# Simplifying Analyses of Chemical Reaction Networks for Approximate Majority

Anne Condon, Monir Hajiaghayi, David Kirkpatrick and Ján Maňuch

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

**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 simpler, and more intuitive proofs of correctness and efficiency for two bi-molecular CRNs for Approximate Majority, including that of Angluin et al. Key to our approach is to show how the bi-molecular CRNs essentially emulate a tri-molecular CRN with just two reactions and two species. Our results improve on those of Angluin et al. in that they hold even with an initial gap of $\Omega(\sqrt{n\log n})$. Our analysis approach, which leverages the simplicity of a tri-molecular CRN to ultimately reason about bi-molecular CRNs, may be useful in analyzing other CRNs too.

**Keywords:** Approximate Majority, Chemical Reaction Networks, Population Protocols

## 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 analyzed 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 a new, simpler analysis of CRNs for Approximate Majority.

### 1.1 CRNs and Population Protocols

A CRN is specified as a finite set of chemical reactions, such as those in Figure 1. 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]. Second, 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 1(c) are equally likely reactions triggered by an interaction involving one molecule of species $X$ and one of species $Y$. Soloveichik et al. [8]'s method 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].

| | | |
|---|---|---|
| | | $X+Y \xrightarrow{1/2} X+B$ (0'x) |
| | $X+Y \to B+B$ (0') | $X+Y \xrightarrow{1/2} Y+B$ (0'y) |
| $X+X+Y \to X+X+X$ (1) | $X+B \to X+X$ (1') | $X+B \to X+X$ (1') |
| $X+Y+Y \to Y+Y+Y$ (2) | $Y+B \to Y+Y$ (2') | $Y+B \to Y+Y$ (2') |
| (a) Tri-molecular CRN. | (b) Double-B CRN. | (c) Single-B CRN. |

**Fig. 1.** A tri-molecular and two 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.

### 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 analyze are symmetric with respect to $X$ and $Y$.

Angluin et al. [10] proposed and analyzed the Single-B CRN of Figure 1(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., resulting in some chemical changes) 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 Single-B reaches $X$-majority consensus. They also showed *efficiency*: with "high" probability $1 - n^{-\Omega(1)}$, for any initial gap value $x - y$, Single-B reaches consensus within $O(n \log n)$ interaction events. They also proved correctness and efficiency in more general settings, such as in the presence of $o(\sqrt{n})$ Byzantine agents.

Doerr et al.'s [11] "median rule"protocol for stabilizing consensus with two choices in a distributed setting involves rules that are identical to the interactions of our tri-molecular protocol of Figure 1(a). Their model differs somewhat from that of CRNs in that interactions happen in rounds, in which each process (molecule) initiates exactly one interaction with two other processes chosen uniformly at random. They provide a simple and elegant analysis of the protocol, showing that it achieves consensus with high probability in their model within $O(\log n)$ rounds. They note that the consensus value agrees with that of the initial majority when the initial gap is $\omega(\sqrt{n \log n})$. Doerr et al. did not analyze protocols in which interactions involve just two processes.

Several others have subsequently and independently studied the problem; we'll return to related work after describing our own contributions.

### 1.3 Our Contributions

We analyze three CRNs for Approximate Majority: a simple tri-molecular CRN whose reactions involve just the two species $X$ and $Y$ that are present initially, and two bi-molecular CRNs, which we call Double-B and Single-B, that use an additional "blank" species $B$ – see Figure 1. As noted earlier, the Single-B CRN is the same as that of Angluin et al. The Double-B CRN is symmetric even in

the PP setting, and was among the earliest CRN algorithms constructed with strand displacement chemistry, by Chen et al. [12].

Our primary motivation is to provide the simplest and most intuitive proofs of correctness and efficiency that we can, with the hope that simple techniques can be adapted to reason about CRNs for other problems. A bonus is that our results apply with high probability when the initial gap is $\Omega(\sqrt{n \log n})$, and thus are a factor of $\sqrt{\log n}$ stronger than Angluin et al.'s results in this situation. We do not concern ourselves with smaller initial gaps, but note that even with no initial gap we can still expect efficiency, since the expected number of interaction events until a gap of $\sqrt{n \log n}$ is reached is $O(n \log n)$. This would be true even if there were no bias in favour of reaction (1') as $x$, the majority species, increases. We suspect that the complexity of Angluin et al.'s proof stems from the case when the initial gap is small ($o(\sqrt{n \log n})$), and the fact that they show efficiency with high probability, rather than expected efficiency for such an initial setup.

First, in Section 3 we analyze the tri-molecular CRN of Figure 1(a). Intuitively, its 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) 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 exactly the ratio of $X$-molecules to $Y$-molecules.

We analyze the CRN in three phases. In the first phase we model the evolution of the gap $x - y$ as a sequence of random walks with increasing bias of success (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 of each random walk, we model it 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.

Then in Section 4 we analyze the bi-molecular CRNs of Figure 1 by relating them to the tri-molecular CRN. For the Double-B CRN, we show that with high probability, after a short initial start-up period and continuing almost until consensus is reached, the number of $B$'s is at least proportional to $y$ and is at most $n/2$, in which case reaction events are reactions (1') or (2') with probability $\Omega(1)$. Moreover, 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. Essentially the same argument applies to Single-B. We present empirical results in Section 5.

Our analysis of the tri-molecular protocol is quite similar to that of Doerr et al.'s median rule algorithm, although the models of interaction are different. We discuss the similarities in Section 6, as well as directions for future work.

### 1.4 Related Work

Perron et al. [13] analyze 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. [14] 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 [15] 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 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 [9, 16, 14, 11], or analyze more general multi-valued consensus problems [10, 17, 18, 11].

## 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 $s_i$ 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, 19] 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 the Appendix 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). 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 and $v$ denotes the total volume of the solution. Accordingly, if $n$ and $v$ remain fixed, 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 $v = \Theta(n)$, 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 are led to 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, 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 $c \to^r c'$. 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 the CRNs that we analyze, there is some order $o$ such that for every reaction $(s, t)$ of $\mathcal{R}$, $s_1 + s_2 + \ldots s_m = t_1 + t_2 + \ldots t_m = o$. 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 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 $c \to^r c'$.

### 2.3 Analysis Tools

We will use the following well-known property of random walks, Chernoff tail bounds on functions of independent random variables, and Azuma's inequality.

**Lemma 1 (Asymmetric one-dimensional random walk [20](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 [21]).** *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) $\mathbb{P}[S_N \geq (1+\delta)\mu] \leq \exp(-\frac{\delta^2\mu}{3})$.*

**Lemma 3 (Azuma's inequality [22]).** *Let $Q_1, \ldots, Q_k$ be independent random variables, with $Q_r$ taking values in a set $A_r$ for each $r$. Suppose that the (measurable) function $f : \Pi A_r \to R$ satisfies $|f(x) - f(x')| \leq c_r$ whenever the vectors $x$ and $x'$ differ only in the $r$th coordinate. Let $Y$ be the random variable $f(Q_1, \ldots, Q_k)$. Then , for any $t > 0$,*

$$\mathbb{P}[|Y - E[Y]| \geq t] \leq 2\exp\Big(-2t^2/\sum_{r=1}^{k} c_r^2\Big).$$

## 3 Approximate Majority Using Tri-molecular Reactions

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

**Theorem 1.** *For any constant $\gamma > 0$, there exists a constant $c_\gamma$ such that, provided the initial molecular count of $X$ exceeds that of $Y$ by at least $c_\gamma\sqrt{n \lg n}$, a computation of the tri-molecular CRN reaches a consensus of $X$-majority, with probability at least $1 - n^{-\gamma}$, in at most $c_\gamma n \lg n$ interaction events.*

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 even 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 $(x-1)/(x+y-2) \geq x/(x+y)$, provided $x \geq y$.

We divide the computation into a sequence of three, slightly overlapping and possibly degenerate, phases, where $c_\gamma$, $d_\gamma$ and $e_\gamma$ are constants depending on $\gamma$:

**phase 1** $c_\gamma/2\sqrt{n \lg n} < x - y \leq n(d_\gamma - 2)/d_\gamma$. It ends as soon as $y \leq n/d_\gamma$.
**phase 2** $e_\gamma \lg n < y < 2n/d_\gamma$. It ends as soon as $y \leq e_\gamma \lg n$.
**phase 3** $0 \leq y < 2e_\gamma \lg n$. It ends as soon as $y = 0$.

Of course the assertion that a computation can be partitioned in such a way that these phases occur in sequence holds only with sufficiently high probability. To facilitate this argument, as well as the subsequent efficiency analysis, 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 in phase 1 starts with $x \geq y + 2^t\sqrt{n \lg n}$ and ends with $x \geq y + 2^{t+1}\sqrt{n \lg n}$, where $\lg c_\gamma \leq t \leq (\lg n - \lg \lg n)/2 + \lg((d_\gamma - 2)/(2d_\gamma))$.
- A typical stage in phase 2 starts with $y \leq n/2^s$ and ends with $y \leq n/2^{s+1}$, where $\lg d_\gamma \leq s \leq \lg n - \lg \lg n - \lg e_\gamma - 1$.

Our proof of correctness (the computation proceeds through the specified phases as intended) and our timing analysis (how many interaction events does it take to realize the required number of productive reaction events) 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) [Low probability of unintended phase/stage completion] The relative probability of reactions (1) and (2) is determined by the relative counts of $X$ and $Y$. This allows us to argue, using a biased random walk analysis (Lemma 1 above), that, with high probability, there is no back-sliding; when the computation leaves a phase/stage it is always to a higher indexed phase/stage (cf. Corollaries 1, 2 and 3, below).
(b) [High probability of intended 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, an upper bound, for each phase/stage, on the probability that the phase/stage completes within a specified number of productive reaction events (cf. Corollaries 4, 5 and 6, below).
(c) [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, 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) (cf. Corollary 7, below).

**Lemma 4.** *At any point in the computation, if $x - y = \Delta$ then the probability that $x - y \leq \Delta/2$ at some subsequent point in the computation is less than $(1/e)^{\Delta^2/(2n+2\Delta)}$.*

*Proof.* Since $x - y > \Delta/2$ up to the point when we first have $x - y \leq \Delta/2$, it follows that $x \geq n/2 + \Delta/4$ and $y \leq n/2 - \Delta/4$. We can view the change in $x - y$ resulting from productive reaction events as a random walk, starting at $\Delta$, with success (an increase in $x - y$, following reaction (1)) probability $p$ satisfying $p \geq 1/2 + \Delta/(4n)$.

It follows from Lemma 1 that reaching a configuration where $x - y \leq \Delta/2$ (which entails an excess of $\Delta/2$ failures to successes) is less than $(\frac{1}{1+\Delta/n})^{\Delta/2}$ which is at most $(1/e)^{\Delta^2/(2n+2\Delta)}$.

**Corollary 1.** *In stage $t$ of phase 1, $x - y$ reduces to $2^{t-1}\sqrt{n \lg n}$ with probability less than $1/n^{2^{2t-2}}$.*

**Lemma 5.** *At any point in the computation, if $y = n/k$ then the probability that $y > 2n/k$ at some subsequent point in the computation is less than $(2/(k-2))^{n/k}$.*

*Proof.* Since $y \leq 2n/k$ up to the point when we first have $y > 2n/k$, 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 $y$, following reaction (1)) probability $p$ satisfying $p \geq 1 - 2/k$.

It follows from Lemma 1 that reaching a configuration where $y > 2n/k$ (which entails an excess of $n/k$ failures to successes) is less than $(2/(k-2))^{n/k}$.

**Corollary 2.** *In stage $s$ of phase 2, $y$ increases to $n/2^{s-1}$ with probability less than $(2/(2^s - 2))^{n/2^s}$.*

**Corollary 3.** *In phase 3, $y$ increases to $2e_\gamma \lg n$ with probability less than $(2e_\gamma \lg n/(n - 2e_\gamma \lg n))^{e_\gamma \lg n}$.*

**Lemma 6.** *At any point in the computation, if $x - y = \Delta \leq n/2$ then, assuming that $x - y$ never reduces to $\Delta/2$, the probability that $x - y$ increases to $2\Delta$ within at most $\lambda n$ productive reaction events is at least $1 - \exp(-\frac{(\lambda-2)\Delta^2}{\lambda(2n+\Delta)})$.*

*Proof.* We view the choice of productive reaction as an independent trial with success corresponding to reaction (1), and failure to reaction (2). We start with $x - y = \Delta$ and run until either $x - y = \Delta/2$ or we have completed $\lambda n$ productive reactions. By Lemma 2, the probability that we complete $\lambda n$ productive reactions with fewer than $\lambda n/2 + \Delta/2$ successes, which is necessary under our assumptions if we finish with $x - y < 2\Delta$, is at most $\exp(-\frac{(\lambda-2)\Delta^2}{\lambda(2n+\Delta)})$.

**Corollary 4.** *In stage $t$ of phase 1, assuming that $x - y$ never reduces to $2^{t-1}\sqrt{n \lg n}$, the probability that $x - y$ increases to $2^{t+1}\sqrt{n \lg n}$ within at most $\lambda n$ productive reaction events is at least $1 - \exp(-\frac{(\lambda-2)2^{2t} \lg n}{3\lambda})$.*

**Lemma 7.** *At any point in the computation, if $y = n/k$ then, assuming that $y$ never increases to $2n/k$, the probability that $y$ decreases to $n/k - r$ within $f(n) > 2r$ productive reaction events is at least $1 - \exp(-\Theta(f(n))$.*

*Proof.* We view the choice of productive reaction as an independent trial with success corresponding to reaction (1), and failure to reaction (2). We start with $y = n/k$ and run until either $y = n/k - r$ or we have completed $f(n)$ productive reaction events. (We assume, by Lemma 5, that $y < 2n/k$, and so $p > 1 - 2/k$, throughout.)

By Lemma 2, the probability that we complete $f(n)$ productive reactions with fewer than $(f(n) + r)/2$ successes, which is necessary under our assumptions if we finish with $y > \frac{n}{k-r}$, is at most

$$\exp(-\frac{f(n)(k-2)/2 - (f(n)+r)/2]^2}{2f(n)(k-2)/k}),$$

which is at most $\exp(-\Theta(f(n))$, when $f(n) > 2r$.

**Corollary 5.** *In stage $s$ of phase 2, assuming that $y$ never increases to $n/2^{s-1}$, $y$ decreases to $n/2^{s+1}$, ending stage $s$, in at most $\lambda n/2^s$ productive reaction events, with probability at least $1 - \exp(-\Theta(\lambda n/2^s))$.*

**Corollary 6.** *In phase 3, assuming that $y$ never increases to $2e_\gamma \lg n$, $y$ decreases to 0, ending phase 3 (and the entire computation), in at most $\lambda \lg n$ productive reaction events, with probability at least $1 - \exp(-\Theta(\lambda \lg n))$.*

The following is an immediate consequence of Lemma 2:

**Lemma 8.** *If during some sequence of $m$ interaction events the total probability of all productive reactions is at least $p$ then the probability that the sequence gives rise to fewer than $mp/2$ productive reaction events is no more than $\exp(-mp/8)$.*

**Corollary 7.**
*(i) The $\lambda n$ productive reaction events of each stage of phase 1 occur within $(8/3)d_\gamma \lambda n$ interaction events, with probability at least $1 - \exp(-\lambda n/4)$.*
*(ii) The $\lambda(n/2^s)$ productive reaction events of stage $s$ of phase 2 occur within $(16/3)\lambda n$ interaction events, with probability at least $1 - \exp(-\lambda n/2^{s+2})$.*
*(iii) The $\lambda \lg n$ productive reaction events of phase 3 occur within $(8/3)\lambda n \lg n$ interaction events, with probability at least $1 - \exp(\lambda \lg n/4)$.*

*Proof.* It suffices to observe the following lower bounds on the probability that an interaction event triggers reaction (1) in individual phases/stages:
(i) in phase 1, $x > y \geq n/d_\gamma$, so this probability is greater than $3/(4d_\gamma)$;
(ii) in stage s of phase 2, $x > n(1 - 2^{s-1})$ and $y \geq n/2^{s+1} \geq (\lg n)/2$, so this probability is at least $3/2^{s+3}$;
(iii) in phase 3, $x \geq n - \lg n$ and $y \geq 1$, so this probability is at least $3/(4n)$.

Finally, we prove Theorem 1 using the pieces proved until now.

*Proof (of Theorem 1).*
(i) [Correctness] It follows directly from Corollaries 1 and 4 (respectively, 2 and 5, 3 and 6) that phase 1 (respectively phase 2, phase 3) completes in the

intended fashion, within at most $\lambda n \lg n$ (respectively, $\lambda n$, $\lambda \lg n$) productive reaction events, with probability at least $1 - \exp(-\Theta(c_\gamma \lg n))$ (respectively, $1 - \exp(-\Theta(\lambda n/d_\gamma))$, $1 - \exp(-\Theta(\lambda \lg n))$).
(ii) [Efficiency] It is immediate from Corollary 7 that the required number of productive reaction events in phases 1 2 and 3 occur within $\Theta(\lambda n \lg n)$ interaction events, with probability at least $1 - \exp(-\Theta(\lambda \lg n))$.

## 4 Approximate Majority Using Bi-molecular Reactions

Here we show correctness and efficiency of the Double-B and Single-B CRNs, essentially by showing that both CRNs respect the more abstact tri-molecular CRN of the previous section.

### 4.1 The Double-B CRN

In this section we analyse the behaviour of the Double-B CRN of Figure 1(b):

**Theorem 2.** *For any constant $\gamma > 0$, there exists a constant $c_\gamma$ such that, provided (i) the initial molecular count of $X$ and $Y$ together is at least $n/2$, and (ii) the count of $X$ exceeds that of $Y$ by at least $c_\gamma \sqrt{n \lg n}$, a computation of Double-B reaches a consensus of $X$-majority, with probability at least $1 - n^{-\gamma}$, in at most $c_\gamma n \lg n$ interaction events.*

Comparing with Theorem 1, it becomes clear that the role of the molecule $B$ is simply to facilitate a bimolecular emulation of the tri-molecular CRN. The sense in which Double-B can be seen as emulating the earlier tri-molecular CRN is that we can analyse its behaviour using exactly the same three phases (and the same sub-phase stages) that we used in our tri-molecular analysis.

**Correctness of the emulation** We measure progress throughout in terms of the change in the molecular counts $\hat{x}$, defined as $x + b/2$, and $\hat{y}$, defined as $y + b/2$, noting that reaction (0') leaves these counts unchanged and reactions (1') and (2') change $\hat{x}$ and $\hat{y}$ at exactly half the rate 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 or exceeds the relative probability of reaction (1) to that of (2) in the tri-molecular CRN, and we argue that the total probability of reactions (1') and (2') is at least some constant fraction of the total probability of reactions (1) and (2). This allows us to conclude that Double-B reaches the same conclusion as the tri-molecular CRN, using at most twice as many productive reaction events as the tri-molecular CRN to complete each corresponding phase/stage.

**Efficiency of the emulation** We argue that the productive reaction events needed to carry out the emulation of the tri-molecular CRN occur within a number of interaction events that is at most some constant multiple of the number

of interaction events needed to realize the required productive reaction events in the tri-molecular CRN.

This argument is made most simply by setting out bounds on $b$, the molecular count of molecule $B$ that, with high probability, hold after the first $\Theta(n)$ interaction events, and continue to hold thereafter.

Our bounds are summarized in Lemma 9 below. The proof, a straightforward application of Chernoff bounds, is in the Appendix. In the interests of simplicity, the bounds we provide here are not the tightest possible, but are sufficient for us to conclude immediately that the probability of reactions (1') and (2') of Double-B are each at most a constant factor smaller than those of reactions (1) and (2) in the corresponding phases/stages of the tri-molecular CRN.

**Lemma 9.** *Let $I$ be any interval of $n/64$ interaction events of a computation of Double-B. Let $x_0$, $x_e$, $x_{\min}$ and $x_{\max}$, the initial, final, minimum and maximum values of $x$ in the interval $I$ (similarly, for $y$ and $b$). Then for any constant $\gamma > 0$, there exists a constant $f_\gamma$ such that, if $y_0 \geq f_\gamma \lg n$, the following bounds hold with probability at least $1 - 1/n^\gamma$:*

(a) *[Upper bounds] If $b_0 \leq 15n/32$ then $b_e \leq 15n/32$ and $b_{\max} \leq n/2$.*

(b) *[Lower bounds] Even if $b_0 = 0$, $b_e \geq y_e/265$. Furthermore, if $b_0 \geq y_0/265$ then $b_{\min} \geq y_{\max}/292$.*

The efficiency of Double-B follows similarly from the earlier analysis of the tri-molecular CRN presented in Corollary 7. There we observed that it sufficed to bound from below the probability of reaction (1). For the corresponding analysis of Double-B, we observe that in all corresponding phases/stages the probability of reaction (1') is up to a constant factor the same as that of reaction (1). This follows immediately from the upper bound ($n/2$) on $b$, which ensures that the molecular count of $X$ is at least $n/4$, and the lower bound ($y/292$) on $b$, which ensures that the molecular count of $B$ is at least a constant fraction of that of $Y$. The constant $e_\gamma$ that is used in demarking the end of phase 2 and the start of phase 3 will now depend on the constant $f_\gamma$ of Lemma 9, in order to ensure that this lower bound on $b$ holds throughout phase 2 with high probability.

### 4.2 The Single-B CRN

Here, we study the behaviour of Single-B, originally proposed by Angluin et al. [10] and shown in Figure 1(c):

**Theorem 3.** *For any constant $\gamma > 0$, there exists a constant $c_\gamma > \gamma$ such that, provided (i) the initial molecular count of $X$ and $Y$ together is at least $n/2$, and (ii) the count of $X$ exceeds that of $Y$ by at least $c_\gamma \sqrt{n \lg n}$, a computation of the Single-B CRN reaches a consensus of $X$-majority, with probability at least $1 - n^{-\gamma}$, in $c_\gamma n \log n$ interactions.*

Comparing the Double-B and Single-B CRNs, we notice that the only difference is that reaction (0') is replaced by probabilistic reactions (0'x) and (0'y) which are equally likely and thus on average, have no effect on $\hat{x}$ and $\hat{y}$. An advantage of

Single-B is that $B$-majority consensus is never reached [1]. The analysis of Single-B proceeds in phases that are essentially the same as for Double-B, except for the need to account for drift in the gap $\hat{x} - \hat{y}$ caused by fluctuations in the number of (0'x) vs (0'y) reactions. For example, this drift may cause $\hat{x} - \hat{y}$ to initially dip lower when Single-B executes than it does when Double-B executes. To address this, we show in Lemma 10 that, despite the drift, the gap will remain at least $(c_\gamma - \gamma)/2\sqrt{n \lg n}$ with all but exponentially small probability, and accordingly we change the definition of phase 1 to be:

**phase 1** $(c_\gamma - \gamma)/2\sqrt{n \lg n} < \hat{x} - \hat{y} \leq n(d_\gamma - 2)/d_\gamma$. It ends as soon as $\hat{y} \leq n/d_\gamma$.

Further minor adjustments, described in Appendix A.3, do not require any further changes to the definitions of phases and stages.

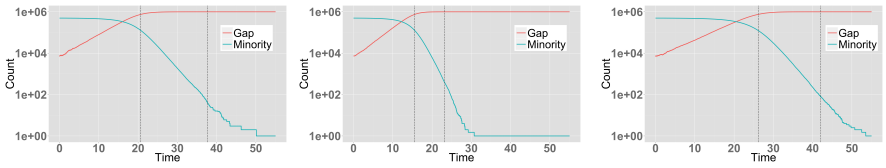
**Lemma 10.** *Starting from $\hat{x} - \hat{y} \geq c_\gamma \sqrt{n \lg n}$, where $c_\gamma > \gamma$, $\hat{x} - \hat{y}$ reduces to $(c_\gamma - \gamma)\sqrt{n \log n}$ within $n$ reaction events with probability less than $1/n^{(\gamma^2)}$.*

*Proof.* Starting from $\hat{x} - \hat{y} \geq c_\gamma \sqrt{n \lg n}$, the probability that $\hat{x} - \hat{y}$ increases is at least as much as the probability that it decreases. As a worst case scenario, we can view the changes in $\hat{x} - \hat{y}$ as an unbiased random walk which starts at $c_\gamma \sqrt{n \lg n}$. Let $Q_1, \ldots, Q_r$ denote independent random variables where $0 \leq r \leq n$ taking values in set $A_r = [1, -1]$. The $Q_r$ satisfy the conditions of Azuma's inequality (Lemma 3) with $c_r = 2$, the expected change $\sqrt{n}$ (assuming an unbiased random walk), and function $Y = f(Q_1, \ldots, Q_n) = \max_{1 \leq r \leq n} |\sum_{i=1}^{r} Q_i|$ which gives us the maximum translation distance over $n$ reaction events. Now, using Azuma's inequality, we can infer that $\mathbb{P}[|Y - \sqrt{n}| \geq \gamma\sqrt{n \lg n}] \leq 1/n^{\gamma^2}$. Thus in our unbiased random walk the maximum distance from the origin is at most $\gamma\sqrt{n \lg n}$ with high probability.

## 5 Empirical Results

Figure 2 illustrates the progress of computations of each of our CRNs in each of the three phases, on a single run. In the first phase, the gap $x - y$ (red line) increases steadily. Once the gap is sufficiently high, phase 2 starts and the count of $y$ for the tri-molecular CRN, and $\hat{y}$ for the bi-molecular CRNs, decrease steadily. In the last phase, as the counts of $y$ and $\hat{y}$ are small, there is more noise in the evolution of $y$ and $\hat{y}$, but they do reach 0. Figure 3 compares time (efficiency) and success rates (probability of correctness) of the three CRNs to reach consensus, as a function of the log of the initial count $n$ of molecules, or the log of the volume. 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 and Single-B CRNs grow as $3.4 \ln n$, $2.4 \ln n$, and $4.0 \ln n$ 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.

[1] We note that although the $B$-majority consensus is reachable in the Double-B CRN, the probability of such an event is easily shown to be very small (i.e., $n^{\Omega(-\lg(n))}$).




(a) (b) (c)

**Fig. 2.** The gap $x - y$ (red line) and minority (count $y$ for tri-molecular CRN and $\hat{y}$ for bi-molecular CRNs) (blue line), as a function of time, of sample runs of the (a) tri-molecular, (b) Double-B, and (c) Single-B CRNs. The initial count is $n = 10^6$, the initial gap $x - y$ is $2\sqrt{n \lg n}$ and parameters $c_\gamma$, $d_\gamma$ and $e_\gamma$ are set to 2, 8, and 2 respectively. The vertical dotted lines demark transitions between phases 1, 2 and 3.

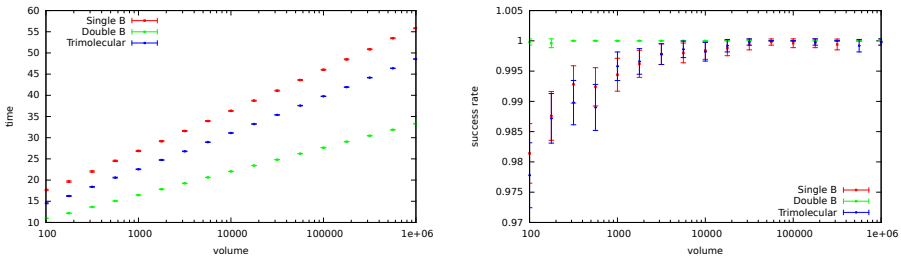



**Fig. 3.** Comparison of the time (left) and success rate, i.e., probability of correctness (right) of Single-B, Double-B and the tri-molecular CRN for Approximate Majority. Each point in the plot is an average over 5,000 trials. The initial configuration has no $B$'s and the imbalance between $X$'s and $Y$'s is $\sqrt{n \ln n}$. Plots show confidence intervals at 99% confidence level.

## 6 Discussion

As noted earlier, Doerr et al. [11] analyse what they call the median rule consensus protocol, which bears strong resemblance to our tri-molecular CRN for approximate majority. The median rule protocol assumes rounds of $n$ concurrent interactions, with each of $n$ participating processes initiating one interaction that involves two additional processes chosen uniformly at random. The result of each such 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 theirs. 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 $Y$ interact and thus are converted to $X$'s. In contrast, the end game is completed in $O(1)$ rounds with

high probability in the Doerr et al. model. More significant differences between the Doerr et al. model versus the CRN and PP models arise when the initial gap $x - y$ is small, a case that we do not analyze and that appears to be significantly harder to handle in the CRN model.

There are several ways in which we can extend our results. Angluin et al. [10] analyze settings in which (i) some agents (molecules) have Byzantine, i.e., adversarial, behaviour upon interactions with others, (ii) some molecules are "activated" (become eligible for reaction) by epidemic spread of signal, and (iii) there are three or more species present initially and the goal is to reach consensus on the most populous species (multi-valued consensus). We believe that our techniques can be generalized to these settings.

Other generalizations are motivated by practicalities of molecular systems. When a CRN is "compiled" to a DNA strand displacement system, it may be that the DNA strand displacement reaction rate constants closely approximate, but are not exactly equal to, the CRN reaction rates. It could be helpful to describe how the initial gap needed to guarantee correct and efficient computations for Approximate Majority with high probability depends on the uncertainty in the rate constants. Also, our techniques may be useful for proving correctness of the Chen et al. strand displacement implementation of Double-B [12], which involves so-called fuel species and waste products in addition to molecules that represent the species of the CRN. Third, it could be useful to analyze variants of the CRNs analyzed here, or other 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 is still 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 $X$s and $Y$s, and some ratio of $B$s. 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'). Again, we believe that our analyses can easily generalize to these scenarios.

## 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. 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.
12. 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, 2013.
13. 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.
14. 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.
15. J. Cruise and A. Ganesh. Probabilistic consensus via polling and majority rules. *Queueing Systems*, 78(2):99–120, 2014.
16. M. Draief and M. Vojnovic. Convergence speed of binary interval consensus. *SIAM Journal on Control and Optimization*, 50(3):1087–1109, 2012.
17. 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.
18. L. Becchetti, A. Clementi, E. Natale, F. Pasquale, R. Silvestri, and L. Trevisan. Simple dynamics for plurality consensus. *Distributed Computing*, pages 1–14, 2016.
19. N. van Kampen. Stochastic processes in physics and chemistry (revised edition), 1997.
20. W. Feller. *An Introduction to Probability Theory and its Applications*, volume 1. Wiley, New York, 3rd edition, 1968.
21. 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.
22. C. McDiarmid. On the method of bounded differences. *London Society Lecture Note Series*, 141:148–188, 1989.

# A Appendix

In this appendix we 1) relate our CRN model to that of Cook et al. [3], 2) prove our lower and upper bounds on the number of $B$ molecules in the Double-B CRN, and 3) prove the lemmas which make the analysis of Single-B parallel to that of the tri-molecular CRN.

## A.1 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. [2] See the example of

[2] 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

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

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).

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

$$X + X + Y \rightarrow^{2/14} X + X + X \ (r_1) \qquad Y + Y + Y \rightarrow^{12/14} X + X + Y \ (r_2)$$
(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 2) 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.

### A.2 Bounds on $b$, the molecular count of $B$, in the Double-B CRN

Here we provide a proof of Lemma 9, omitted from Section 3. We note that the probability that an interaction event in the interval $I$ triggers reaction (0') (respectively, reaction (1'), reaction (2')) is just $xy/\binom{n}{2}$ (respectively, $xb/\binom{n}{2}$, $yb/\binom{n}{2}$). In the following, we simplify calculations by replacing $\binom{n}{2}$ with $n^2/2$.

**Upper bounds on $b$** Note that reaction (0') has probability at most $(n/2)(n/2)/(n^2/2) = 1/2$, so at most $n/64$ new $B$ molecules are produced by reaction (0') over interval $I$, in expectation, and at most $n/32$ are produced, with probability $1 - \exp(\Theta(n))$. Thus, $b_{\max} \leq b_{\min} + n/32$.

Given this, we can clearly assume that $b_{\min} \geq 14n/32$, since otherwise $b_{\max}$ (and, of course $b_e$) is less than $15n/32$. Thus $x + y \leq 18n/32$ throughout the interval, and so reaction (0') has probability at most $(18n/64)^2/(n^2/2)$ which is less than $1/6$. Hence fewer than $2(n/64)/6 = n/192$ new $B$ molecules are produced by reaction (0') over interval $I$, in expectation, and fewer than $n/175$ are produced, with probability $1 - \exp(\Theta(n))$ (here we use a Chernoff upper tail

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 rection 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.

bound). Assuming $b_0 \leq 15n/32$, it follows that $b_{\max} < b_0 + n/175 < n/2$, and so $x + y > n/2$ throughout interval $I$.

It follows that the total probability of reactions (1') and (2') is at least $(n/2)(14n/32)/(n^2/2) = 14/32$ throughout interval $I$, which means that at least $(14/32)(n/64) > n/148$ $B$ molecules are consumed by these reactions, in expectation, and at least $n/160$ are consumed, with probability $1 - \exp(\Theta(n))$, over the course of interval $I$ (here we use a Chernoff lower tail bound). Thus, with probability $1 - \exp(\Theta(n))$, the net change in $b$ is less than $n/175 - n/160 < 0$, and so $b_e < b_0 \leq 15n/32$. We note that this upper bound holds with probability $1 - \exp(-\Theta(n))$, which is stronger than in the statement of the lemma.

**Lower bounds on $b$** Note that $x - y$ is not changed by reaction (0'), and by Lemma 4, it never reaches $(x_0 - y_0)/2$ through reactions (1') and (2'). Therefore, $b_{\max} \leq n/2$, it follows that $x + y \geq n/2$ and hence $x \geq n/4$. We will use this fact throughout.

We first show that even if $b_0 = 0$, $b_e \geq y_e/292$. Since $b_{\max} \leq n/2$ it follows that reaction (2') has probability at most $y_{\max} b_{\max}/(n^2/2) \leq y_{\max}/n$. Thus reaction (2') increases $y$ from its minimum value $y_{\min}$ by at most $y_{\max}/64$, in expectation, and by at most $y_{\max}/32$, with high probability, over the course of interval $I$. Here, the high probability follows from the fact that $y_{\max} \geq y_0 \geq f_\gamma \lg n = \Omega(\log n)$, and application of a Chernoff tail bound. Thus, $y_e \leq y_{\max} \leq y_{\min} + y_{\max}/32$ and so

$$y_{\min} \geq (31/32) y_{\max} = \Omega(\log n). \tag{*}$$

Now suppose that

$$b_{\max} > (1/16) y_{\min}. \tag{**}$$

Thus we also have that $b_{\max} = \Omega(\log n)$, by (*). Since $x + y \leq n$, reactions (1') and (2') together have probability at most $n b_{\max}/(n^2/2)$, and so these reactions reduce $b$ from its maximum value $b_{\max}$ by at most $b_{\max}/32$, in expectation, and by at most $b_{\max}/16$, with high probability, over the course of interval $I$. Here, the high probability follows from the fact that $b_{\max} = \Omega(\log n)$, and application of a Chernoff tail bound. Thus, with high probability,

$$b_e \geq (15/16) b_{\max}. \tag{***}$$

Then,

$$\begin{aligned}
b_e &\geq (15/16) b_{\max} && \text{by (***)} \\
&> (15/16)(1/16) y_{\min} && \text{by (**)} \\
&\geq (15/16)(1/16)(31/32) y_{\max} && \text{by (*)} \\
&> y_{\max}/18 \geq y_e/18.
\end{aligned}$$

On the other hand, suppose that

$$b_{\max} \leq (1/16) y_{\min}. \tag{****}$$

Since reaction (0') has probability at least $x_{\min}y_{\min}/(n^2/2) \geq y_{\min}/(2n)$, reaction (0') increases $b$ by at least $y_{\min}/64$, in expectation, and at least $y_{\min}/128$, with high probability, over the course of interval $I$. Since reactions (1') and (2') together have probability at most $nb_{\max}/(n^2/2) \leq n(y_{\min}/16)/(n^2/2)$ by (****), we know that together they decrease $b$ by at most $y_{\min}/512$, in expectation, and at most $y_{\min}/256$, with high probability, over the course of interval $I$. Here, the high probability follows from the fact that $y_{\min} = \Omega(\log n)$ by (*), and application of a Chernoff tail bound. Thus the net change in $b$ is at least $y_{\min}/128 - y_{\min}/256$, with high probability. Also,

$$\begin{aligned} y_{\min}/128 - y_{\min}/256 &= y_{\min}/256 \\ &\geq (31/32)(1/256)y_{\max} && \text{by (*)} \\ &> y_{\max}/265. \end{aligned}$$

So, $b_e > y_{\max}/265 \geq y_e/265$, even if $b_0 = 0$.

Finally, assume that $b_0 \geq y_0/265$. Let $b'_{\max}$ be the maximum value of $b$ between $b_0$ and $b_{\min}$ in the course of interval $I$. By an argument similar to the one used for equation (***), with high probability, we get

$$b_{\min} \geq (15/16)b'_{\max} \geq (15/16)b_0 \qquad (*****)$$

Therefore, we have

$$\begin{aligned} b_{\min} &\geq (15/16)b_0 && \text{by (*****)} \\ &\geq (15/16)y_0/265 \\ &\geq (15/16)y_{\min}/265 \\ &> (15/16)(31/32)y_{\max}/265. && \text{by (*)} \end{aligned}$$

and so $b > y/292$ throughout interval $I$.

### A.3 Adjustments required for the proof of Single-B

Here we describe additional adjustments to the proof of correctness and efficiency of the tri-molecular CRN that are needed to account for changes to random variables $\hat{x}$ and $\hat{y}$ due to reactions (0'x) and (0'y). Note that reactions (0'x) and (1') increase $\hat{x}$ by 1/2 and decrease $\hat{y}$ by 1/2, while reactions (0'y) and (2') decrease $\hat{x}$ by 1/2 and increase $\hat{y}$ by 1/2.

First, in the proof of the upper $(n/2)$ and lower $(y/292)$ bounds on $b$ in Lemma 9, we simply adjust the probabilities of a change in $\hat{x}$ or $\hat{y}$ to account for reactions (0'x) and (0'y). (We remark that we are able to provide tighter lower and upper bounds on $b$ with respect to variable $y$ , i.e., $\frac{y}{2\alpha} \leq b \leq 2\alpha y$, where $\alpha \geq 20$, and $b = \Omega(\log n)$, for the Single-B CRN - details omitted.) Then, utilizing the lower bound on $b$, Lemma 11 shows that the ratio of total probability of reactions (0'x) and (1') to that of reactions (0'y) and (2') is lower than the ratio of the probability of reaction (1) to that of reaction (2) in the tri-molecular CRN by at most a small constant. Therefore, the analysis of phase 1 of Single-B parallels that of the tri-molecular CRN.

**Lemma 11.** *At any point in the computations, assuming that $\hat{x} - \hat{y} \geq \Delta/2$, the probability that $\hat{x} - \hat{y}$ increases is at least $1/2 + \Theta(\Delta/n)$.*

*Proof.* Let $p$ denote the probability of a success ($\hat{x} - \hat{y}$ increases) and $q$ denote the probability of a failure ($\hat{x} - \hat{y}$ increases). So, given that $x \leq n$, and $y/292 < b$, we have that

$$1)\ \frac{q}{p} = \frac{1/2xy + yb}{1/2xy + xb} \leq 1 - \frac{(\hat{x} - \hat{y})b}{1/2xy + xb} \leq 1 - \frac{(\Delta/2)b}{x(1/2y + b)} \leq 1 - \Theta(\Delta/n),$$

$$2)\ q + p = 1.$$

It follows from equations 1 and 2 that $p \geq 1/2 + \Theta(\Delta/4n)$.

Similarly, we can revise Lemmas 5 and 7 (and their related corollaries) to make the analysis of phases 2 and 3 of Single-B also parallel to those of the tri-molecular CRN–see Lemmas 12 and 13.

**Lemma 12.** *At any point in the computation, if $\hat{y} = n/k$ then the probability that $\hat{y} > 2n/k$ at some subsequent point in the computation is less than $(1 - \Theta(1))^{n/k}$.*

*Proof.* Let $p$ denote the probability of a success ($\hat{y}$ decreases) and $q$ denote the probability of a failure ( $\hat{y}$ increases). So, assuming that $x \leq n$, $\hat{x} - \hat{y} \geq n - n/4k$, and $y < 292b$, we can compute the ratio $q/p$ on a reaction event as follows.

$$\frac{q}{p} = \frac{1/2xy + yb}{1/2xy + xb} \leq 1 - \frac{(\hat{x} - \hat{y})b}{1/2xy + xb} \leq 1 - \frac{(n - 4n/k)b}{n(1/2y + b)} \leq 1 - \Theta(1).$$

By Lemma 1, we conclude that reaching a configuration where $y > 2n/k$ (which entails an excess of $n/k$ failures to successes) is less than $(1 - \Theta(1))^{n/k}$.

**Lemma 13.** *At any point in the computation, if $\hat{y} = n/k$ then, assuming that $\hat{y}$ never increases to $2n/k$, the probability that $\hat{y}$ decreases to $n/k - r$ within $f(n) > \Theta(r)$ reaction events is at least $1 - exp(-\Theta(f(n)))$.*

*Proof.* The proof is completely parallel to the proof of Lemma 7. We only need to compute the probability of a success ($\hat{y}$ decrease). By Lemma 12, $q/p = 1 - \Theta(1)$. So, considering $p + q = 1$, it's straightforward to obtain $p \geq \frac{1}{2} + \Theta(1)$.

Finally, we employ Lemma 8 to complete the proof of efficiency. Using the upper bound on $b$, which confirms that $x \geq n/4$ and the lower bound on $b$, which confirms $b \geq y/292$, we can conclude that the total probability of reactions (0'x), (0'y), (1'), and (2') is at least some constant fraction of the total probability of reactions (1) and (2) in tri-molecular CRN. Therefore, the total number of interactions in Single-B is at most some constant multiple times the required number of interactions in the tri-molecular CRN.