2. D. Angluin, J. Aspnes, Z. Diamadi, M. J. Fischer, and R. Peralta. Computation in networks of passively mobile finite-state sensors. *Distributed Computing*, 18(4):235–253, 2006.
3. M. Cook, D. Soloveichik, E. Winfree, and J. Bruck. Programmability of chemical reaction networks. *Algorithmic Bioprocesses*, pages 543–584, 2009.
4. D. Soloveichik, M. Cook, E. Winfree, and J. Bruck. Computation with finite stochastic chemical reaction networks. *Nat Comput*, 7, 2008.
5. L. Cardelli and A. Csikász-Nagy. The cell cycle switch computes approximate majority. *Nature Scientific Reports*, 2, 2012.
6. D. Angluin, J. Aspnes, and D. Eisenstat. Fast computation by population protocols with a leader. In *Dolev S. (eds) Distributed Computing (DISC), Lecture Notes in Computer Science*, volume 4167, pages 61–75. Springer, Berlin, Heidelberg, 2006.
7. L. Cardelli, M. Kwiatkowska, and L. Laurenti. Programming discrete distributions with chemical reaction networks. In *Rondelez Y., Woods D. (eds) DNA Computing and Molecular Programming, Lecture Notes in Computer Science*, volume 9818, pages 35–51. Springer, Cham, 2016.
8. D. Soloveichik, G. Seelig, and E. Winfree. DNA as a universal substrate for chemical kinetics. *PNAS*, 107(12):5393–5398, 2010.
9. D. Alistarh, J. Aspnes, D. Eisenstat, R. Gelashvili, and R. L. Rivest. Time-space trade-offs in population protocols. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2560–2579, 2017.
10. D. Angluin, J. Aspnes, and D. Eisenstat. A simple population protocol for fast robust approximate majority. *Distributed Computing*, 21(2):87–102, July 2008.
11. Y.-J. Chen, N. Dalchau, N. Srinivas, A. Phillips, L. Cardelli, D. Soloveichik, and G. Seelig. Programmable chemical controllers made from DNA. *Nature Nanotechnology*, 8(10):755–762, 2013.
12. A. Condon, M. Hajiaghayi, D. Kirkpatrick, and J. Manuch. Simplifying analyses of chemical reaction networks for approximate majority. In *23rd International Conference on DNA Computing and Molecular Programming, Lecture Notes in Computer Science, Volume 10467*, pages 189–209. Springer-Verlag, 2017.
13. B. Doerr, L. A. Goldberg, L. Minder, T. Sauerwald, and C. Scheideler. Stabilizing consensus with the power of two choices. In *Proceedings of the Twenty-third Annual ACM Symposium on Parallelism in Algorithms and Architectures*, SPAA '11, pages 149–158, New York, NY, USA, 2011. ACM.
14. L. Becchetti, A. Clementi, E. Natale, F. Pasquale, R. Silvestri, and L. Trevisan. Simple dynamics for plurality consensus. *Distributed Computing*, pages 1–14, 2016.
15. L. Becchetti, A. E. F. Clementi, E. Natale, F. Pasquale, and L. Trevisan. Stabilizing consensus with many opinions. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 620–635, 2016.
16. E. Perron, D. Vasudevan, and M. Vojnovic. Using three states for binary consensus on complete graphs. In *Proceedings of the 28th IEEE Conference on Computer Communications (INFOCOM)*, pages 2527–2535, 2009.
17. G. B. Mertzios, S. E. Nikoletseas, C. L. Raptopoulos, and P. G. Spirakis. Determining majority in networks with local interactions and very small local memory. *Distributed Computing*, 30(1):1–16, 2017.
18. J. Cruise and A. Ganesh. Probabilistic consensus via polling and majority rules. *Queueing Systems*, 78(2):99–120, 2014.
19. M. Draief and M. Vojnovic. Convergence speed of binary interval consensus. *SIAM Journal on Control and Optimization*, 50(3):1087–1109, 2012.
20. N. van Kampen. Stochastic processes in physics and chemistry (revised edition), 1997.
21. W. Feller. *An Introduction to Probability Theory and its Applications*, volume 1. Wiley, New York, 3rd edition, 1968.
22. H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.
23. W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.