Distributed Community Detection via Metastability of the 2-Choices Dynamics

Emilio Cruciani[1], Emanuele Natale[2], and Giacomo Scornavacca[3]

[1]Gran Sasso Science Institute , emilio.cruciani@gssi.it
[2]Max Planck Institute für Informatik & Université Côte d'Azur, CNRS, I3S, Inria , enatale@mpi-inf.mpg.de
[3] Università degli Studi dell'Aquila , giacomo.scornavacca@graduate.univaq.it

Abstract

We investigate the behavior of a simple majority dynamics on networks of agents whose interaction topology exhibits a community structure. By leveraging recent advancements in the analysis of dynamics, we prove that, when the states of the nodes are randomly initialized, the system rapidly and stably converges to a configuration in which the communities maintain internal consensus on different states. This is the first analytical result on the behavior of dynamics for non-consensus problems on non-complete topologies, based on the first symmetry-breaking analysis in such setting.

Our result has several implications in different contexts in which dynamics are adopted for computational and biological modeling purposes. In the context of *Label Propagation Algorithms*, a class of widely used heuristics for *community detection*, it represents the first theoretical result on the behavior of a distributed label propagation algorithm with quasi-linear message complexity. In the context of *evolutionary biology*, dynamics such as the Moran process have been used to model the spread of mutations in genetic populations [LHN05]; our result shows that, when the probability of adoption of a given mutation by a node of the evolutionary graph depends super-linearly on the frequency of the mutation in the neighborhood of the node and the underlying evolutionary graph exhibits a community structure, there is a non-negligible probability for *species differentiation* to occur.

1 Introduction

Dynamics are simple stochastic processes on networks, in which agents update their own state according to a symmetric function of the state of their neighbors and of their current state, with no dependency on time or on the topology of the network [MT17, Nat17]. In previous decades, in the context of automata networks, this kind of systems has been investigated from a computability point of view, attracting the interest of mathematicians and physicists. Recently it has been subject to a renewed interest from computer scientists, as new techniques for analyzing this class of processes have made possible to answer questions regarding their efficiency and capability as distributed algorithms [DGM$^+$11, BCN$^+$15, BCN$^+$16, BCN$^+$17a, CER$^+$15, CRRS17].

In this work we consider the 2-Choices dynamics (Definition 2), in which at each discrete-time step each agent samples two random neighbors with replacement and, if the two have the same state, the agent adopts that state. The process rapidly converges to *consensus*, i.e., a configuration where all agents have the same state, if the proportion of agents supporting one state exceeds a given function of the second eigenvalue of the graph [CER$^+$15, CRRS17]. Their proofs leverage an interesting property of the 2-Choices dynamics, i.e., that the expected number of agents supporting one state can be expressed as a quadratic form of the transition matrix of a simple random walk on the underlying graph. This fact allows to relate the behavior of the process to the eigenspaces of the graph.

Motivated by questions arising in *graph clustering* and *evolutionary biology*, we exploit the aforementioned relation to show a more fine-grained understanding of the *consensus* behavior of the 2-Choices dynamics. Our new analysis combines symmetry-breaking techniques [BCN$^+$16, CGG$^+$18] and concentration of probability arguments with a linear algebraic approach [CER$^+$15, CRRS17] to obtain the first symmetry-breaking analysis for dynamics on non-complete topologies.

Informal description of Theorem 1. Let the agents of a network initially pick a random binary state and then run the 2-Choices dynamics. If the network has a *community structure* there is a significant probability that it will rapidly converge to an *almost-clustered* configuration, where almost all nodes within each community share the same state, but the predominant states in the communities are different. In other words, with constant probability, after a short time the states of the nodes constitute a labeling which reveals the clustered structure of the network.

The aforementioned probability for the labeling to reveal the community structure can be amplified via *Community-Sensitive Labeling* [BCN$^+$17a], transforming the 2-Choices dynamics into a *distributed label propagation algorithm* with quasi-linear message complexity.

We remark that, because of the stochastic and time-independent behavior of the 2-Choices dynamics, the process eventually leaves almost-clustered configurations and reaches a *monochromatic* configuration in which all agents have the

same state. However, before that happens, we prove that the process remains in almost-clustered configurations for a time equal to a large-degree polynomial in n. Hence, the event that the process leaves the almost-clustered configuration is negligible for most practical applications. This key transitory property of some stochastic processes, called *metastability* [AFPP12, FV15], has recently attracted a lot of attention in the Theoretical Computer Science community.

1.1 Label Propagation Algorithms

Label Propagation Algorithms (LPAs) are a widely used class of algorithms used for *community detection* and inspired by epidemic processes on networks. The generic pattern of such algorithms can be described as follows: First, a label taken from a finite set is assigned to each node according to some *initialization rule*; then the nodes are activated following some *activation rule*; active nodes interact with their neighbors and update their labels according to some *local majority-based update rule*.

After the first algorithm, known in literature as LPA, has been proposed and its effectiveness empirically assessed [RAK07], a new line of research started with the goal of improving the quality of the detected communities and the efficiency of the algorithm [LHLC09, LM10, BRSV11, ŠB11a, ŠB11b, XS13, ZRS$^+$17], and to investigate more general settings, e.g., dynamic networks [XCS13, CDIG$^+$15]. Many variants with small variations on initialization rule, activation rule, and local update rule have been proposed, but they have only been validated experimentally. On the other hand, there exist only few theoretical works. One shows the equivalence of LPA with finding the minima of a generalization of the Ising model, used in statistical mechanics to describe the spin interaction of electrons on a crystalline lattice [TK08]. Another is the first and only rigorous analysis of a variant of LPA on the *Stochastic Block Model*[1] [KPS13]: They propose Max-LPA, i.e., a synchronous version of LPA that follows a deterministic majority rule, and analyze its behavior on $\mathcal{G}_{2n,p,q}$ graphs with parameters $p = \Omega(n^{-1/4+\varepsilon})$ and $q = \mathcal{O}(p^2)$, i.e., on graphs that present very dense communities of constant diameter separated by a sparse cut.

The absence of substantial theoretical progress in the analysis of LPAs is largely due to the lack of techniques for handling the interplay between the non-linearity of the local update rules and the topology of the graph. In this work we look at the 2-Choices dynamics as a *distributed label propagation algorithm*. The randomized nature of the 2-Choices dynamics introduces a major challenge with respect to deterministic rules such as the one of Max-LPA.

1.1.1 Comparison with our result.

Let a and b respectively be the number of neighbors of each agent in its own community and in the other community; let $d := a + b$. The analysis of Max-LPA [KPS13] essentially requires $a \geq n^{3/4-\varepsilon}$ and $b \leq ca^2/n$, for some

[1] The *Stochastic Block Model* is a generative model for random graphs, that produces graphs with community structure.

arbitrary constants ε and c. Our analysis requires[2] $\lambda \leq n^{-1/4}$, which implies $a \geq n^{1/2}$ because of the extremality of Ramanujan graphs, and $b/d \leq n^{-1/2}$. Compared to the analysis of Max-LPA, Theorem 1 holds for much sparser communities at the price of a stricter condition on the cut. Moreover, given the distributed nature of the two algorithms, Max-LPA has a message complexity of $\Omega(m)$, with m the number of edges in the graph that is at least $n^{7/4}$; instead, the message complexity of the 2-Choices dynamics is $\mathcal{O}(n \log n)$ regardless of the actual density of the edges on the graph, since the local update rule only looks at 2 labels. Our algorithm performs an implicit *sparsification* of the graph, an interesting property for the design of sparse clustering algorithms [SZ17], in particular for opportunistic network settings [BCM+18].

1.2 Evolutionary dynamics

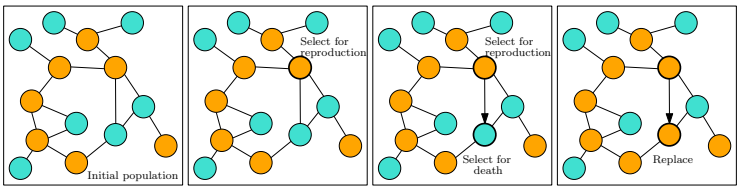

Figure 1: Visual representation of the *Moran process* (adapted from [LHN05]). At each time step an individual is randomly chosen for reproduction according to its *fitness*, and a second individual adjacent to it is randomly chosen for death; the offspring of the first individual then replaces the second. When the underlying network is regular, the process is equivalent to the Voter dynamics [BGKMT16].

Evolutionary dynamics is the branch of genetics which studies how populations evolve genetically as a result of the interactions among the individuals [Dur11]. The study of evolutionary dynamics on graphs started with the investigation of the *fixation probability* of the *Moran process* (Figure 1) on different families of graphs, namely the probability that a new mutation with increased fitness eventually spreads across all individuals in the population [LHN05]. The Moran process has since then attracted the attention of the computer science community due to the algorithmic questions associated to its fixation probability [Gia16, GGG+17].

However, no simple dynamics has been proposed so far in the context of evolutionary graph theory for explaining one of evolution's fundamental phenomena, namely *speciation* [CO04]. Two fundamental classes of driving forces for speciation can be distinguished: *allopatric speciation* and *sympatric/parapatric speciation*. The former, which refers to the divergence of species resulting from geographical isolation, is nowadays considered relatively well understood [SAL+06];

[2] λ is the maximum eigenvalue, in absolute value and different from 1, of the transition matrices of the subgraphs induced by the communities.

on the contrary, the latter, namely divergence without complete geographical isolation, is still controversial [SAL+06, BF07]. In several evolutionary settings the spread of a mutation appears nonlinear with respect to the number of interacting individuals carrying the mutation, exhibiting a drift towards the most frequent phenotypes [CO04]. In this work we look at the 2-Choices dynamics as a quadratic evolutionary dynamics on a clustered graph representing sympatric and parapatric scenarios. We regard the random initialization of the 2-Choices process as two inter-mixed populations of individuals with different genetic pools. The interactions for reproduction purposes between the two populations can be categorized in frequent interactions among individuals within an equal-size bipartition of the populations, i.e., the *communities*, and less frequent interactions between these two communities which, in later stages of the differentiation process, may be interpreted as genetic admixture, i.e. interbreeding between two genetically-diverging populations [MDN+13].

Within the aforementioned framework our Theorem 1 provides an analytical evolutionary graph-theoretic proof of concept on how speciation can emerge from the simple nonlinear underlying dynamics of the evolutionary process at the population level.

1.3 Computational dynamics

Dynamics are rules to update an agent's state according to a function which is invariant with respect to time, network topology, and identity of an agent's neighbors, and whose arguments are only the agent's current state and those of its neighbors [MT17, Nat17]. Simple models of interaction between pairs of nodes in a network have been studied since the first half of the 20th century in statistical mechanics [Lig12] and in the second half in diverse sciences, such as economics and sociology, where averaging-based opinion dynamics such as the DeGroot model have been investigated [Fre56, Deg74, Jac10]. The first study in computer science of a dynamics from a computational point of view is that of a synchronous-time version of the Voter dynamics, where, in each discrete-time round, each node looks at a random neighbor and copies its opinion [HP01]. The Voter dynamics can be regarded as the simplest dynamics, in the sense that there is arguably no simpler rule by which nodes may meaningfully update their state as a function of their neighbors' states. Examples of other dynamics are: Undecided-State [CGG+18], 3-Majority [BCN+17a, BCN+16], 2-Median [DGM+11], Averaging. The Averaging dynamics has been employed for solving the Community Detection task [BCN+17b]. However, we remark that the resulting protocol is not classifiable within LPAs: The configuration space in not described in terms of the finite set of labels initially used by nodes, but by rational values generated from the averaging update rule. Other examples of problems for which dynamics have been successfully employed in order to design an efficient solution are Noisy Rumor Spreading [FN16], Exact Majority [MNRS17], and Clock Synchronization [BKN17].

We now focus on the 2-Choices dynamics, which is the subject of this work. It can arguably be considered the simplest type of dynamics after the

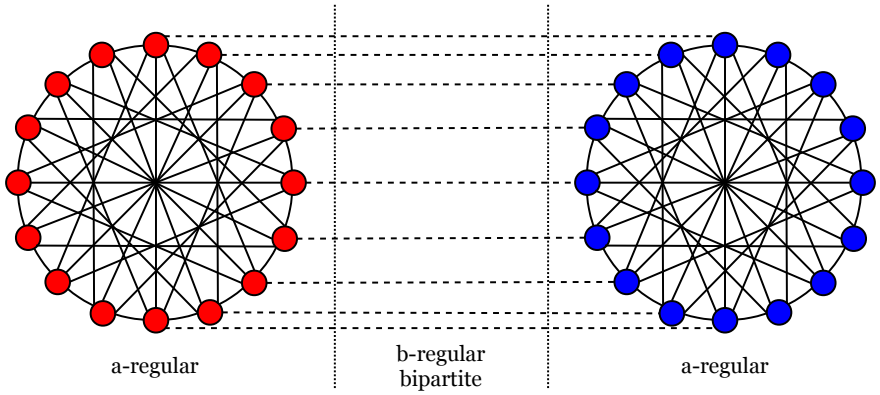

Figure 2: Representation of a $(2n, d, b)$-clustered regular graph where $a := d - b$. Each community induces an a-regular graph while the cut between the two communities induces a b-regular bipartite graph.

VOTER dynamics and, until now, it constitutes one of the few processes whose behavior has been characterized on non-complete topologies [CER14, CER$^+$15, CRRS17]. It has been proven that a network of agents, each with a binary state, will support the initially most frequent opinion with high probability after a polylogarithmic number of rounds whenever the initial *bias* (the advantage of a state on the other) is greater than a function of the network's *expansion* [CER14]. Such result was later refined with milder assumptions on the initial bias with respect to the network's expansion [CER$^+$15] and generalized to more opinions [CRRS17]. Moreover, in core-periphery networks, depending on the strength of the cut between the core and the periphery, a phase-transition phenomenon occurs [CNNS18]: Either one of the colors rapidly spreads over the rest of the network, or a metastable phase takes place, in which both the colors coexist in the network for superpolynomial time.

2 Notation

Let $G = (V, E)$ be a $(2n, d, b)$-*clustered regular graph* (Definition 1) and let us define $a := d - b$. Notice that G is composed by two a-regular communities connected by a b-regular cut (Figure 2) and that when $a > b$ the graph G exhibits a well-clustered structure, i.e., each node has more neighbors in its community than in the other one.

Definition 1 ([BCN$^+$17b])**.** *A $(2n, d, b)$-clustered regular graph is a graph $G = (V, E)$ such that:*

- *$V = V_1 \cup V_2$, $V_1 \cap V_2 = \emptyset$, and $|V_1| = |V_2| = n$;*
- *every node has degree d;*

- *every node in V_1 has b neighbors in V_2 and every node in V_2 has b neighbors in V_1.*

Each node of G maintains a binary *state* that we represent as a color: either *red* or *blue*. We denote the vector of states of all nodes in G at time t as the *configuration* vector $\mathbf{c}^{(t)}$ and we refer to the state of a node $u \in V$ at time t as $\mathbf{c}_u^{(t)} \in \{red, blue\}$. We call $B^{(t)}$ the set of nodes colored *blue* at time t and $R^{(t)}$ the set of nodes colored *red* at time t. For each community $i \in \{1, 2\}$ we define $B_i^{(t)} := V_i \cap B^{(t)}$ and $R_i^{(t)} := V_i \cap R^{(t)}$. We call $s_i^{(t)} = |R_i^{(t)}| - |B_i^{(t)}|$ the *bias* in community i toward color *red*. Given some initial configuration $\mathbf{c}^{(0)}$, we let the nodes of G run the following 2-Choices dynamics.

Definition 2. *The* 2-Choices *dynamics is a local synchronous protocol that works as follows: In each round, each node u chooses two neighbors v, w uniformly at random with replacement; if v and w support the same color, then u updates its own color to their color, otherwise u keeps its previously supported color.*

Notice that the random sequence of configurations $\{\mathbf{c}^{(t)}\}_{t\in\mathbb{N}}$ generated by multiple iterations of the 2-Choices dynamics on G is a Markov Chain with two absorbing states, namely the configurations where all the nodes support the same color, either *red* or *blue*.

Let us now introduce the notion of *almost-clustered configuration*.

Definition 3. *A configuration $\mathbf{c}^{(t)}$ is* almost-clustered *if*

$$|s_i| \geq n - \mathcal{O}\left(\tfrac{\log n}{\log\log n}\right)$$

for each $i \in \{1, 2\}$ and the sign of the biases is different, i.e., $s_1 s_2 < 0$.

Intuitively, *almost-clustered* configurations are such that the vast majority of the nodes in one community is supporting one of the two colors, and the vast majority of nodes in the other community is supporting the other color.

In the rest of the section we introduce the notation used to describe the spectral properties of the transition matrix of the underlying graph G: The analysis in expectation of the process (Lemma 2) exploits such spectral properties and our main result (Theorem 1) makes assumptions on the spectrum of the transition matrix of G.

Let $P = \frac{1}{d}A$ be the transition matrix of a simple random walk on G, where we denote with d the degree of the nodes and with A the adjacency matrix of G. Note that the transition matrix P can be decomposed as follows:

$$P = \begin{pmatrix} P_{1,1} & P_{1,2} \\ P_{2,1} & P_{2,2} \end{pmatrix} = A + B = \begin{pmatrix} P_{1,1} & 0 \\ 0 & P_{2,2} \end{pmatrix} + \begin{pmatrix} 0 & P_{1,2} \\ P_{2,1} & 0 \end{pmatrix},$$

where A is the transition matrix of the communities if we disconnect them, while B is the transition matrix of the bipartite graph connecting the two communities. Note that since the cut is regular B is symmetric and $P_{1,2}^{\intercal} = P_{2,1}$.

We denote with $\lambda_1 \geq \ldots \geq \lambda_n$ the eigenvalues of the transition matrix of the subgraph induced by the first community $\bar{P}_{1,1} := \frac{d}{a} P_{1,1}$ and with $\mu_1 \geq \ldots \geq \mu_n$ the eigenvalues of the transition matrix of the subgraph induced by the second community $\bar{P}_{2,2} := \frac{d}{a} P_{2,2}$. Since both $\bar{P}_{1,1}$ and $\bar{P}_{2,2}$ are stochastic matrices we have that $\lambda_1 = \mu_1 = 1$. We consider the case in which both the subgraphs induced by the communities are connected and not bipartite; thus it holds that $\lambda_2 < 1$, $\mu_2 < 1$ and that $\lambda_n > -1$, $\mu_n > -1$.

We define $\lambda := \max(|\lambda_2|, |\lambda_n|, |\mu_2|, |\mu_n|)$. The value of λ is a representative of the second largest eigenvalues for both the subgraphs induced by the communities and is closely related to the third largest eigenvalue of P.

In addition to the analysis in expectation, we also provide concentration bounds for the behavior of the process. In this context, we say that an event $\mathcal{E}$ happens *with high probability* (for short, *w.h.p.*) if $\mathbf{P}(\mathcal{E}) \geq 1 - \mathcal{O}(n^{-\gamma})$, for some constant $\gamma > 0$.

3 Analysis of the 2-Choices dynamics

In this section we give a high-level overview of the main steps and ideas used for the analysis of the process.

Let G be a clustered regular graph (Definition 1). Let each node in G initially pick a color $\mathbf{c}_u^{(0)} \in \{red, blue\}$ uniformly at random and independently from the other nodes. Then let the nodes of G run the 2-Choices dynamics (Definition 2).

The variance in the initialization suggests that with some constant probability the distribution of the two colors will be slightly asymmetric w.r.t. the two communities, i.e., the first community will have a bias toward a color, while the second community will have a bias toward the other color. Without loss of generality, we consider the case in which s_1 is positive and s_2 is negative, i.e., the first community is unbalanced toward color *red* while the second community is unbalanced toward color *blue*.

Roughly speaking, we show that when the initialization is "lucky", i.e., the biases in the two communities are toward different colors, there is a significant probability that the process will rapidly make the distribution more and more asymmetric until converging to an *almost-clustered* configuration (Definition 3), i.e., a configuration in which, apart from a small number of outliers, the nodes in the two communities support different colors. This behavior of the 2-Choices dynamics is formalized in the following theorem.

Theorem 1 (Constant probability of clustering)**.** *Let $G = (V, E)$ be a connected $(2n, d, b)$-clustered regular graph such that $\frac{b}{d} = \mathcal{O}(n^{-1/2})$ and $\lambda = \mathcal{O}(n^{-1/4})$. Let $c \in \mathbb{N}$ be any constant; let us define the two following events about the* 2-Choices *dynamics on G:*

> *ξ: "Starting from a random initialization the process reaches an* almost-clustered *configuration within $\mathcal{O}(\log n)$ rounds."*

ξ_c: *"Starting from an* almost-clustered *configuration the process stays in* almost-clustered *configurations for n^c rounds."*

For two suitable positive constants γ_1 and γ_2 it holds that

$$\mathbf{P}(\xi) \geq \gamma_1 \text{ and } \mathbf{P}(\xi_c) \geq 1 - n^{-\gamma_2}.$$

Proof. The proof is divided in the following steps:

1. The bias in each community is initially $|s_i| = \Theta(\sqrt{n})$, for each $i \in \{1, 2\}$, and the sign of the biases is different, with constant probability (Lemma 1);
2. The bias in each community becomes $|s_i| = \Theta(\sqrt{n} \log n)$, for each $i \in \{1, 2\}$, in $\mathcal{O}(\log \log n)$ rounds and the sign of the biases is preserved, with constant probability (Lemma 4);
3. The bias in each community becomes $|s_i| \geq n - \mathcal{O}(\log n)$, for each $i \in \{1, 2\}$, in $\mathcal{O}(\log n)$ rounds and the sign of the biases is preserved, with high probability (Lemma 5);
4. The process enters an *almost-clustered* configuration in one single round and lies in the set of *almost-clustered* configurations for the next n^c rounds, with high probability (Lemma 6).

For lack of space the proofs of the lemmas used in Theorem 1 are omitted, but they can be found in the full version of the paper that is publicly available online.

Before starting with the proof, let us introduce some extra notation. Let $\frac{b}{d} \leq c_1 \cdot n^{-1/2}$ for some positive constant c_1, i.e., let every node in each community have at most c_1 neighbors in the opposite community for every $\sqrt{n}$ neighbors in their own. Let $\lambda \leq c_2 \cdot n^{-1/4}$, for some positive constant c_2; note that the hypothesis on λ implies that the subgraph induced by each community is a good expander. Let us define the constant $h := 4(2\sqrt{2}c_1 + c_2^2)$.

We start the analysis of the process by looking at the initialization phase. In particular, in Lemma 1 we show that there is a probability at least constant that the initialization is "lucky", i.e., that the biases in the two communities are $\Theta(\sqrt{n})$ toward different colors. This is true because the Binomial distribution, i.e., the initial distribution of the colors in the graph, is well approximated by a Gaussian distribution, and the latter has a constant probability to deviate from the mean by the standard deviation. The Central Limit Theorem establishes the approximation of the distribution and we are able to quantify it using the Berry-Esseen Theorem.

Lemma 1 (Lucky initialization)**.** *Let $G = (V, E)$ be a $(2n, d, b)$-clustered regular graph and let each node $u \in V$ choose a color $\mathbf{c}_u^{(0)} \in \{red, blue\}$ uniformly at random and independently from the others. Let c_1 and c_2 be two positive constants. Then, there exists a constant γ_1 such that*

$$\mathbf{P}\left(s_1^{(0)} \geq h\sqrt{n} \wedge -s_2^{(0)} \geq h\sqrt{n}\right) \geq \gamma_1.$$

Then, considering a configuration $\mathbf{c}^{(t)}$ at a generic time t, we look at the expected evolution of the process observing the behavior of one single community, but also taking into account the influence of the other. Informally, Lemma 2 gives a bound to the number of nodes that will support the minority color in each community at the next round as a function of all the parameters involved in the process: the number of nodes supporting the minority color in each community at the current round; the number of nodes supporting the same color in the other community at the current round; the expansion of the communities $\lambda \leq c_2 \cdot n^{-1/4}$; the cut density $\frac{b}{d} \leq c_1 \cdot n^{-1/2}$.

The proof of Lemma 2 leverages the fact that the expected number of nodes supporting a given color can be expressed as a quadratic form of the transition matrix of a simple random walk on the graph, allowing to relate the behavior of the process to the expansion of the communities, as exploited in [CER+15, CRRS17].

Lemma 2 (Expected decrease of the minority color)**.** *Let G be a $(2n, d, b)$-clustered regular graph. For any configuration $\mathbf{c}^{(t)}$ we have that*

$$\mathbf{E}\left[|B_1^{(t+1)}| \;\middle|\; \mathbf{c}^{(t)}\right] < |B_1^{(t)}| \left[1 - \tfrac{s_1}{2n} + \tfrac{c_2^2}{\sqrt{n}} + \tfrac{2c_1}{\sqrt{n}}\sqrt{\tfrac{|B_2^{(t)}|}{|B_1^{(t)}|}\left(\tfrac{1}{2} - \tfrac{s_1}{2n} + \tfrac{c_2^2}{\sqrt{n}} + \tfrac{c_1^2|B_2^{(t)}|}{n|B_1^{(t)}|}\right)}\right]$$

and

$$\mathbf{E}\left[|R_2^{(t+1)}| \;\middle|\; \mathbf{c}^{(t)}\right] < |R_2^{(t)}| \left[1 + \tfrac{s_2}{2n} + \tfrac{c_2^2}{\sqrt{n}} + \tfrac{2c_1}{\sqrt{n}}\sqrt{\tfrac{|R_1^{(t)}|}{|R_2^{(t)}|}\left(\tfrac{1}{2} + \tfrac{s_2}{2n} + \tfrac{c_2^2}{\sqrt{n}} + \tfrac{c_1^2|R_1^{(t)}|}{n|R_2^{(t)}|}\right)}\right].$$

It follows from Lemma 2 that the asymmetry in the coloring of the nodes in the two communities continues to grow in expectation. In fact, when in a certain range of values, the bias in the first community increases in expectation at each round while the bias in the second community decreases in expectation at each round, since the minority color in each community decreases. With Lemma 3 we prove that the increase of the bias in the first community and the decrease of the bias in the second community we have shown in expectation in Lemma 2 is multiplicative w.h.p. whenever s_1 satisfies $s_1 \in [h\sqrt{n}, \frac{n}{2}]$ and s_2 satisfies $s_2 \in [-\frac{n}{2}, -h\sqrt{n}]$. With the use of concentration of probability arguments, namely a multiplicative form of the Chernoff bounds [DP09, Lemma 1.1], we show that the number of nodes with the minority color in each community decreases and we use this fact to prove Lemma 3.

Lemma 3 (Probability of multiplicative growth of the bias)**.** *Let $\mathbf{c}^{(t)}$ be a configuration such that $h\sqrt{n} \leq s_1 \leq \frac{n}{2}$ and $h\sqrt{n} \leq -s_2 \leq \frac{n}{2}$. Then, it holds that*

$$\mathbf{P}\left(s_1^{(t+1)} \geq (1 + 1/16)\, s_1 \;\middle|\; \mathbf{c}^{(t)}\right) \geq 1 - e^{-2s_1^2/32^2 n}$$

and

$$\mathbf{P}\left(s_2^{(t+1)} \leq (1 + 1/16)\, s_2 \;\middle|\; \mathbf{c}^{(t)}\right) \geq 1 - e^{-2s_2^2/32^2 n}.$$

Now we know that there is a constant probability that the initialization of the process starts is "lucky" (Lemma 1); we also know that the bias in the first community will increase in expectation and the bias in the second community will decrease in expectation (Lemma 2); moreover, when in a given range, we know that the biases will follow their expected behavior with high probability (Lemma 3).

Then we need to show that the asymmetry in the coloring of the two communities will rapidly increase up to a configuration such that $|s_i| = \Theta(\sqrt{n}\log n)$, for each $i \in \{1, 2\}$, while the sign of the biases is preserved. More formally, with Lemma 4 we prove the internal symmetry breaking of each community. This is possible by applying Lemma 1, and by iterating the application of Lemma 3 for $\mathcal{O}(\log\log n)$ rounds, i.e., until the bias is large enough; finally we handle the stochastic dependency between the two biases during their respective increases in opposite directions.

Lemma 4 (Clustering – Symmetry Breaking)**.** *Starting from an initial configuration where each node $u \in V$ chooses a color $\mathbf{c}_u^{(0)} \in \{red, blue\}$ uniformly at random and independently from the others, it holds that, with constant probability, within $\mathcal{O}(\log\log n)$ rounds the process reaches a configuration $\mathbf{c}^{(t)}$ such that*

$$s_1^{(t)} \geq \sqrt{n}\log n \;\; \textit{and} \;\; -s_2^{(t)} \geq \sqrt{n}\log n.$$

Once the internal symmetry of each community is broken, we show that, with high probability, both biases keep increasing while preserving their sign until they rapidly reach a configuration in which the minority color in each community has at most logarithmic size. This behavior is formally proved in Lemma 5, again through the application of Lemma 2 and Lemma 3.

Lemma 5 (Convergence)**.** *Starting from a configuration $\mathbf{c}^{(t)}$ such that $|s_i| \geq \sqrt{n}\log n$, for each $i \in \{1, 2\}$, there exist two rounds $\tau_1, \tau_2 = \mathcal{O}(\log n)$ such that*

$$|s_1^{(\tau_1)}| \geq n - \log n \;\; \textit{and} \;\; |s_2^{(\tau_2)}| \geq n - \log n$$

and the sign of the biases is preserved, with high probability.

Finally, with Lemma 6 we show that the number of wrongly colored nodes in each community drops to $\mathcal{O}(\log n/\log\log n)$ in one single round (by approximating it with a Poisson random variable through an application of Le Cam's Theorem) and then, with high probability, the process enters a *metastable* phase in which the only possible configurations are *almost-clustered*; this will last for any polynomial number of rounds. In other words, even if a few nodes in each community will continue to change color, almost all the nodes in one community will support one color while almost all the nodes in the other community will support the other color. Note that this quantity is *tight*: It is possible to prove that, within any polynomial number of rounds, there will be a round in which at least $\Omega(\log n/\log\log n)$ nodes in each community will have the wrong color.

Lemma 6 (Metastability)**.** *Let $c \in \mathbb{N}$ be any constant. Starting from a configuration $\mathbf{c}^{(t)}$ such that $|s_i| \geq n - \log n$ for each $i \in \{1, 2\}$, for the next n^c rounds the process lies in the set of configurations such that*

$$|s_i| \geq n - \mathcal{O}\left(\frac{\log n}{\log \log n}\right)$$

and the sign of the bias is preserved, with high probability.

More formally, through Lemma 5 and Lemma 6 we can finally prove that $\mathbf{P}(\xi) \geq \gamma_1$ and $\mathbf{P}(\xi_c) \geq 1 - n^{-\gamma_2}$ for any constant c, concluding the proof of Theorem 1. □

4 Distributed Label Propagation Algorithm via Community-Sensitive Labeling

We showed that, starting from a random initialization, the 2-Choices dynamics reaches an *almost-clustered* configuration within $\mathcal{O}(\log n)$ rounds with constant probability. This result is tight, given that there is constant probability that the two communities converge to the same color. Similarly to Lemma 1, it holds that with constant probability both the biases are unbalanced toward the same color, i.e., $s_1^{(0)} \geq h\sqrt{n}$ and $s_2^{(0)} \geq h\sqrt{n}$. It means that a suitable variant of Lemma 4 shows that there is constant probability that within $\mathcal{O}(\log \log n)$ rounds the process reaches a configuration such that $s_1^{(t)} \geq \sqrt{n} \log n$ and $s_2^{(t)} \geq \sqrt{n} \log n$. Then, Lemma 5 and Lemma 6 show that the system gets quickly stuck in a configuration where almost all nodes have the same color. This is a proof that, given the symmetric nature of the process, we need some luck in the initialization to reach an *almost-clustered* configuration.

In order to get an algorithm that works w.h.p. we sketch how to use the results of the previous sections to build a Community-Sensitive Labeling [BCM+18] within $\Theta(\log n)$ rounds. A Community-Sensitive Labeling (CSL) is made up by a labeling of the nodes and a predicate that can be applied to pairs of labels; it holds that, for all but a small number of outliers, the predicate is satisfied if the nodes belong to the same community, and it is not satisfied if the nodes belong to different communities.

Theorem 2 (LPA via CSL)**.** *Let $G = (V, E)$ be a connected and nonbipartite $(2n, d, b)$-clustered regular graph such that $\frac{b}{d} = \mathcal{O}(n^{-1/2})$ and $\lambda = \mathcal{O}(n^{-1/4})$. Let $\mathbf{c}^{(0)}$ be the initial configuration, where each node $u \in V$ picks a* vector *of colors $\mathbf{c}_u^{(0)} \in \{red, blue\}^\ell$ sampled uniformly at random and independently from the other nodes, such that $\ell = c \log n$ for some positive constant c. Consider the resulting vector after $\Theta(\log n)$ rounds of independent parallel runs of the* 2-Choices *dynamics, each one working on a different component of the vector: For all the pairs of nodes but a polylogarithmic number, it holds that the vectors of nodes in the same community are equal while the vectors of nodes in different communities are different.*

Sketch of proof. As for the first part of the predicate, it is a simple application of Theorem 1. Indeed, at least one of the $\Theta(\log n)$ runs of the 2-Choices dynamics ends in an *almost-clustered* configuration with probability $1 - \gamma^{-\Theta(\log n)} = 1 - n^{-\Theta(1)}$. As for the second part we show that no matter if the process reaches an almost-clustering, nodes in the same community will have the same color with high probability. This is consequence of Lemma 5 and of the following one, which we can prove by applying a general tool for Markov Chains [CGG$^+$18, Lemma 4.5].

Lemma 7 (Consensus – Symmetry Breaking)**.** *Starting from any initial configuration* $\mathbf{c}^{(0)}$*, within* $\mathcal{O}(\log n)$ *rounds the system reaches a configuration* $\mathbf{c}^{(t)}$ *such that*

$$|s_1^{(t)}| \geq \sqrt{n}\log n \;\; \textit{and} \;\; |s_2^{(t)}| \geq \sqrt{n}\log n,$$

with high probability.

Thus, most pairs of nodes can locally distinguish if they are in the same community with high probability by checking whether their vectors differ on any component. □

5 Conclusions and future work

We focused on providing a proof of concept of how spectral techniques and concentration of probability results can be combined to provide a rigorous analysis of the behavior of dynamics converging to metastable configurations that reflect structural properties of the network. In turns, we identified two important implications of our result, which we discussed in the Introduction and we briefly recall here. In the context of *graph clustering*, it constitute the first analytical result on a distributed *label propagation algorithm* with quasi-linear message complexity, contributing to a deeper understanding of such class of widely applied heuristics to detect communities in networks. In the framework of *evolutionary biology*, it provides a simplistic model of how *species differentiation* may occur as the result of the interplay between the local interaction rule at the population level and the underlying topology that describes such interaction.

A limitation of our approach is the restriction to regular topologies. The regularity assumption greatly simplifies the calculations, which are still quite involved. However, it has been shown in [CER$^+$15] that a similar analysis can be performed for general topologies. Thus, it should be possible to extend our analysis to the irregular case, at the price of a much greater amount of technicalities. For example, it should be possible to prove a generalization of our result to the class of $(2n, d, b, \gamma)$-clustered graphs investigated in [BCN$^+$17b], which relaxes the class of $(2n, d, b)$-clustered graphs by assuming that each node has $d \pm \gamma d$ neighbors of which $b \pm \gamma d$ belongs to the other community. In fact it is possible to bound the second eigenvalue of the graph in a way which approximates (depending on γ) the $(2n, d, b)$-clustered graphs case considered here using [BCN$^+$17b, Lemma C.2]. Another important issue is to get a denser

cut, at least parametrized w.r.t. the number of edges inside each community. This cannot be achieved by slightly changing the analysis of this paper, but requires a different approach, since it is possible to show that the technique used in Lemma 2 brings to a sparse cut. Finally, an interesting direction is the use of domination arguments, perhaps based on coupling techniques, to generalize our result to more general dynamics which interpolates between the *quadratic* 2-Choices dynamics and the *linear* Voter dynamics [BCE+17]. In particular, this latter direction would have more general implications in the practical contexts discussed in this work, namely label propagation algorithms and evolutionary dynamics.

References

[AFPP12] Vincenzo Auletta, Diodato Ferraioli, Francesco Pasquale, and Giuseppe Persiano. Metastability of Logit Dynamics for Coordination Games. In *33rd Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1006–1024, Kyoto, Japan, 2012. SIAM.

[BCE+17] Petra Berenbrink, Andrea Clementi, Robert Elssser, Peter Kling, Frederik Mallmann-Trenn, and Emanuele Natale. Ignore or Comply?: On Breaking Symmetry in Consensus. In *ACM Symposium on Principles of Distributed Computing*, pages 335–344, 2017.

[BCM+18] Luca Becchetti, Andrea E. F. Clementi, Pasin Manurangsi, Emanuele Natale, Francesco Pasquale, Prasad Raghavendra, and Luca Trevisan. Average whenever you meet: Opportunistic protocols for community detection. In *26th Annual European Symposium on Algorithms*, pages 7:1–7:13, 2018.

[BCN+15] Luca Becchetti, Andrea Clementi, Emanuele Natale, Francesco Pasquale, and Riccardo Silvestri. Plurality Consensus in the Gossip Model. In *26th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 371–390, 2015.

[BCN+16] Luca Becchetti, Andrea Clementi, Emanuele Natale, Francesco Pasquale, and Luca Trevisan. Stabilizing Consensus with Many Opinions. In *27th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 620–635, 2016.

[BCN+17a] Luca Becchetti, Andrea Clementi, Emanuele Natale, Francesco Pasquale, Riccardo Silvestri, and Luca Trevisan. Simple dynamics for plurality consensus. *Distributed Computing*, 30(4), August 2017.

[BCN+17b] Luca Becchetti, Andrea Clementi, Emanuele Natale, Francesco Pasquale, and Luca Trevisan. Find Your Place: Simple Distributed Algorithms for Community Detection. In *28th Annual*

ACM-SIAM Symposium on Discrete Algorithms, pages 940–959, January 2017.

[BF07] Daniel I. Bolnick and Benjamin M. Fitzpatrick. Sympatric Speciation: Models and Empirical Evidence. *Annual Review of Ecology, Evolution, and Systematics*, 38:459–487, 2007.

[BGKMT16] Petra Berenbrink, George Giakkoupis, Anne-Marie Kermarrec, and Frederik Mallmann-Trenn. Bounds on the Voter Model in Dynamic Networks. In Ioannis Chatzigiannakis, Michael Mitzenmacher, Yuval Rabani, and Davide Sangiorgi, editors, *43rd International Colloquium on Automata, Languages, and Programming*, volume 55, pages 146:1–146:15, 2016.

[BKN17] Lucas Boczkowski, Amos Korman, and Emanuele Natale. Minimizing Message Size in Stochastic Communication Patterns: Fast Self-Stabilizing Protocols with 3 bits. In *28th Annual ACM-SIAM Symposium on Discrete Algorithms*, January 2017.

[BRSV11] Paolo Boldi, Marco Rosa, Massimo Santini, and Sebastiano Vigna. Layered label propagation: A multiresolution coordinate-free ordering for compressing social networks. In *20th International Conference on World Wide Web*, pages 587–596, 2011.

[CDIG+15] Andrea Clementi, Miriam Di Ianni, Giorgio Gambosi, Emanuele Natale, and Riccardo Silvestri. Distributed community detection in dynamic graphs. *Theoretical Computer Science*, 2015.

[CER14] Colin Cooper, Robert Elsässer, and Tomasz Radzik. The power of two choices in distributed voting. In *41st International Colloquium on Automata, Languages, and Programming*, pages 435–446, 2014.

[CER+15] Colin Cooper, Robert Elsässer, Tomasz Radzik, Nicolas Rivera, and Takeharu Shiraga. Fast consensus for voting on general expander graphs. In *29th International Symposium on Distributed Computing*, pages 248–262, 2015.

[CGG+18] Andrea E. F. Clementi, Mohsen Ghaffari, Luciano Gualà, Emanuele Natale, Francesco Pasquale, and Giacomo Scornavacca. A Tight Analysis of the Parallel Undecided-State Dynamics with Two Colors. In *43st International Symposium on Mathematical Foundations of Computer Science*, 2018.

[CNNS18] Emilio Cruciani, Emanuele Natale, André Nusser, and Giacomo Scornavacca. Phase transition of the 2-Choices dynamics on core-periphery networks. In *17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 777–785, 2018.

[CO04] Jerry A Coyne and H Allen Orr. *Speciation*. Sinauer Associates, Inc, Sunderland, Mass, 1 edition edition, 2004.

[CRRS17] Colin Cooper, Tomasz Radzik, Nicolás Rivera, and Takeharu Shiraga. Fast plurality consensus in regular expanders. In *31st International Symposium on Distributed Computing*, 2017.

[Deg74] Morris H. Degroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

[DGM+11] Benjamin Doerr, Leslie Ann Goldberg, Lorenz Minder, Thomas Sauerwald, and Christian Scheideler. Stabilizing consensus with the power of two choices. In *23rd Annual ACM Symposium on Parallelism in Algorithms and Architectures*, pages 149–158, 2011.

[DP09] Devdatt P Dubhashi and Alessandro Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

[Dur11] Richard Durrett. *By Richard Durrett - Probability Models for DNA Sequence Evolution: 2nd (second) Edition*. Springer-Verlag New York, LLC, November 2011.

[FN16] Pierre Fraigniaud and Emanuele Natale. Noisy Rumor Spreading and Plurality Consensus. In *ACM Symposium on Principles of Distributed Computing*, pages 127–136, 2016.

[Fre56] John R. French. A formal theory of social power. *Psychological Review*, 63(3):181–194, 1956.

[FV15] Diodato Ferraioli and Carmine Ventre. Metastability of Asymptotically Well-Behaved Potential Games. In *Mathematical Foundations of Computer Science 2015*, Lecture Notes in Computer Science, pages 311–323. Springer Berlin Heidelberg, 2015.

[GGG+17] Andreas Galanis, Andreas Gbel, Leslie Ann Goldberg, John Lapinskas, and David Richerby. Amplifiers for the Moran Process. *J. ACM*, 64(1):5:1–5:90, March 2017.

[Gia16] George Giakkoupis. Amplifiers and Suppressors of Selection for the Moran Process on Undirected Graphs. *arXiv:1611.01585 [cs, math, q-bio]*, November 2016. arXiv: 1611.01585.

[HP01] Yehuda Hassin and David Peleg. Distributed probabilistic polling and applications to proportionate agreement. *Information and Computation*, 171(2):248 – 268, 2001.

[Jac10] Matthew O. Jackson. *Social and Economic Networks*. Princeton Univers. Press, 2010.

[KPS13] Kishore Kothapalli, Sriram V Pemmaraju, and Vivek Sardeshmukh. On the analysis of a label propagation algorithm for community detection. In *International Conference on Distributed Computing and Networking*, pages 255–269, 2013.

[LHLC09] Ian X. Y. Leung, Pan Hui, Pietro Liò, and Jon Crowcroft. Towards real-time community detection in large networks. *Phys. Rev. E*, 79:066107, Jun 2009.

[LHN05] Erez Lieberman, Christoph Hauert, and Martin A. Nowak. Evolutionary dynamics on graphs. *Nature*, 433(7023):312–316, January 2005.

[Lig12] Thomas M. Liggett. *Interacting Particle Systems*. Springer Science & Business Media, 2012.

[LM10] Xin Liu and Tsuyoshi Murata. Advanced modularity-specialized label propagation algorithm for detecting communities in networks. *Physica A: Statistical Mechanics and its Applications*, 389(7):1493 – 1500, 2010.

[MDN+13] Simon H. Martin, Kanchon K. Dasmahapatra, Nicola J. Nadeau, Camilo Salazar, James R. Walters, Fraser Simpson, Mark Blaxter, Andrea Manica, James Mallet, and Chris D. Jiggins. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Research*, 23(11):1817–1828, November 2013.

[MNRS17] George B Mertzios, Sotiris E Nikoletseas, Christoforos L Raptopoulos, and Paul G Spirakis. Determining majority in networks with local interactions and very small local memory. *Distributed Computing*, 30(1):1–16, 2017.

[MT17] Elchanan Mossel and Omer Tamuz. Opinion exchange dynamics. *Probability Surveys*, 14:155–204, 2017.

[Nat17] Emanuele Natale. *On the Computational Power of Simple Dynamics*. Ph.D. Thesis, Sapienza University of Rome, 2017.

[RAK07] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, 76:036106, Sep 2007.

[SAL+06] Vincent Savolainen, Marie-Charlotte Anstett, Christian Lexer, Ian Hutton, James J. Clarkson, Maria V. Norup, Martyn P. Powell, David Springate, Nicolas Salamin, and William J. Baker. Sympatric speciation in palms on an oceanic island. *Nature*, 441(7090):210–213, May 2006.

[ŠB11a] Lovro Šubelj and Marko Bajec. Robust network community detection using balanced propagation. *The European Physical Journal B*, 81(3):353–362, 2011.

[ŠB11b] Lovro Šubelj and Marko Bajec. Unfolding communities in large complex networks: Combining defensive and offensive label propagation for core extraction. *Physical Review E*, 83(3):036103, 2011.

[SZ17] He Sun and Luca Zanetti. Distributed Graph Clustering and Sparsification. *arXiv:1711.01262 [cs]*, November 2017.

[TK08] Gergely Tibly and Jnos Kertsz. On the equivalence of the label propagation method of community detection and a potts model approach. *Physica A: Statistical Mechanics and its Applications*, 387(19):4982 – 4984, 2008.

[XCS13] Jierui Xie, Mingming Chen, and Boleslaw K. Szymanski. Labelrankt: Incremental community detection in dynamic networks via label propagation. In *ACM Workshop on Dynamic Networks Management and Mining*, pages 25–32, 2013.

[XS13] Jierui Xie and Boleslaw K. Szymanski. Labelrank: A stabilized label propagation algorithm for community detection in networks. In *IEEE Network Science Workshop*, pages 138–143, 2013.

[ZRS+17] Xian-Kun Zhang, Jing Ren, Chen Song, Jia Jia, and Qian Zhang. Label propagation algorithm for community detection based on node importance and label influence. *Physics Letters A*, 381(33):2691–2698, 2017.