Epsilon-stability and the speed of learning in network games

Théophile T. Azomahou

Maastricht University—UNU-MERIT, 6211 TC Maastricht, The Netherlands

Daniel C. Opolot*

Maastricht University—UNU-MERIT, 6211 TC Maastricht, The Netherlands

This Version—April 2014

Abstract: This paper introduces *epsilon-stability* as a generalization of the concept of stochastic stability in learning and evolutionary game dynamics. An outcome of a model of stochastic evolutionary dynamics is said to be epsilon-stable in the long-run if for a given model of mistakes it maximizes its invariant distribution. We construct an efficient algorithm for computing epsilon-stable outcomes and provide conditions under which epsilon-stability can be approximated by stochastic stability. We also define and provide tighter bounds for *contagion rate* and *expected waiting time* as measures for characterizing the short-run and medium-run behavior of a typical stochastic evolutionary model.

Keywords: Stochastic evolution, network games, epsilon-stable sets, expected waiting time, metastability, contagion rate.

JEL classification: C70, D00.

1. Introduction

Multiplicity of equilibria is a well known phenomenon in economic models of strategic interactions. A typical example is economic environments with strategic complementarities, whereby the payoff a strategy or an action generates is a non-decreasing function of the number of players who adopt it. The question as to which among the possible equilibria gets selected or is much more likely in a long-run is of high interest in game theoretic modeling. In their pioneering work Foster and Young (1990), Kandori *et al.* (1993) and Young (1993) showed that a disequilibrium process in which agents learn their opponent's play and subsequently revise their strategies—generally known as *a stochastic evolutionary model*—can

*Corresponding author: Opolot D. C., Maastricht University—UNU-MERIT, Keizer Karelplein 19, 6211 TC Maastricht, The Netherlands, (opolot@merit.unu.edu), Tel. +31 433884440, Fax +31 433884499.

be employed as a mechanism for equilibrium selection.[1] The basic idea of stochastic evolution is that agents play their "optimal" strategies with high probability and with a small probability—also known as the *mutation rate*—they randomize among the strategy set—according to the *mistakes distribution*.[2] The solution concept that is normally employed in such evolutionary models is that of *stochastic stability*, according to Foster and Young (1990). An outcome is said to be stochastically stable if the long-run probability with which it occurs does not vanish as the mutation rates tend to zero. It is therefore that within which the learning process spend the most time in the long-run.

The present paper aims to develop a general framework and convergence measures that circumvent the limitations of stochastic evolutionary models. The first limitation concerns the robustness of stochastic stability as a solution concept. Bergin and Lipman (1996) pointed out that the long-run stable outcome of an evolutionary model strictly depends on the specifications made about the mutation rates. Stochastic stability relies on the assumption that the stationary distribution converges uniformly in the limit of mutation rates. When this assumption is relaxed it is possible to make any outcome stable in the long-run by suitably choosing the mutation rates structure.[3] Another criticism in relation to robustness, that has not been discussed in the literature, concerns the mistakes distribution. In most papers (if not all) in the literature, it is assumed that the mistakes distribution does not play a significant role in determining the long-run stable set. We demonstrate in the motivational example of section 2 below that such an assumption is misleading, and that the long-run stable outcome also strictly depends on the assumptions made about the mistakes distribution. More specifically, we show that for every specification of the mutation rates, there exists a mistakes distribution for which an arbitrary outcome is stable in the long-run. The second limitation of stochastic stability, in particular the assumption of vanishing mutation rates, concerns the convergence rate to stationarity. This assumption generally implies that the convergence rate of the learning process to its stationary distribution becomes too low, casting doubts on whether the long-run properties of such models carry any realizable economic implications.

The contribution of this paper is twofold. First, we define a solution concept—to be called *epsilon-stability* or in short *ε-stability*—that is a generalization of stochastic stability. An outcome is said to be *ε-stable* if for a given model of mistakes, it maximizes the stationary

[1] Since its development as an equilibrium selection mechanism, several applications and similar approaches to stochastic evolutionary dynamics (such as bargaining, signaling, contagion and social innovation) have been explored. See for example Vega-Redondo (1997), Nöldeke and Samuelson (1997), Kandori and Rob (1998), Alós-Ferrer and Weidenholzer (2008), Huck *et al.* (2012).

[2] By optimal we mean a strategy that is prescribed by the given learning rule, such as best-reply and imitation dynamics.

[3] The response in the literature to the criticism by Bergin and Lipman (1996) has been to construct classes of models within which the outcomes of stochastic stability are robust. Examples include the learning models with adjustment costs (van Damme and Weibull, 2002), and models in which the mutation rates are a function of a single parameter (Maruta, 2002; Blume, 2003).

distribution of the model. It is therefore a state or a subset of states within which the learning process spends the most time in the long-run. But unlike stochstic stability, ε-stability does not require the assumption of vanishing mutation rates. Implying that a stochastically stable set is ε-stable only under strict restrictions on the model of mistakes, and in Proposition 2 we show conditions under which ε-stability can be approximated by stochastic stability.

We construct an algorithm for computing ε-stable sets, that is independent of the model of mistakes. Unlike the computation of stochastically stable sets, the computational process for ε-stable sets requires determining both the sizes and "depths" of basins of attraction. More specifically, we define the *cost of a transition* from one basin of attraction to another as a function of the distance between them—to be called the *diameter*—and the collective probability associated with the transition. The collective probability is captured by the average of individual revision probabilities. We demonstrate that the average probability is a sufficient statistic for capturing the effect of the collective probabilities. In addition to its computational convenience, average probabilities also permit heterogeneity in probabilities of mistakes across agents. For example they can depend on agents' positions in the network for the case of local interactions.

We then use the constructed costs of transitions between basins of attraction to define four algorithmic measures that can be employed to compute the long-run ε-stable sets. The *resistance* of the basin of attraction, is defined as the minimum cost of exiting the basin of attraction starting from its corresponding *metastable set*.[4] The *coresistance* of the basin of attraction is the maximum cost of a direct transition from any other metastable set into its boundaries. The third measure is the *path potential* of the directed path connecting two basins of attraction, and is defined as the total cost of that path minus the total of minimum deviations from it. It is a measure of how accessible or reachable a basin of attraction is from another through a given path. The fourth measure is the *stochastic potential* of a basin of attraction, which is computed following the combinatorial methods of Freidlin and Wentzell (1984). An efficient algorithm for computing the long-run ε-stable set is then that which combines the above four algorithmic measures in ascending order of computational complexity.

The second contribution of this paper is to provide convergence measures that can be used to characterize the short-run and medium-run behavior of a typical stochastic evolutionary model. ε-stability concerns the long-run properties of the learning process (more specifically the stationary distribution), but as as noted above in the motivational example of section 2 below, the time it takes to attain the stationary distribution can be unrealistically long in some cases. Under such circumstances, it becomes essential to focus on the short-run and medium-run behavior of the process. To this end, we introduce the measures of *metastability* and *contagion rate*. The metastability of a metastable set is the minimum expected time the

[4] Metastable sets are an equivalent of limit sets in the model of learning with mistakes. They are subsets in which the process spends extended amounts of time before making a transition to another.

process spends inside the boundaries of its basin of attraction, and the contagion rate within a basin of attraction is the speed at which the process converges to its quasi-stationary distribution once it has entered the boundaries of a given basin of attraction. The metastability captures the transition between basins of attraction and hence the medium-run behavior of the learning process. The contagion rate on the other hand captures the dynamics within the basins of attraction, and hence the short-run behavior of the learning process. In particular, it can be viewed as a measure of how fast a strategy diffuses across the population once its threshold has been attained. In Theorem 2 we derive a tight lower bound for metastability, while Theorem 3 provides the relationship between the contagion rate and the spectral properties of the associated transition matrix and the interaction structure.

The definition of metastability in this paper is similar to the definition of the *waiting time* in Ellison (1993, 2000). Ellison (1993, 2000) provides an upper bound for waiting time while assuming vanishing mutation rates, which makes the result rather specific to the case of state-independent mutation rates. The direct implication of the result in Ellison (1993, 2000) is that the metastable set with the largest basin of attraction is the most metastable. On a contrary, Theorem 2 of this paper provides a tighter lower bound for metastability that is independent of the model of mistakes and the assumption of vanishing mutation rate is not necessary in the proof of the result. The expression we provide also implies that the metastable set with the largest basin of attraction is not necessarily the most metastable.

Apart from stochastic stability, there exist other forms of evolutionary equilibrium selection. Most notably the models that consider the equilibrium properties of the learning process as the population size becomes large, such as Binmore and Samuelson (1997). The expressions we provide for the expected waiting time in Theorems 2 can be directly employed in such analysis. Finally, Theorems 2 and 3 can together be employed to characterize strategic diffusion in networks. The specific cases in the literature are Morris (2000), Lee *et al.* (2003), Montanari and Saberi (2010) and Young (2011). For example the finding in Morris (2000), Montanari and Saberi (2010) and Young (2011) that strategic diffusion is faster if the network structure is made up of cohesive subgroups, can be derived as a corollary of Theorems 2 and 3.

The remainder of the paper is organized as follows. Section 2 provides a motivational example demonstrating the non-robustness of stochastic stability. In Section 3, we outline the general model for noisy stochastic evolutionary dynamics. Section 4 provides algorithmic characterization for epsilon-stability and its applications to both random matching and local interactions. Section 5 provides definitions and bounds for expected waiting time, metastability and contagion rate. The main proofs are relegated to the Appendix.

2. Motivational example

The criticism with regard to the structure of mutation rates is well discussed in Bergin and Lipman (1996), so we focus on that concerning the mistakes distribution and convergence rates. In particular, we demonstrate that even in situations where the mutation rates are state-independent and identical for all agents, there always exists a mistakes distribution for every value of the mutation rate (except zero) for which an arbitrary metastable set gets selected. And that the convergence time of the learning process increases exponentially with the inverse of the mutation rate.

Consider a set of three players $N = \{1, 2, 3\}$ who are uniformly and randomly matched to play the coordination game in Table 1. Let the learning model be that defined in Young (1993), in which each player follows the "Best Reply" dynamics with probability $1 - \varepsilon$ and with probability ε randomizes among the action set. Given the action set $X = \{A, B\}$, let $\mathbf{x}$ be the action profile, $u_i(x, \mathbf{x})$ be the payoff to $i \in N$ when playing action x, and denote by $BR_i(x, \mathbf{x})$ for the probability that i plays the Best Reply action. That is $BR_i(x, \mathbf{x}) = 1$ if $x \in \arg\max_{x_i} u_i(x_i, \mathbf{x})$ and zero otherwise. Let $\mathcal{P}(x, \mathbf{x})$ be the mistakes distribution (or more specifically the probability mass function) identical to all players. Assume also that the payoff is identical for all players, such that the probability of playing action $x \in X$ given the mutation rate ε and profile $\mathbf{x}$ is

$$\mathbb{P}(x, \mathbf{x}) = (1 - \varepsilon) BR(x, \mathbf{x}) + \varepsilon \mathcal{P}(x, \mathbf{x}) \tag{1}$$

From the payoff structure in Table 1, the states $\vec{A}$ and $\vec{B}$ in which all agents play A and B respectively are the equilibria (limit singleton sets) of the best reply dynamics without mistakes. They are also the metastable sets of the model of learning with mistakes whose dynamics is governed by (1). When players are matched uniformly and randomly, there are only four relevant states or action profiles; $\vec{A}$, $\vec{B}$, BBA in which one agent plays A and the other two play B, and AAB in which one agent plays B and the other two play A. Write $\mathbf{X}$ for the state space in the order $\mathbf{X} = (\vec{A}, AAB, BBA, \vec{B})$. Let P_ε be the transition matrix of the Markov chain $(\mathbf{X}, P_\varepsilon)$ induced by the dynamics in (1). P_ε is irreducible meaning that $(\mathbf{X}, P_\varepsilon)$

Table 1: The action profile (A, A) is risk-dominant.

		player j	
		A	B
player i	A	4 , 4	3 , 0
	B	0 , 3	5 , 5

has a unique stationary distribution π_ε whose structure is determined by the (normalized) left eigenvector corresponding to the leading eigenvalue of P_ε. That is $\pi_\varepsilon P_\varepsilon = \pi_\varepsilon$.

The first step in the computation of the stochastically stable set normally is to determine the basins of attraction limit sets. From the payoff in Table 1, it can easily be shown through best reply argument that the basin of attraction $\tilde{A}$ of $\vec{A}$ is $\tilde{A} = \{\vec{A}, AAB, BBA\}$, and for $\vec{B}$, $\tilde{B} = \{\vec{B}\}$. Let $e = \varepsilon\mathcal{P}(B, \vec{A})$, $f = \varepsilon\mathcal{P}(B, AAB)$ and $h = \varepsilon\mathcal{P}(A, \vec{B})$. Then the transition matrix P_ε induced by the dynamics in (1) is given by,

$$P_\varepsilon = \left(\begin{array}{ccc|c} (1-e)^3 & 3(1-e)^2e & 3(1-e)e^2 & e^3 \\ (1-e)(1-f)^2 & (1-f)(e+2f-3ef) & f(2e+f-3ef) & ef^2 \\ (1-f)(2f+h-3fh) & (1-f)^2(1-h) & f^2h & f(f+2h-3fh) \\ \hline h^3 & 3(1-h)h^2 & 3(1-h)^2h & (1-h)^3 \end{array}\right)$$

The dashed lines partition P_ε into block matrices describing transitions within basins of attraction (diagonal block matrices) and between basins of attraction (off-diagonal block matrices). Since the size of $\tilde{A}$ is greater than that of $\tilde{B}$, the computational algorithms in Young (1993) and Ellison (2000) imply that the long-run stable equilibrium (stochastically stable state) is $\vec{A}$.

Now, consider the case in which $\varepsilon = 0.01$. Then there exists a mistakes distribution, for example $\mathcal{P}(B, \vec{A}) = 0.9$, $\mathcal{P}(B, AAB) = 0.9$ and $\mathcal{P}(A, \vec{B}) = 10^{-5}$ for which $\pi(\vec{B}) = 0.71 > \pi(\vec{A}) = 0.28$. If $\varepsilon = 10^{-4}$, then substituting $\mathcal{P}(B, \vec{A}) = 0.9$, $\mathcal{P}(B, AAB) = 0.9$ and $\mathcal{P}(A, \vec{B}) = 10^{-9}$ into P_ε yields $\pi(\vec{B}) = 0.70 > \pi(\vec{A}) = 0.29$.

In other words, for every value of $\varepsilon \in (0, 1)$ there exists a mistakes distribution for which an arbitrary limit set maximizes the stationary distribution (or gets selected in the long-run, in the language of stochastic stability). In general, it is fair to say that when the mistakes distribution is bounded then there exists an ε' such that for any $\varepsilon < \varepsilon'$, stochastic stability is a good approximation to the behavior of the process $(\mathbf{X}, P_\varepsilon)$ in the long-run. We elaborate on this argument in section 4

The question then becomes, how small should ε' be for the convergence (mixing) time of the process to have any economic implications? For the cases above where $\varepsilon = 0.01$ and $\varepsilon = 10^{-4}$, we find the mixing times through simulations to be over 10^8 and 10^{15} respectively. That is the mixing time scales exponentially with the inverse of mutation rate. Such time scales are definitely too high for the properties of the stationary distribution to have any economic implications. Moreover this example has only four states and two metastable sets.

3. The model

Let Γ be an $n-$person game that can be either in strategic form or additively separable preferences. Let $N = \{1, \cdots, i, \cdots, n\}$ be the set of agents, and let X_i be a discrete and finite set of actions available to i. Denote by $t = 1, 2, \cdots$ for the successive periods of play. Agents simultaneously play the game Γ once each period. The strategies that each agent reacts to depends on whether they interact with the entire population or locally through a social network. Let the strategy (or simply the action) chosen by each $i \in N$ at period t be denoted by $x_{i,t}$, and for $x_{i,t} \in X_i$ let $\mathbf{x}_t = (x_{1,t}, \cdots, x_{n,t})$ denote the strategy profile at t. Each $\mathbf{x}_t \in \mathbf{X} = \prod_{i=1}^{n} X_i$ will also be referred to as the *population state* or simply the *state* of the learning process at t, where $\mathbf{X}$ is the state space.

The agents' interaction can be global or local. Under global interactions, each agent is randomly and uniformly matched with every other agent in the population while under local interactions agents react to the strategies of a subset of the population referred to as their *neighborhood*. Local interactions are defined in a graph theoretical manner or simply by a social network. Let $G(n, E)$ be a graph with n vertices, representing the number of agents and E edges linking different pairs of agents such that a graph g_{ij} defines the connection between i and j. If $g_{ij} = 1$ then a directed link exists from i to j, and zero implies otherwise. We thus have a directed network $G(n, E)$ describing the relationship of any one agent with every other agent in the population. The *adjacency matrix* G of an interaction structure with a network topology given by $G(n, E)$ is basically an $n \times n$ matrix with entries being the elements of g_{ij}. The neighborhood of agent i, $\mathscr{N}_i$, is defined as $\mathscr{N}_i = \{j \in n | g_{ij} = 1\}$, and gives the set of players to which i is linked to. The cardinality $\#\mathscr{N}_i = k_i$, is the *degree* of i.

3.1. *Payoff structure*

Let $\mathbf{x}_{-i}$ be the strategy profile of all agents excluding i. The functions $u_i : \mathbf{X} \to \mathbb{R}$ for each i define the payoffs of the game, such that $u_i(x_i, \mathbf{x}_{-i})$ is the payoff of i when he plays x_i and the other players follow strategy profile $\mathbf{x}_{-i}$. In this paper we focus on games in which $u_i(x_i, \mathbf{x}_{-i})$ exhibits strategic complementarity and substitutes. These include multi-action coordination games and linear-quadratic network games. The following examples belong to the category of games that satisfy such conditions.

Consider a coordination game with a binary action set $X_i = \{A, B\}$ homogeneous to all $i \in N$ with the payoff structure in Table 2. Let $v_i(x_i, x_j)$ be the payoff to i from playing action x_i when his opponent $j \in \mathscr{N}_i$ plays actions x_j. Then the total payoff to i from playing x_i when the other players follow strategy $\mathbf{x}_{-i}$ is of the form.

$$u_i(x_i, \mathbf{x}_{-i}) = \sum_{j \in \mathscr{N}_i} J_{ij} v_i(x_i, x_j), \tag{2}$$

The parameter J_{ij} depends on whether players are randomly and uniformly matched with every other in the population (global interactions) in which case $J_{ij} = \frac{1}{n}$, or they interact locally through a network. Note that local interactions can also be random and uniform, in which case $J_{ij} = \frac{1}{k_i}$. Otherwise, $J_{ij} \in [0, 1]$ for all $j \in \mathscr{N}_i$ and for each $i \in N$.

A more general case of network games of strategic complements and substitutes are the linear-quadratic games with a general payoff structure of the form

$$u_i(x_i, \mathbf{x}_{-i}) = s_i(x_i) + \sum_{j \in \mathscr{N}_i} S_i^j(x_i, x_j). \tag{3}$$

Table 2: Payoff structure for the pure coordination game between i and j

		player j	
		A	B
player i	A	a , a	d , c
	B	c , d	b , b

The first term in the sum, $u_i(x_i)$ is the intrinsic utility to i from playing strategy x_i and $S_i^j(x_i, x_j)$ is the network externality or social utility to i from playing action x_i when the neighbor j's action is x_j. The notable examples include the *status* model of Akerlof (1997), the social interactions model of Brock and Durlauf (2001) and the neighborhood effects models discussed in Glaeser and Scheinkman (2001).

3.2. Revision probabilities

The main behavioral assumption of our model is that agents follow the Darwinian dynamics, that is they respond myopically to the past strategies of their opponents. This assumption is for the sake of clarity and the main theorems we present can be easily extended to other forms of learning, such as that in Young (1993) in which agents respond to a bounded history of their opponents strategies. Based on this assumption, we consider general evolutionary dynamics in which agents play the "optimal" strategy with high probability and with a small probability play that which is not necessarily optimal. By "optimal" we mean a strategy that would be prescribed by a given learning rule. For example under Best-Reply dynamics it would be the strategy which maximizes the associated utility function, and under imitation dynamics it would be that which is the most successful in the population or neighborhood. We focus on the case of Best-Reply dynamics in this paper.

Let $BR_i(x_{t+1} = x|\mathbf{x}_t)$ be the probability that i plays action x in the next period under best-reply dynamics given that the current state is $\mathbf{x}_t$. Then

$$BR_i(x_{t+1} = x|\mathbf{x}_t) = \begin{cases} 1 & \text{if } x \in \arg\max_{x_i \in X_i} u_i(x_i, \mathbf{x}_t) \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Let $\varepsilon_i(\mathbf{x})$ be i's state-dependent mutation rate, the probability that i randomizes among the elements of X_i with the conditional distribution defined by $\mathcal{P}_i(x|\mathbf{x})$. Then the revision probabilities for each $i \in N$ under the model of learning with mistakes is defined as

$$\mathbb{P}_i(x_{t+1} = x|\mathbf{x}_t) = (1 - \varepsilon_i(\mathbf{x}_t))BR_i(x_{t+1} = x|\mathbf{x}_t) + \varepsilon_i(\mathbf{x}_t)\mathcal{P}_i(x_{t+1} = x|\mathbf{x}_t) \tag{5}$$

where for each $i \in N$ and $\mathbf{x}_t \in \mathbf{X}$, $\sum_{x \in X_i} \mathcal{P}_i(x_{t+1} = x|\mathbf{x}_t) = 1$. Throughout this paper, we refer to $\varepsilon_i(\mathbf{x}_t)$ as the mutation rate, $\mathcal{P}_i(x_{t+1} = x|\mathbf{x}_t)$ as the mistakes distribution (or more specifically mistakes probability mass function (PMF)) and the product $\mathscr{P}_i(x|\mathbf{x}_t) = \varepsilon_i(\mathbf{x}_t)\mathcal{P}_i(x_{t+1} = x|\mathbf{x}_t)$ as the probability of playing action x by mistake. We denote the vector of mutation rates by $\boldsymbol{\varepsilon} = (\varepsilon_1(\mathbf{x}), \cdots, \varepsilon_n(\mathbf{x}))$. The structural assumptions made about $\boldsymbol{\varepsilon}$ and $\mathcal{P}_i(x_{t+1} = x|\mathbf{x}_t)$ entail what we refer to as the *model of mistakes.*

For the applications in section 4.2 below we work with the revision probabilities of the form

$$\mathbb{P}_i(x_{t+1} = x|\mathbf{x}_t) = (1-\varepsilon)BR_i(x_{t+1} = x|\mathbf{x}_t) + \varepsilon\frac{\exp\left[\beta u_i(x, \mathbf{x}_t)\right]}{\sum_{y \in X_i} \exp\left[\beta u_i(y, \mathbf{x}_t)\right]} \tag{6}$$

The revision probability in (6) captures most of the learning models that have been analyzed in the literature. The models with a state-independent mutation rate such as those in Kandori *et al.* (1993), Young (1993) and Ellison (2000) then corresponds to that in which ε is the mutation rate that is identical to all agents and states, and the mistakes assume the logit distribution with parameter β. When $\beta = 0$ one obtains the case in which the mistakes are uniformly and randomly distributed. To recover the state-dependent mutation rates learning models from (6), such as that in van Damme and Weibull (2002), we simply assume that $\varepsilon \in (0, 1]$ is some constant and the parameter β determines the mutation rate. More specifically, under state-dependent mutation rates we take the limit of the process $(\mathbf{X}, P_{\boldsymbol{\varepsilon}})$ for $\beta \to \infty$ rather than $\varepsilon \to 0$. Setting $\varepsilon = 1$ recovers the *Logit* learning models such as that in Blume (1995).

For the remainder of the paper the discussion will be centered on two Markov chains, $(\mathbf{X}, P)$ and $(\mathbf{X}, P_{\boldsymbol{\varepsilon}})$ induced by best-reply and best-reply with mistakes dynamics respectively. Hereafter, the *mutationless* and *mutation* models respectively. P and $P_{\boldsymbol{\varepsilon}}$ are the transition matrices whose elements $P(\mathbf{x}, \mathbf{y})$ and $P_{\boldsymbol{\varepsilon}}(\mathbf{x}, \mathbf{y})$ are defined by

$$P_{\boldsymbol{\varepsilon}}(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} \mathbb{P}_i(x_{t+1} = y_i|\mathbf{x}_t = \mathbf{x}) \quad \text{for each } y_i \in \mathbf{y} \tag{7}$$

3.3. *Limit sets, $\boldsymbol{\varepsilon}$-stability and stochastic stability*

It is well known and can be easily verified that the dynamics of the mutationless model generates multiple equilibria in games of strategic complements and substitutes such as those listed in subsection 3.1 above. The resulting equilibria are generally referred to as *limit sets* or *recurrent classes*, formally defined as follows.

Definition 1: *A set $\Omega \subset \mathbf{X}$ is a limit set of $(\mathbf{X}, P)$ if $\forall \mathbf{y} \in \Omega$, $\mathbb{P}(\mathbf{x}_{t+1} \in \Omega|\mathbf{x}_t = \mathbf{y}) = 1$, and that $\forall \mathbf{y}, \mathbf{z} \in \Omega$, there exists a $\tau > 0$ such that $\mathbb{P}(\mathbf{x}_{t+\tau} = \mathbf{z}|\mathbf{x}_t = \mathbf{y}) > 0$*

In the coordination game of Table 1 above for example, the limit sets include the singleton sets in which all players play strategy A and in which they all play strategy B. Generally, the limit sets of $(\mathbf{X}, P)$ can include sets that are cycles and those in which players use different strategies. When the interactions are governed by a social network , the number of limit sets is enhanced. In particular, there will exists singleton limit sets in which strategies co-exists and different *cohesive* subgroups adopt different strategies. The limit set of $(\mathbf{X}, P)$ that results depends on the initial state of the process. We denote the set of limit sets of $(\mathbf{X}, P)$ by $\mathbf{\Omega}$.

We refer to the equivalents of limit sets in the mutation model as *metastable sets*, denoted by Ω_ε. The set of all metastable sets of a given $(\mathbf{X}, P_\varepsilon)$ will be denoted by $\mathbf{\Omega}_\varepsilon$. The perturbed process $(\mathbf{X}, P_\varepsilon)$ has a unique invariant distribution $\pi_\varepsilon = \lim_{t\to\infty} \mathbf{q}_0 P_\varepsilon^t$, where $\mathbf{q}_t$ is the vector of probability mass functions at period t . The existence of a unique stationary distribution is a standard fact about *aperiodic-irreducible* Markov chains. The stationary distribution of the process $(\mathbf{X}, P_\varepsilon)$ describes the amount of time it spends in each state in the long-run. The standard approach in the stochastic evolutionary literature is to assume that for small values of the mutation rates, the stationary distribution $(\mathbf{X}, P_\varepsilon)$ can be approximated by its limit $\pi = \lim_{\varepsilon\to 0} \pi_\varepsilon$. Or in the case of state-dependent mutation rates for the dynamics in (6) above $\pi = \lim_{\beta\to\infty} \pi_\varepsilon$. The metastable sets Ω_ε for which $\pi(\Omega_\varepsilon) > 0$ are then said to be *stochastically stable.*

The principal argument behind stochastically stable sets is that, they are sets within which the process $(\mathbf{X}, P_\varepsilon)$ spends the most time in the long-run. They are therefore assumed to correspond to the metastable sets that maximize the stationary distribution. But as discussed in the introduction and in section 2, this approximation is not robust to various models of mistakes. Rather than focusing on the limit properties of π_ε, we characterize its properties and algorithmic computation for any given model of mistakes. We then refer to the metastable sets that maximize π_ε as $\boldsymbol{\varepsilon}$-*stable sets.* Formally.

Definition 2: *Let $\boldsymbol{\varepsilon} = (\varepsilon_1(\mathbf{x}), \cdots, \varepsilon_n(\mathbf{x}))$ be the mutation rates of the process $(\mathbf{X}, P_\varepsilon)$, and let $\mathcal{P}_i(x|\mathbf{x})$ for all $i \in N$, for each $x \in X$ and $\mathbf{x} \in \mathbf{X}$ be the mistakes distribution. Then a metastable set Ω_ε^* is said to be $\boldsymbol{\varepsilon}$-stable if $\Omega_\varepsilon^* = \arg\max_{\Omega_\varepsilon \in \mathbf{\Omega}_\varepsilon} \pi_\varepsilon(\Omega_\varepsilon)$*

An $\boldsymbol{\varepsilon}$-stable set can thus be equivalently defined as that for which $\pi_\varepsilon(\Omega_\varepsilon^*) > \pi_\varepsilon(\Omega_\varepsilon)$ for all $\Omega_\varepsilon \neq \Omega_\varepsilon^*$. A stochastically stable set is $\boldsymbol{\varepsilon}$-stable only under some strict restrictions on the model of mistakes. We construct an algorithm for computing an $\boldsymbol{\varepsilon}$-stable set in next section, and provide applications demonstrating its relationship with stochastically stable sets.

4. Epsilon-stable sets

In this section we prove the first theorem of this paper which provides the conditions for a set to be $\boldsymbol{\varepsilon}$-stable in the long-run. These conditions are then used to construct the computational algorithm for identifying the $\boldsymbol{\varepsilon}$-stable sets. The computation of an $\boldsymbol{\varepsilon}$-stable set can be performed by first determining the stationary distribution π_ε through the *combinatorial methods* of Freidlin and Wentzell (1984), or by use of the properties of quotients of the stationary distribution. The former can identify the unique $\boldsymbol{\varepsilon}$-stable set (if it exists) but it is more computationally demanding than the later, and the reverse is true for the latter. We thus construct an algorithm that combines both methods. In the next subsection we explicitly define the concepts of *resistance, co-resistance*, and *path potential* all of which are associated with the notion of quotients, and the *stochastic potential* that is associated with the notion of spanning trees. But first we formally define the notion of *basins of attraction* and introduce the concept of the *collapsed Markov chain* of the process $(\mathbf{X}, P_\varepsilon)$ that will be the basis of analysis in the following sections.

The basin of attraction $D(\Omega)$ of a limit set Ω of the unperturbed process $(\mathbf{X}, P)$ is $D(\Omega) = \{\mathbf{x} \in \mathbf{X} | \mathbb{P}(\exists T \text{ s.t } \mathbf{x}_t \in \Omega \, \forall \, t > T | \mathbf{x}_0 = \mathbf{x}) = 1\}$.

Similarly, the basin of attraction $D(\Omega_\varepsilon)$ of a metastable set Ω_ε of the perturbed process $(\mathbf{X}, P_\varepsilon)$ is $D(\Omega_\varepsilon) = \{\mathbf{x} \in \mathbf{X} | \mathbb{P}(\exists T \text{ s.t } \mathbf{x}_t \in \Omega \, \forall \, T < t < \infty | \mathbf{x}_0 = \mathbf{x}) > \mathbb{P}(\exists T \text{ s.t } \mathbf{x}_t \in \Omega \, \forall \, T < t < \infty | \mathbf{x}_0 = \mathbf{x})\}$. Without loss of generality, the model of mistakes is such that $D(\Omega_\varepsilon)$ is equal to $D(\Omega)$ in composition. In other words $D(\Omega_\varepsilon)$ is the equivalent of $D(\Omega)$ in the process $(\mathbf{X}, P_\varepsilon)$.

The basins of attraction induce a partition on the state space into disjoint subsets $D(\Omega) \subset \mathbf{X}$. Let $\tilde{\mathbf{x}}$ be the shorthand for $D(\Omega)$ and let $\tilde{\mathbf{X}}$ be the state space consisting of $\tilde{\mathbf{x}}$'s as its states. A *collapsed Markov chain* $(\tilde{\mathbf{X}}, \tilde{P}_\varepsilon)$, derived from $(\mathbf{X}, P_\varepsilon)$ consists of $\tilde{\mathbf{x}}$ as its states and the transition probabilities among them defined as follows (a generalization of the collapsed Markov chain in Aldous and Fill (1994, Chapter 2)):

$$\tilde{P}_\varepsilon(\mathbf{x}, \mathbf{y}) = P_\varepsilon(\mathbf{x}, \mathbf{y}), \tag{8a}$$

$$\tilde{P}_\varepsilon(\mathbf{y}, \tilde{\mathbf{x}}) = \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} P_\varepsilon(\mathbf{y}, \mathbf{x}), \tag{8b}$$

$$\tilde{P}_\varepsilon(\tilde{\mathbf{x}}, \mathbf{y}) = \frac{1}{\pi_\varepsilon(\tilde{\mathbf{x}})} \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} \pi_\varepsilon(\mathbf{x}) P_\varepsilon(\mathbf{x}, \mathbf{y}), \tag{8c}$$

$$\tilde{P}_\varepsilon(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = \frac{1}{\pi_\varepsilon(\tilde{\mathbf{x}})} \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} \sum_{\mathbf{y} \in \tilde{\mathbf{y}}} \pi_\varepsilon(\mathbf{x}) P_\varepsilon(\mathbf{x}, \mathbf{y}) \tag{8d}$$

The following lemma is an immediate consequence of the above definition of a collapsed Markov chain.

Lemma 1: *Let π_ε and $\tilde{\pi}_\varepsilon$ be the stationary distributions of $(\mathbf{X}, P_\varepsilon)$ and $(\tilde{\mathbf{X}}, \tilde{P}_\varepsilon)$ respectively. Then for any $\tilde{\mathbf{x}} \in \tilde{\mathbf{X}}$, $\tilde{\pi}_\varepsilon(\tilde{\mathbf{x}}) = \pi_\varepsilon(\tilde{\mathbf{x}}) = \sum_{\mathbf{x} \in \tilde{\mathbf{X}}} \pi_\varepsilon(\mathbf{x})$*

Proof. See Appendix A.1 □

The following proposition establishes a bound on the ratios or quotients of the stationary distributions of the collapsed process $(\tilde{\mathbf{X}}, \tilde{P}_\varepsilon)$, and will be pivotal in the characterization of long-run $\boldsymbol{\varepsilon}$-stable sets below.

Proposition 1: *Let $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ be any two states in $\tilde{\mathbf{X}}$ such that $\tilde{\mathbf{x}} \cap \tilde{\mathbf{y}} = \emptyset$. Then*

$$\frac{\pi_\varepsilon(\tilde{\mathbf{y}})}{\pi_\varepsilon(\tilde{\mathbf{x}})} \leq \frac{\max_{\mathbf{x} \in \tilde{\mathbf{x}}} P_\varepsilon(\mathbf{x}, \tilde{\mathbf{x}}^c)}{\min_{\mathbf{y} \in \tilde{\mathbf{y}}} P_\varepsilon(\mathbf{y}, \tilde{\mathbf{x}})} \tag{9}$$

where $\tilde{\mathbf{x}}^c$ is the complement of $\tilde{\mathbf{x}}$.

Proof. See Appendix A.2 □

The right hand side of (9) has the following interpretation. Consider any two metastable sets Ω_ε and Ω'_ε and let $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ be their respective basins of attraction. Then $\max_{\mathbf{x}\in\tilde{\mathbf{x}}} P_\varepsilon(\mathbf{x},\tilde{\mathbf{x}}^c)$ is equivalent to the probability of making a transition to $\tilde{\mathbf{x}}^c$ in a single time step given that the process starts from $\mathbf{x}\in\Omega_\varepsilon$. The denominator $\min_{\mathbf{y}\in\tilde{\mathbf{y}}} P_\varepsilon(\mathbf{y},\tilde{\mathbf{x}})$ is equivalent to the probability of making a transition to $\tilde{\mathbf{x}}$ in a single time step given that the process starts from $\Omega'_\varepsilon \in \tilde{\mathbf{y}}$.

The quotients $\frac{\pi_\varepsilon(\tilde{\mathbf{y}})}{\pi_\varepsilon(\tilde{\mathbf{x}})}$ can be easily extended to the notion of paths whenever a path exists from $\tilde{\mathbf{y}}$ to $\tilde{\mathbf{x}}$ whose total probability is less than the direct transition $\tilde{\mathbf{y}}\to\tilde{\mathbf{x}}$. This can be achieved by use of the chain-rule argument. That is, if there exists an intermediate metastable set $\bar{\Omega}_\varepsilon$ with a basin of attraction $\tilde{\mathbf{z}}$ then,

$$(10)\qquad \frac{\pi_\varepsilon(\tilde{\mathbf{y}})}{\pi_\varepsilon(\tilde{\mathbf{x}})} = \frac{\pi_\varepsilon(\tilde{\mathbf{y}})}{\pi_\varepsilon(\tilde{\mathbf{z}})}\frac{\pi_\varepsilon(\tilde{\mathbf{z}})}{\pi_\varepsilon(\tilde{\mathbf{x}})} \le \frac{\max_{\mathbf{z}\in\tilde{\mathbf{z}}} P_\varepsilon(\mathbf{z},\tilde{\mathbf{z}}^c)}{\min_{\mathbf{y}\in\tilde{\mathbf{y}}} P_\varepsilon(\mathbf{y},\tilde{\mathbf{z}})}\frac{\max_{\mathbf{x}\in\tilde{\mathbf{x}}} P_\varepsilon(\mathbf{x},\tilde{\mathbf{x}}^c)}{\min_{\mathbf{z}\in\tilde{\mathbf{z}}} P_\varepsilon(\mathbf{z},\tilde{\mathbf{x}})}$$

In which case $(\tilde{\mathbf{X}},\tilde{P}_\varepsilon)$ starts from $\tilde{\mathbf{y}}$ then to $\tilde{\mathbf{z}}$ and finally to $\tilde{\mathbf{x}}$.

4.1. *Resistance, path potential and stochastic potential*

In this section we introduce and define concepts of *resistance, co-resistance, path potential* and *stochastic potential* that will be used in characterizing the long-run ε-stable sets. Given the definition of the collapsed process $(\tilde{\mathbf{X}},\tilde{P}_\varepsilon)$, all the four concepts are defined on the state space $\tilde{\mathbf{X}}$ of the basins of attraction. The identification of the ε-stable sets is then based on the notion that if a given point set $\tilde{\mathbf{x}}\in\tilde{\mathbf{X}}$ satisfies the conditions of ε-stability, then so does its corresponding metastable set.

Define the (normalized) *diameter* $d(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j)$ of the directed relation $\tilde{\mathbf{x}}_i\to\tilde{\mathbf{x}}_j$ as the fraction of mistakes required to enter the basin of attraction of $\tilde{\mathbf{x}}_j$ starting from the metastable set Ω^i_ε of $\tilde{\mathbf{x}}_i$. Equivalently, $d(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j)$ is the fraction of players required to simultaneously play a different action by mistake for the process $(\mathbf{X},P_\varepsilon)$ to enter the boundary of $\tilde{\mathbf{x}}_j$ given that it is in the state $\mathbf{x}\in\Omega^i_\varepsilon$.

The collective probability associated with the diameter $d(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j)$ or a direct transition $\tilde{\mathbf{x}}_i\to\tilde{\mathbf{x}}_j$ is captured by its average probability. Formally, let $\mathbb{P}_i(x_{t+1}=y|\mathbf{x}_t=\mathbf{x})$ for $y\in\mathbf{y}\in\Omega^j_\varepsilon\in\tilde{\mathbf{x}}_j$ and $y\notin\mathbf{x}\in\Omega^i_\varepsilon\in\tilde{\mathbf{x}}_i$ be the probability that i plays an action belonging to the state in a metastable set $\Omega^j_\varepsilon\in\tilde{\mathbf{x}}_j$ that is different from that played in the state belonging to the metastable set $\Omega^i_\varepsilon\in\tilde{\mathbf{x}}_i$, given that $(\mathbf{X},P_\varepsilon)$ is in $\mathbf{x}\in\Omega^i_\varepsilon$. We can the define the average probability $\mathbb{P}_A(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j)$ as

$$(11)\qquad \mathbb{P}_A(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_i(x_{t+1}=y|\mathbf{x}_t=\mathbf{x}) \quad \text{for } y\in\mathbf{y}\in\Omega^j_\varepsilon \text{ and } y\notin\mathbf{x}\in\Omega^i_\varepsilon$$

Take an example of the coordination game in Table 1 above where $(\Omega_\varepsilon=\vec{A}$ and $\Omega'_\varepsilon=\vec{B}$ are the two main metastable sets under uniform random interactions. The corresponding average probability of the transition between the basins of attraction $\tilde{A} \rightarrow \tilde{B}$ is

$$\mathbb{P}_A(\tilde{A},\tilde{B}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}_i(x_{t+1} = B|\mathbf{x}_t = \vec{A}) \tag{12}$$

Clearly, under mutationless model $(\mathbf{X}, P)$, $\mathbb{P}_A(\tilde{A},\tilde{B}) = 0$ and generally $\mathbb{P}_A(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) = 0$ for all $\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j \in \tilde{\mathbf{X}}$. Under $(\mathbf{X}, P_\varepsilon)$ on the other hand $\mathbb{P}_A(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) = \frac{1}{n}\sum_{i=1}^{n} \mathscr{P}_i(x_{t+1} = y|\mathbf{x}_t = \mathbf{x})$, where $\mathscr{P}_i(x_{t+1} = y|\mathbf{x}_t)$ is defined in (5) above.

Given the diameter and the average probability, we then define the *cost* $c(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$ of a direct transition $\tilde{\mathbf{x}}_i \rightarrow \tilde{\mathbf{x}}_j$ as

$$c(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) = -d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \ln\left(\mathbb{P}_A(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)\right)$$

Given all direct transitions out of a basin of attraction $\tilde{\mathbf{x}}_i$, its resistance $\mathscr{R}(\tilde{\mathbf{x}}_i)$ is defined as the minimum cost over all $\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i$. That is

$$\mathscr{R}(\tilde{\mathbf{x}}_i) = \min_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} \{c(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)\}$$

The *co-resistance* $\mathscr{C}\mathscr{R}(\tilde{\mathbf{x}}_i)$ of $\tilde{\mathbf{x}}_i$ is defined as the maximum cost over all direct transitions to $\tilde{\mathbf{x}}_i$. That is

$$\mathscr{C}\mathscr{R}(\tilde{\mathbf{x}}_i) = \max_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} \{c(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_i)\}$$

To define the path potential of a basin of attraction, we first define the potential associated with each path terminating at the given basin of attraction. Let $\mathscr{H}_{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_j}$ be the set of all directed paths starting from $\tilde{\mathbf{x}}_i \in \tilde{\mathbf{X}}$ and terminating at $\tilde{\mathbf{x}}_j \in \tilde{\mathbf{X}}$, and let $H_{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_j} = (\tilde{\mathbf{x}}_i, \cdots, \tilde{\mathbf{x}}_\kappa, \cdots, \tilde{\mathbf{x}}_j)$ be the typical path in $\mathscr{H}_{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_j}$. Then the path potential $\phi(H_{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_j})$ of $H_{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_j}$ is defined as

$$\phi(H_{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_j}) = \sum_{\kappa=i}^{j-1} \Big(c(\tilde{\mathbf{x}}_\kappa, \tilde{\mathbf{x}}_{\kappa+1}) - \mathscr{R}(\tilde{\mathbf{x}}_{\kappa+1}) \Big)$$

That is, the total cost of the path minus the total of minimum deviations from the path. The path potential is thus a measure of how accessible or reachable a given basin of attraction is from another through that particular path. The logic behind the definition is that the accessibility of $\tilde{\mathbf{x}}_j$ from $\tilde{\mathbf{x}}_i$ through $H_{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_j}$ depends on the total cost associated with $H_{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_j}$ and the likelihood of deviating from $H_{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_j}$. The higher the cost the more difficult it is to reach $\tilde{\mathbf{x}}_j$ through $H_{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_j}$ and the lower the resistance of the basins of attraction that the path traverses to reach $\tilde{\mathbf{x}}_j$ the more likely that the process will follow such a path.

We then define the maximum path potential of any given basin of attraction $\tilde{\mathbf{x}}_j \in \tilde{\mathbf{X}}$ as follows

$$\phi(\tilde{\mathbf{x}}_j) = \max_{\tilde{\mathbf{x}}_i \neq \tilde{\mathbf{x}}_j} \min_{H_{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_j} \in \mathscr{H}_{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_j}} \phi(H_{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_j}) \tag{13}$$

To define the stochastic potential of a given basin of attraction, we need to define a graph $\mathscr{W}$ on the state space $\tilde{\mathbf{X}}$, in which each state $\tilde{\mathbf{x}} \in \tilde{\mathbf{X}}$ is a vertex and the weight $w(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$ of a directed edge $\tilde{\mathbf{x}}_i \to \tilde{\mathbf{x}}_j$ defined as the minimum total cost over all paths $H_{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_j} \in \mathscr{H}_{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_j}$. That is

$$w(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) = \min_{H_{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_j} \in \mathscr{H}_{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_j}} \sum_{\kappa=i}^{j-1} c(\tilde{\mathbf{x}}_\kappa, \tilde{\mathbf{x}}_{\kappa+1})$$

Given $\mathscr{W}$, we can then define the concept of an $\tilde{\mathbf{x}}_i$-tree according to Freidlin and Wentzell (1984) as the spanning tree such that from every $\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i$, there is a unique path directed from $\tilde{\mathbf{x}}_j$ to $\tilde{\mathbf{x}}_i$. Denote by $\mathscr{T}_{\tilde{\mathbf{x}}_i}$ as the set of all $\tilde{\mathbf{x}}_i$-trees on $\mathscr{W}$ and index by τ for a typical $\tilde{\mathbf{x}}_i$-tree in $\mathscr{T}_{\tilde{\mathbf{x}}_i}$. The stochastic potential of $\tilde{\mathbf{x}}_i$ is then defined as

$$\mathscr{S}(\tilde{\mathbf{x}}_i) = \min_{\tau \in \mathscr{T}_{\tilde{\mathbf{x}}_i}} \sum_{(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_\kappa) \in \tau} w(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_\kappa) \tag{14}$$

The following theorem and corollary show how the resistance, co-resistance and path potential of basins of attraction can be used to compute the long-run $\boldsymbol{\varepsilon}$-stable sets. It also shows that the measures of the costs of transitions among basins of attraction (or metastable sets) also satisfy the conditions of the usual spanning tree algorithm.

Theorem 1: *Let* $(\mathbf{X}, P_\varepsilon)$ *be a model of learning with mistakes, and let* $\tilde{\Omega}_\varepsilon$ *be the union of metastable sets* $\Omega_\varepsilon \subset \tilde{\mathbf{x}}$ *for which* $\phi(\tilde{\mathbf{x}}) < 0$.

(i) Then the long-run $\boldsymbol{\varepsilon}$*-stable set of* $(\mathbf{X}, P_\varepsilon)$ *lies in* $\tilde{\Omega}_\varepsilon$.

(ii) The long-run $\boldsymbol{\varepsilon}$*-stable set* Ω^*_ε *of* $(\mathbf{X}, P_\varepsilon)$ *is that among all sets in* $\tilde{\Omega}_\varepsilon$ *whose corresponding basin of attraction* $\tilde{\mathbf{x}}^*$ *is such that* $\mathscr{S}(\tilde{\mathbf{x}}^*) = \min_{\tilde{\mathbf{x}}_i \in \tilde{\mathbf{X}}} \mathscr{S}(\tilde{\mathbf{x}}_i)$.

Proof. See Appendix A.3 □

Corollary 1: *Suppose* $\bar{\Omega}_\varepsilon$ *is the union of metastable sets whose resistance and co-resistance are such that* $\mathscr{R}(\tilde{\mathbf{x}}) > \mathscr{CR}(\tilde{\mathbf{x}})$. *Then the long-run* $\boldsymbol{\varepsilon}$*-stable set of* $(\mathbf{X}, P_\varepsilon)$ *lies in* $\bar{\Omega}_\varepsilon$.

Proof. See Appendix A.3

□

Theorem 1 and Corollary 1 provide necessary and sufficient conditions for identifying the long-run $\boldsymbol{\varepsilon}$-stable sets. Theorem 1 (i) states that all metastable sets whose basins of attraction are such that the maximum path potential is negative, belong to the union set containing the long-run $\boldsymbol{\varepsilon}$-stable set. Corollary 1 considers only direct transitions rather than paths among basins of attraction, and is thus a coarser measure compared to the maximum path potential.

Ellison (2000) also derived conditions that are closely related to those in Theorem 1 (i) and Corollary 1, but for computing the long-run stochastically stable set rather than the general case of $\boldsymbol{\varepsilon}$-stable sets. Ellison (2000) uses the radii and co-radii of limit sets as a measure of their persistence and attractiveness respectively, which are based on counting the number of mutation required to exit the basins of attraction. As discussed in the introduction and section 2, such measure are specific to the model of mistakes in which the mutation rates are state-independent and identical for all agents. The resistance, coresistance and path potentials above are independent of the model of mistakes. Theorem 1 (i) together with its constructive proof also provides a theoretical foundation to the notion of step-by-step evolution in Ellison (2000). In particular, the proof of Theorem 1 (i) which is based on the quotients of stationary distributions of basins of attraction demonstrates that the modified coradius argument of Ellison (2000) is basically a consequence of the chain-rule applied to quotients.

Since the condition in Corollary 1 is coarser and hence more restrictive than that of Theorem 1 (i), it is therefore computationally economical to first consider the resistances and co-resistances of direct transitions to identify the union set containing the $\boldsymbol{\varepsilon}$-stable set. If this procedure leads to a null set, only then should it be necessary to compute the maximum path potentials of basins of attraction. A conjecture we state here but whose proof is beyond the scope of this paper, is that at least one among conditions in Corollary 1 and Theorem 1 (i) must result into a non-empty union set provided that $\varepsilon_i(\mathbf{x}) > 0$ for all i and $\mathbf{x}$.

Theorem 1 (ii) shows that the measures of resistance we construct can also be used in the usual spanning tree algorithm employed in Kandori *et al.* (1993) and Young (1993). But as in Ellison (2000), the measure of costs of transitions among metastable sets employed in Kandori *et al.* (1993) and Young (1993) are specific to the model of mistakes in which the mutation rates are state-independent and identical for all agents. We also demonstrate below that such measures require an additional restriction that the mistakes distribution must be bounded. The condition in Theorem 1 (ii) is tighter than those in Corollary 1 and Theorem 1 (i). It selects more than one $\boldsymbol{\varepsilon}$-stable set only if there exist two or more metastable sets with identical minimum stochastic potential, otherwise it selects a unique $\boldsymbol{\varepsilon}$-stable set. A computationally economical algorithm for identifying the long-run $\boldsymbol{\varepsilon}$-stable set would then entail the following three steps.

The first step is to identify the limits sets of the process $(\mathbf{X}, P)$ and the associated basins of attraction. This can usually be done heuristically by combining the properties of the payoff function and the topology of the interaction structure. The limit sets of $(\mathbf{X}, P)$ are also the metastable sets of $(\mathbf{X}, P_{\boldsymbol{\varepsilon}})$ with the respective basins of attraction. The second step is to compute the resistances and co-resistances of direct transitions associated with each basins of attraction. Identify the metastable sets whose basins of attraction have a resistance greater than their co-resistance. If this procedure identifies one metastable set, then it is the unique $\boldsymbol{\varepsilon}$-stable set of the process $(\mathbf{X}, P_{\boldsymbol{\varepsilon}})$. If the procedure leads to a null set, then compute the maximum path potentials for each metastable set and identify those for which the potential is negative. A set that is a union of all metastable sets whose maximum path potential is negative, contains the long-run $\boldsymbol{\varepsilon}$-stable set. If this procedure identifies one metastable stable set, then it is the unique $\boldsymbol{\varepsilon}$-stable set of $(\mathbf{X}, P_{\boldsymbol{\varepsilon}})$. If either of the procedures selects more than one metastable set then proceed to the third step, which employs the spanning tree algorithm to select a unique $\boldsymbol{\varepsilon}$-stable set if it exists. More specifically, compute the stochastic potentials of all sets that survived the second step and select those with the minimum stochastic potential.

The spanning tree algorithm is therefore only necessary if the first two steps fail to select a unique $\boldsymbol{\varepsilon}$-stable set. Moreover, the first two steps reduce the number of metastable sets for which one has to construct their spanning trees. This reduces the computational burden of constructing the spanning trees of all metastable sets and of the associated stochastic potentials.

The following proposition establishes the conditions under which stochastic stability can be used to approximate $\boldsymbol{\varepsilon}$-stability.

Proposition 2: *Let* $(\mathbf{X}, P_{\boldsymbol{\varepsilon}})$ *be a model of learning with state-independent homogeneous mutation rates* $\boldsymbol{\varepsilon} = (\varepsilon, \cdots, \varepsilon)$*, and action space* X*. Let also* $0 < \mathcal{P}_i(x|\mathbf{x}) < 1 \ \forall \ i \in N$*,* $\forall \ x \in X$ *and* $\forall \ \mathbf{x} \in \mathbf{X}$ *be the bounded mistakes probability mass function.*

(i) Then there exists an $\varepsilon' > 0$ *such that for every* $\varepsilon < \varepsilon'$*,* $\pi_{\boldsymbol{\varepsilon}} \sim \lim_{\varepsilon \to 0} \pi_{\boldsymbol{\varepsilon}}$*. Where "*$\sim$*" stands for "can be approximated by".*

(ii) If $\mathcal{P}_i(x|\mathbf{x}) = \frac{1}{\#X} \ \forall \ i \in N$*,* $\forall \ x \in X$ *and* $\forall \ \mathbf{x} \in \mathbf{X}$*, then* $\pi_{\boldsymbol{\varepsilon}} \sim \lim_{\varepsilon \to 0} \pi_{\boldsymbol{\varepsilon}}$ *for all values of* ε*.*

Proof. See Appendix A.4 □

Proposition 2 provides conditions under which $\boldsymbol{\varepsilon}$-stability can be approximated by stochastic stability. First and most importantly, it is necessary that the mutation rates be state-independent and homogeneous to all agents. If the mistakes are uniformly and randomly distributed, then stochastic stability can be used to approximate $\boldsymbol{\varepsilon}$-stability for any values of $\varepsilon \in (0, 1)$. The second necessary condition in the case of arbitrary mistakes distributions is that the mistakes distribution must be bounded, but even so, the approximation is strictly valid only for a range of values of ε.

The straightforward generalization of the state-independent and homogeneous mutation rates condition above is the situations in which the stationary distribution converges uniformly in the limit of mutation rates, such as van Damme and Weibull (2002). A less general case is the models whereby the mutation rates are a function of a single parameter, such as Maruta (2002) and Blume (2003). Examples include the logit dynamics in (6) in which $\varepsilon = 1$ and β is the single parameter that controls the mutation rates.

4.2. Applications

We now provide applications of Theorem 1 to the learning process whose dynamics is governed by (6). These applications will also demonstrate the robustness of the measures we have constructed above and the generality of Theorem 1. The game in the first example is also used by Young (1993) and Ellison (2000) to show that the risk-dominant equilibrium need not be selected in 3×3 games, and we use it here to demonstrate the relationship between $\boldsymbol{\varepsilon}$-stability and stochastic stability.

Example 1: *Consider a normal form game Γ in Table 3, played by a sufficiently large number n of agents that are uniformly and randomly matched over time. Let the learning model be that prescribed by (6), and let $\vec{x}$ denote the state in which all players play x.*

(i) Let the mutation rates $\boldsymbol{\varepsilon} = (\varepsilon, \cdots, \varepsilon)$ be state-independent and homogeneous.

(a) If the mistakes are uniformly and randomly distributed (that is $\beta = 0$) then $\pi_{\boldsymbol{\varepsilon}}(\vec{B}) \sim \lim_{\varepsilon \to 0} \pi_{\boldsymbol{\varepsilon}}(\vec{B}) = 1$ for all values of $\varepsilon \in (0, 1)$.

(b) If $\beta = 1$, then $\pi_{\boldsymbol{\varepsilon}}(\vec{B}) \sim \lim_{\varepsilon \to 0} \pi_{\boldsymbol{\varepsilon}}(\vec{B}) = 1$ if and only if $\varepsilon \lesssim 10^{-7}$, otherwise $\vec{C}$ is the $\boldsymbol{\varepsilon}$-stable set.

(c) If $\beta = 10$, then $\pi_{\boldsymbol{\varepsilon}}(\vec{B}) \sim \lim_{\varepsilon \to 0} \pi_{\boldsymbol{\varepsilon}}(\vec{B}) = 1$ if and only if $\varepsilon \lesssim 10^{-57}$, otherwise $\vec{C}$ is the $\boldsymbol{\varepsilon}$-stable set.

(ii) Let the mutation rates be state-dependent and determined by the parameter β. Consider the case of pure logit dynamics, that is set $\varepsilon = 1$. Then $\pi_{\boldsymbol{\varepsilon}}(\vec{C}) \sim \lim_{\beta \to \infty} \pi_{\boldsymbol{\varepsilon}}(\vec{C}) = 1$

Table 3: For any pair of agent $i, j \in N$ the profile (C, C) is risk dominant.

	A	B	C
A	6 , 6	0 , 5	0 , 0
B	5 , 0	7 , 7	5 , 5
C	0 , 0	5 , 5	8 , 8

Proof. The first step is to identify the metastable sets of the process $(\mathbf{X}, P_{\boldsymbol{\varepsilon}})$ which are exactly the same as the limit sets of $(\mathbf{X}, P)$. Under uniform-random matching, the metastable sets for the game in Table 3 are the singleton sets $\vec{A}$, $\vec{B}$ and $\vec{C}$. We then compute the diameters and average probabilities of direct transitions between pairs of basins of attractions of $\vec{A}$, $\vec{B}$ and $\vec{C}$ denoted by $\tilde{A}$, $\tilde{B}$ and $\tilde{C}$ respectively. These can both be computed from the payoff structure. For example the diameter $d(\tilde{A}, \tilde{B})$ of the directed relation $\tilde{A} \to \tilde{B}$ is given by[5]

[5] The expression on the right hand side of the inequality (15) follows from the fact that under best response

$$d(\tilde{A},\tilde{B}) > \max\left\{\frac{v(A,A)-v(B,A)}{(v(A,A)-v(B,A))+(v(B,B)-v(A,B))},\right.$$
$$\left.\frac{v(A,A)-v(B,A)}{(v(C,A)-v(B,A))+(v(B,B)-v(C,B))}\right\} \tag{15}$$

where $v(x,y)$ is the payoff derived from playing x when the opponent plays y. By substituting the payoffs from Table 3 it follows that $d(\tilde{A},\tilde{B}) = 1/8$. A similar argument applies to all directed relations between all pairs of basins of attraction.

In the case of the average probability, recall that for any pair $\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j \in \tilde{\mathbf{X}}$

$$\mathbb{P}_A(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j) = \frac{1}{n}\sum_{i=1}^{n}\mathscr{P}_i(x_{t+1}=y|\mathbf{x}_t=\mathbf{x}),$$

By making use of the definition of $\mathscr{P}_i(x_{t+1}=y|\mathbf{x}_t=\mathbf{x})$ in (6) it follows for the transition $\tilde{A}\to\tilde{B}$ that

$$\mathscr{P}_i(x_{t+1}=B|\mathbf{x}_t=\vec{A}) = \varepsilon\frac{1}{1+e^{-5\beta}+e^{\beta}} \quad \text{for all } i\in N \tag{16}$$

where the powers of the exponents on the right hand side of (16) follow from the fact that under uniform-random matching $u_i(B,\vec{A}) = \frac{1}{n}\sum_{j=1}^{n} v(B,A) = v(B,A)$ for all $i\in N$. Consequently,

$$\mathbb{P}_A(\tilde{A},\tilde{B}) = \varepsilon\frac{1}{1+e^{-5\beta}+e^{\beta}}$$

We write $f_\beta(a,b)$ as the shorthand for $1/(1+e^{a\beta}+e^{b\beta})$. The average probabilities for the transitions between each pair of basins of attraction can then be computed in a similar manner, and are generally of the form,

$$\mathbb{P}_A(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j) = \varepsilon\frac{1}{1+e^{a\beta}+e^{b\beta}} \tag{17}$$

The cost of each direct transitions among basins of attraction are as follows: $c(\tilde{A},\tilde{B}) = -\frac{1}{8}\ln(\varepsilon f_\beta(-5,1))$, $c(\tilde{B},\tilde{A}) = -\frac{7}{8}\ln(\varepsilon f_\beta(5,7))$, $c(\tilde{B},\tilde{C}) = -\frac{2}{5}\ln(\varepsilon f_\beta(-5,2))$, $c(\tilde{C},\tilde{B}) = -\frac{3}{8}\ln(\varepsilon f_\beta(-5,3))$, $c(\tilde{A},\tilde{C}) = -\frac{5}{8}\ln(\varepsilon f_\beta(5,6))$ and $c(\tilde{C},\tilde{A}) = -\frac{5}{6}\ln(\varepsilon f_\beta(5,8))$.

(i) (a) To recover the situation in which the mistakes are uniformly and randomly distributed, we simply substitute for $\beta = 0$. The substitution results into $\mathscr{R}(\tilde{B}) = -\frac{2}{5}\ln(\frac{\varepsilon}{3}) > -\frac{3}{8}\ln(\frac{\varepsilon}{3}) = \mathscr{C}\mathscr{R}(\tilde{B})$, $\mathscr{R}(\tilde{A}) < \mathscr{C}\mathscr{R}(\tilde{A})$ and $\mathscr{R}(\tilde{C}) < \mathscr{C}\mathscr{R}(\tilde{C})$ for all values of $\varepsilon\in(0,1)$. Implying that $\pi_\varepsilon(\vec{B}) \sim \lim_{\varepsilon\to 0}\pi_\varepsilon(\vec{B}) = 1$.

dynamics $d(\tilde{A},\tilde{B})$ must satisfy the following inequalities

$$d(\tilde{A},\tilde{B})v(B,B) + (1-d(\tilde{A},\tilde{B}))v(B,A) > d(\tilde{A},\tilde{B})v(A,B) + (1-d(\tilde{A},\tilde{B}))v(A,A)$$

$$d(\tilde{A},\tilde{B})v(B,B) + (1-d(\tilde{A},\tilde{B}))v(B,A) > d(\tilde{A},\tilde{B})v(C,B) + (1-d(\tilde{A},\tilde{B}))v(C,A)$$

(b) When $\beta = 1$, it follows directly from the values of costs of direct transitions among basins of attraction above that $\mathscr{R}(\tilde{B}) > \mathscr{CR}(\tilde{B})$, $\mathscr{R}(\tilde{A}) < \mathscr{CR}(\tilde{A})$ and $\mathscr{R}(\tilde{C}) < \mathscr{CR}(\tilde{C})$ if and only if $\varepsilon \lesssim 10^{-7}$. In which case $\pi_{\boldsymbol{\varepsilon}}(\vec{B}) \sim \lim_{\varepsilon\to 0} \pi_{\boldsymbol{\varepsilon}}(\vec{B}) = 1$. Otherwise, for all values of $\varepsilon > 10^{-7}$ we have that $\mathscr{R}(\tilde{B}) < \mathscr{CR}(\tilde{B})$, $\mathscr{R}(\tilde{A}) < \mathscr{CR}(\tilde{A})$ and $\mathscr{R}(\tilde{C}) < \mathscr{CR}(\tilde{C})$. We thus need to proceed to the second step of the computational algorithm since the first step leads to a null set.

The second step of the algorithm involves computing the maximum path potentials for each basin of attraction. By substituting the costs of direct transitions among basins of attraction into (13), it follows that (though a tedious exercise) for all $\varepsilon > 10^{-7}$, $\phi(\tilde{A}) > 0$, $\phi(\tilde{B}) > 0$ and

$$\phi(\tilde{C}) = -\left(\frac{2}{5}\ln(\varepsilon f_\beta(-5,2)) - \frac{3}{8}\ln(\varepsilon f_\beta(-5,3))\right) \tag{18}$$

$$= -\left(0.3 + \frac{1}{40}\ln(\varepsilon)\right) < 0 \tag{19}$$

where the second equality follows from substituting for $\beta = 1$. We thus have $\pi_{\boldsymbol{\varepsilon}}(\vec{B}) \sim \lim_{\varepsilon\to 0} \pi_{\boldsymbol{\varepsilon}}(\vec{B}) = 1$ for all values of $\varepsilon \lesssim 10^{-7}$, and $\vec{C} = \arg\max_{\Omega_{\boldsymbol{\varepsilon}} \in \boldsymbol{\Omega}_{\boldsymbol{\varepsilon}}} \pi_{\boldsymbol{\varepsilon}}(\Omega_{\boldsymbol{\varepsilon}})$ for all $\varepsilon > 10^{-7}$.

(c) Similarly, when $\beta = 10$, we have that $\mathscr{R}(\tilde{B}) > \mathscr{CR}(\tilde{B})$, $\mathscr{R}(\tilde{A}) < \mathscr{CR}(\tilde{A})$ and $\mathscr{R}(\tilde{C}) < \mathscr{CR}(\tilde{C})$ if and only if $\varepsilon \lesssim 10^{-57}$. And that for all $\varepsilon > 10^{-57}$, $\phi(\tilde{A}) > 0$, $\phi(\tilde{B}) > 0$ and

$$\phi(\tilde{C}) = -\left(3.25 + \frac{1}{40}\ln(\varepsilon)\right) < 0 \tag{20}$$

Implying that $\pi_{\boldsymbol{\varepsilon}}(\vec{B}) \sim \lim_{\varepsilon\to 0} \pi_{\boldsymbol{\varepsilon}}(\vec{B}) = 1$ for all values of $\varepsilon \lesssim 10^{-57}$, and $\vec{C} = \arg\max_{\Omega_{\boldsymbol{\varepsilon}} \in \boldsymbol{\Omega}_{\boldsymbol{\varepsilon}}} \pi_{\boldsymbol{\varepsilon}}(\Omega_{\boldsymbol{\varepsilon}})$ for all $\varepsilon > 10^{-57}$.

(ii) When $\varepsilon = 1$ in (6), we have for all values of β that $\phi(\tilde{A}) > 0$, $\phi(\tilde{B}) > 0$ and

$$\phi(\tilde{C}) = -\left(\frac{2}{5}\ln(f_\beta(-5,2)) - \frac{3}{8}\ln(f_\beta(-5,3))\right) < 0 \tag{21}$$

Note that in the computation of the maximum path potential, direct transitions are also considered. The conditions in Corollary 1 and Theorem 1 (i) thus sufficiently identify the unique long-run $\boldsymbol{\varepsilon}$-stable set, and the computation of the minimum stochastic potential is not necessary. Nevertheless, we demonstrate how the third step of the computational algorithm can be applied when it necessitates.

The graphs $\mathscr{W}$ for the cases in which the mistakes are uniformly and randomly distributed (CASE1), and that in which $\varepsilon = 1$ (CASE2) are shown in Figure 1. The weights $w(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$ on the directed arrows correspond to the minimum cost over all paths (including direct transitions) between $\tilde{\mathbf{x}}_i \to \tilde{\mathbf{x}}_j$. For example the path of minimum cost from $\tilde{\mathbf{A}}$ to $\tilde{\mathbf{C}}$ is $\tilde{\mathbf{A}} \to \tilde{\mathbf{B}} \to \tilde{\mathbf{C}}$ with the costs of

$$w(\tilde{A}, \tilde{C}) = -\left(\frac{1}{8}\ln\left(\frac{\varepsilon}{3}\right) + \frac{2}{5}\ln\left(\frac{\varepsilon}{3}\right)\right)$$

$$w(\tilde{A}, \tilde{C}) = -\left(\frac{1}{8}\ln(f_\beta(-5,1)) + \frac{2}{5}\ln(f_\beta(-5,2))\right)$$

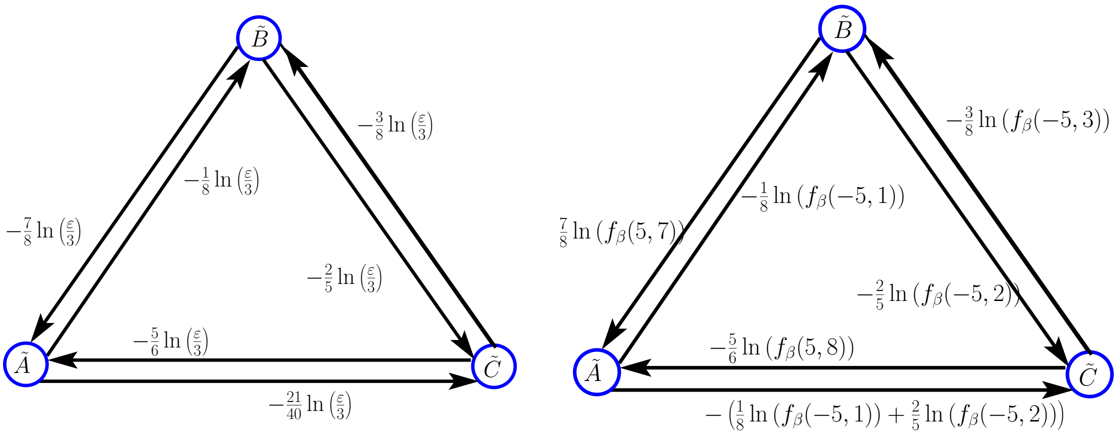
$\tilde{B}$
$-\frac{3}{8}\ln\left(\frac{\varepsilon}{3}\right)$
$-\frac{1}{8}\ln\left(\frac{\varepsilon}{3}\right)$
$-\frac{7}{8}\ln\left(\frac{\varepsilon}{3}\right)$
$-\frac{2}{5}\ln\left(\frac{\varepsilon}{3}\right)$
$-\frac{5}{6}\ln\left(\frac{\varepsilon}{3}\right)$
$\tilde{A}$
$\tilde{C}$
$-\frac{21}{40}\ln\left(\frac{\varepsilon}{3}\right)$
$\tilde{B}$
$-\frac{3}{8}\ln\left(f_\beta(-5,3)\right)$
$-\frac{1}{8}\ln\left(f_\beta(-5,1)\right)$
$\frac{7}{8}\ln\left(f_\beta(5,7)\right)$
$-\frac{2}{5}\ln\left(f_\beta(-5,2)\right)$
$-\frac{5}{6}\ln\left(f_\beta(5,8)\right)$
$\tilde{A}$
$\tilde{C}$
$-\left(\frac{1}{8}\ln\left(f_\beta(-5,1)\right)+\frac{2}{5}\ln\left(f_\beta(-5,2)\right)\right)$

Figure 1: The left hand side figure is for CASE1 and that on the right hand side is for CASE2.

under CASE1 and CASE2 respectively.

From Figure 1 it follows that the minimum stochastic potential under CASE1 and CASE2 are respectively

$$\mathscr{S}(\tilde{B}) = -\left(\frac{1}{8}\ln\left(\frac{\varepsilon}{3}\right) + \frac{3}{8}\ln\left(\frac{\varepsilon}{3}\right)\right) = \frac{1}{2}\ln\left(\frac{\varepsilon}{3}\right)$$
$$\mathscr{S}(\tilde{C}) = -\left(\frac{1}{8}\ln(f_\beta(-5,1)) + \frac{2}{5}\ln(f_\beta(-5,2))\right)$$

□

Example 1 clearly shows that not only does the long-run stochastically stable depend on whether or not the mutation rate is state-dependent but also on the structure of the mistakes distribution. This calls for extra caution when approximating $\boldsymbol{\varepsilon}$-stability by stochastic stability. The result in Example 1 (ii), that the risk-dominat equilibrium is selected ($\boldsymbol{\varepsilon}$-stable) under logit dynamics, is consistent with others in the literature such as Blume (1995).

The next example presents a game in which there does not exists a unique risk-dominant strategy such that $\boldsymbol{\varepsilon}$-stability can be approximated by stochastic stability under both state-dependent and state-independent mutation rates. It also further demonstrates the situation in which the condition in Corollary 1 leads to a null set but the maximum path potential argument identifies a unique $\boldsymbol{\varepsilon}$-stable set.

Example 2: *Consider a normal form game* Γ *in Table 4, played by* n *agents that are uniformly and randomly matched over time. Let the learning model be that prescribed by* (6)*. Then for* n *sufficiently large,* $\pi_\varepsilon(\vec{B}) \sim \lim_{\varepsilon\to 0}\pi_\varepsilon(\vec{B}) = \lim_{\beta\to\infty}\pi_\varepsilon(\vec{B}) = 1$

Proof. By following the same steps as in the proof of Example 1, the costs of transitions among basins of attraction are as follows: $c(\tilde{A},\tilde{B}) = -\frac{4}{7}\ln(\varepsilon f_\beta(2,4))$, $c(\tilde{B},\tilde{A}) = -\frac{3}{7}\ln(\varepsilon f_\beta(-1,3))$, $c(\tilde{B},\tilde{C}) = -\frac{2}{3}\ln(\varepsilon f_\beta(1,4))$, $c(\tilde{C},\tilde{B}) = -\frac{1}{3}\ln(\varepsilon f_\beta(-2,2))$, $c(\tilde{A},\tilde{C}) = -\frac{1}{3}\ln(\varepsilon f_\beta(-2,2))$ and $c(\tilde{C},\tilde{A}) = -\frac{2}{3}\ln(\varepsilon f_\beta(2,4))$.

Table 4: For any pair of agent $i, j \in N$ the profile (C, C) is risk dominant.

	A	B	C
A	5 , 5	3 , 1	0 , 3
B	1 , 3	6 , 6	2 , 2
C	3 , 0	2 , 2	4 , 4

The condition in Corollary 1 leads to a null set for all values of $\varepsilon \in (0,1)$ and $0 \leq \beta < \infty$. We thus proceed to step two of the algorithm, which computes the maximum path potentials. The maximum path potentials are such that for all values of $\varepsilon \in (0,1)$ and $0 \leq \beta < \infty$, $\phi(\tilde{A}) > 0$, $\phi(\tilde{C}) > 0$ and

$$\phi(\tilde{B}) = -\left(\frac{3}{7}\ln(\varepsilon f_\beta(-1,3)) - \frac{1}{3}\ln(\varepsilon f_\beta(-2,2))\right) < 0.$$

Implying that $\pi_\varepsilon(\vec{B}) \sim \lim_{\varepsilon \to 0} \pi_\varepsilon(\vec{B}) = \lim_{\beta \to \infty} \pi_\varepsilon(\vec{B}) = 1$. □

The next example will illustrate how local interactions can affect the resistance of basins of attraction. The network topology can affect both the diameter of any pair of basins of attraction and the associated average probabilities, and hence the speed of evolution between them. In the case of complex and random network structures the normalized diameter between pairs of basins of attraction can be approximated by taking the average over the fraction of neighbors each agent requires to simultaneously switch strategies for that agent to do the same. For example in the game of Table 1, if the process $(\mathbf{X}, P)$ is in state $\vec{B}$ then each $i \in N$ requires at least $\frac{1}{3}k_i$ neighbors to simultaneous switch to play A before i does so in the period after. The diameter $d(\tilde{B}, \tilde{A})$ can then be approximated by

$$d(\tilde{B}, \tilde{A}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{3}\frac{k_i}{n} = \frac{1}{3}\frac{1}{n}\sum_{k=1}^{\infty} kp(k) = \frac{1}{3}\frac{\langle k \rangle}{n} \tag{22}$$

where $\{p(k)\}_{k \geq 1}$ is the degree distribution of the network and $\langle k \rangle$ is its average degree. If the network assumes a simple topology, then it is possible to compute $d(\tilde{\mathbf{x}}, \tilde{\mathbf{x}})$ by first identifying the role each agent plays with respect to their degree.

Example 3: *Consider a normal form game Γ in Table 1 and let the learning model be that governed by the dynamics in (6). Denote by $\mathscr{R}_u(\tilde{\mathbf{x}})$ and $\phi_u(\tilde{\mathbf{x}})$ for the resistance and maximum path potential of $\tilde{\mathbf{x}}$ respectively under uniform and random matching, and by $\mathscr{R}_l(\tilde{\mathbf{x}})$ and $\phi_l(\tilde{\mathbf{x}})$ under uniform and local random matching with network topology in Figure 2. Then under both uniform-random and uniform-local-random matching, $\pi_\varepsilon(\vec{A}) \sim \lim_{\varepsilon \to 0} \pi_\varepsilon(\vec{A}) = \lim_{\beta \to \infty} \pi_\varepsilon(\vec{A}) = 1$. We also have that $\mathscr{R}_u(\tilde{B}) > \mathscr{R}_l(\tilde{B})$, $\mathscr{R}_u(\tilde{A}) > \mathscr{R}_l(\tilde{A})$, $\phi_u(\vec{B}) > \phi_l(\vec{B})$ and $\phi_u(\vec{A}) < \phi_l(\vec{A})$.*

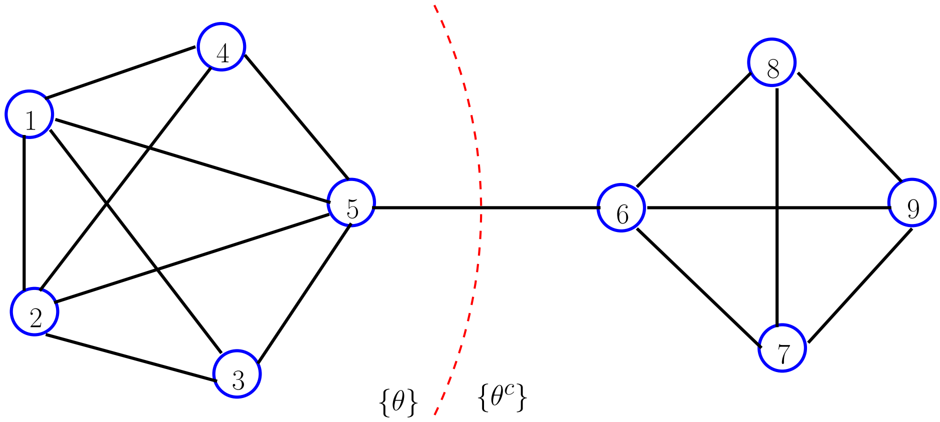

Figure 2: The red dashed line divides the network into subgroups $\theta = \{1, 2, 3, 4, 5\}$ and $\theta^c = \{6, 7, 8, 9\}$.

Proof. Under uniform and random matching, when the process $(\mathbf{X}, P_\varepsilon)$ is in $\vec{B}$, at least $\frac{1}{3}$ of the population must play B for the transition $\vec{B} \to \vec{A}$ to occur. Implying that $d(\tilde{B}, \tilde{A}) = \frac{1}{3}$. After computing the associated average probability $\mathbb{P}_A(\tilde{B}, \tilde{A})$, the cost of the transition $\vec{B} \to \vec{A}$ is $c(\tilde{B}, \tilde{A}) = -\frac{1}{3}\ln(\varepsilon f_\beta(2))$, where $f_\beta(a) = 1/(1 + e^{a\beta})$. Similarly, $c(\tilde{B}, \tilde{A}) = -\frac{2}{3}\ln(\varepsilon f_\beta(4))$. Implying that $\mathscr{R}_u(B) < \mathscr{C}\mathscr{R}_u(B)$. The corresponding maximum path potentials are for all values of $\varepsilon \in (0, 1)$ and $0 \leq \beta < \infty$:

$$\phi_u(\tilde{A}) = -\left(\frac{1}{3}\ln(\varepsilon f_\beta(2)) - \frac{2}{3}\ln(\varepsilon f_\beta(4))\right) < 0 \tag{23}$$

$$\phi_u(\tilde{B}) = -\left(\frac{2}{3}\ln(\varepsilon f_\beta(4)) - \frac{1}{3}\ln(\varepsilon f_\beta(2))\right) > 0 \tag{24}$$

Now consider the case of uniform local random matching for the network in Figure 2. Under the mutationless model $(\mathbf{X}, P)$ there are two additional limit sets. Let $\theta = \{1, 2, 3, 4, 5\}$ be the subset of agents belonging to the subgroup on the left hand side of the network in Figure 2. Then there exist two limit sets of $(\mathbf{X}, P)$ induced by the network topology; $\vec{B}_\theta$ in which all players in subgroup θ play B and those in θ^c play A, and $\vec{A}_\theta$ in which all players in subgroup θ play A and those in θ^c play B. By considering intermediate metastable sets $\vec{B}_\theta$ and $\vec{A}_\theta$ together with their basins of attraction $\tilde{B}_\theta$ and $\tilde{A}_\theta$, the resistance of $\tilde{B}$ is associated with the transition $\tilde{B} \to \tilde{B}_\theta$ and is given by $\mathscr{R}_l(\tilde{B}) = \frac{1}{9}\ln(\varepsilon f_\beta(2))$. Similarly for $\tilde{A}$, we have that $\mathscr{R}_l(\tilde{A}) = \frac{2}{9}\ln(\varepsilon f_\beta(4))$, which is associated with the transition $\tilde{A} \to \tilde{A}_\theta$. The corresponding co-resistances are: $\mathscr{C}\mathscr{R}_l(\tilde{B}) = \frac{4}{9}\ln(\varepsilon f_\beta(4))$ and $\mathscr{C}\mathscr{R}_l(\tilde{A}) = \frac{1}{3}\ln(\varepsilon f_\beta(2))$.

It can also be easily shown (though a bit tedious exercise) that the maximum path potentials for $\vec{B}$ and $\vec{A}$ are given by

$$\phi_u(\tilde{A}) = -\left(\frac{2}{9}\ln(\varepsilon f_\beta(2)) - \frac{2}{9}\ln(\varepsilon f_\beta(4))\right) < 0 \tag{25}$$

$$\phi_u(\tilde{B}) = -\left(\frac{2}{9}\ln(\varepsilon f_\beta(4)) - \frac{1}{9}\ln(\varepsilon f_\beta(2))\right) > 0 \tag{26}$$

Clearly, $\vec{A}$ is both the long-run $\boldsymbol{\varepsilon}$-stable and stochastically stable set, and that $\mathscr{R}_u(\tilde{B}) > \mathscr{R}_l(\tilde{B})$, $\mathscr{R}_u(\tilde{A}) > \mathscr{R}_l(\tilde{A})$, $\phi_u(\vec{B}) > \phi_l(\vec{B})$ and $\phi_u(\vec{A}) < \phi_l(\vec{A})$ □

The presence of metastable sets induced by the network act to reduce the resistance of the basins of attraction of other metastable sets and hence speeding up the evolution of the process $(\mathbf{X}, P_\varepsilon)$ between every pair of basins of attraction. This is precisely the effect of step-by-step evolution discussed by Ellison (2000).

5. Local stability and the speed of learning

In this section, we focus on characterizing the convergence rates of the process $(\mathbf{X}, P_\varepsilon)$, and in particular we construct three measures of convergence that can be used to characterize the short-run and medium-run behavior of the system. That is, the *expected waiting time* between metastable sets, *metastability* as a measure of how locally stable a metastable set is, and the *contagion rate* as the measure of the convergence rate of the process to its quasi-stationary distribution within a given basin of attraction. The contagion rate is equivalently the measure of how fast a strategy diffuses across the population once its threshold has been attained.

The rate at which the process $(\mathbf{X}, P_\varepsilon)$ convergence to its long-run stationary distribution is well studied in the literature under the concept of *mixing time*. The lower and upper bounds for the mixing time of discrete Markov chains are well established in the literature, generally as functions of the second largest eigenvalue and of the diameter of the associated transition matrix.[6] For this reason, we focus on measures of convergence that capture the short-run and medium-run behavior of the process $(\mathbf{X}, P_\varepsilon)$. Moreover, as demonstrated in section 2, the mixing time of the process can be unrealistically long when the probabilities of mistakes are small.

5.1. Expected waiting time and metastability

The expected waiting time between any pair of metastable sets is basically the expected time it takes the process $(\mathbf{X}, P_\varepsilon)$ to enter the boundary of the basin of attraction of the second metastable set given that it starts from the state belonging to the first metastable set. Metastability is a measure of how long it takes the process $(\mathbf{X}, P_\varepsilon)$ to exit the basin of attraction of a metastable set once it has entered its boundaries. Equivalently, it is the measure of the persistence of the limit sets of $(\mathbf{X}, P)$ to perturbations and can thus be used to rank limit sets in terms of how locally stable they are. Formally,

Definition 3: *Let Ω^i_ε and Ω^j_ε be two metastable sets of the process $(\mathbf{X}, P_\varepsilon)$ with respective basins of attraction $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_j$. Let $n(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$ be the number of agents that must play a different*

[6] The interested reader can refer to Aldous and Fill (1999) for detailed expositions concerning mixing times of Markov chains.

action for the transition from Ω_ε^i to the boundary of $\tilde{\mathbf{x}}_j$ to occur. Then the expected waiting time $\mathbb{E}T(\Omega_\varepsilon^i, \tilde{\mathbf{x}}_j)$ associated with such a transition and the metastability $\mathscr{M}(\Omega_\varepsilon^i)$ of Ω_ε^i are defined as follows.

$$\mathbb{E}T(\Omega_\varepsilon^i, \tilde{\mathbf{x}}_j) = \mathbb{E}\left[\min\left\{t \middle| \, n(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \geq d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n\right\}\right]$$

$$\mathscr{M}(\Omega_\varepsilon^i) = \min_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} \mathbb{E}T(\Omega_\varepsilon^i, \tilde{\mathbf{x}}_j)$$

The following theorem provides lower bounds on both the expected waiting time and metastability of metastable sets.

THEOREM 2: *Let $(\mathbf{X}, P_\varepsilon)$ be a model of learning with mistakes, and let Ω_ε^i and Ω_ε^j be any two of its metastable sets with corresponding basins of attraction $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_j$. Then*

$$\begin{aligned}\mathbb{E}T(\Omega_\varepsilon^i, \tilde{\mathbf{x}}_j) &\geq e^{n[c(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) + f(d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j))]} \\ \mathscr{M}(\Omega_\varepsilon^i) &\geq e^{n[f(d(\tilde{\mathbf{x}}_i)) + \mathscr{R}(\tilde{\mathbf{x}}_i)]}.\end{aligned}$$

where $f(\alpha) = \alpha \ln\left(\frac{\alpha}{1-\alpha}\right) + \ln(1-\alpha)$, and $d(\tilde{\mathbf{x}}_i) = \arg\min_{d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)} c(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$

Proof. See Appendix A.5 □

The metastability of any metastable set is bounded from below by the exponential function of the resistance of its basin of attraction and the diameter $d(\tilde{\mathbf{x}}_i)$ that minimizes the cost of exiting its basin of attraction. The direct implication is that, depending on the model of mistakes, the metastable set with the largest size of basin of attraction is not necessarily that with the highest level of metastability. By the size of basin of attraction we mean the minimum number (or fraction) of mistakes required to exit a given basin of attraction. That is for any $\tilde{\mathbf{x}}_i \in \tilde{\mathbf{X}}$, the size of basin of attraction of $\tilde{\mathbf{x}}_i$ is given by $\min_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n$. In relation to the discussion of the step-by-step evolution argument above, it then follows that the presence of intermediate metastable sets induced by the network structure act to reduce the expected waiting time and metastability of other metastable sets.

The definitions of expected waiting time and metastability above are closely related to the notion of hitting time in discrete Markov chains. Theorem 2 together with its constructive proof thus provide a novel and tighter lower bounds to hitting time in general. Ellison (1993, 2000) also studies the concept of waiting time and provides an upper bound while assuming vanishing mutation rates, which is rather specific to the case of state-independent mutation rates. On the contrary, Theorem 2 provides tighter lower bounds that are independent of the model of mistakes and the assumption of vanishing mutation rates is not necessary in proof of the result.

The expression we provide for metastability can also be independently employed as a measure for equilibrium selection in some specific settings of stochastic evolutionary dynamics. For example, in some evolutionary models such as those pioneered by Binmore and Samuelson (1997), the interest is in the properties of the learning process as the population size becomes arbitrarily large. The expressions for the expected waiting time and metastability in Theorem 2 imply that the equilibria that are stable in a long-run are those for which the diameter of the basins of attraction are independent of the population size. To see this, let Ω^*_ε be a metastable set that satisfies such conditions, and $\tilde{\mathbf{x}}^*$ be its respective basin of attraction. Then

$$\mathscr{M}(\Omega^*_\varepsilon) \propto e^{n\ln(\mathbb{P}_A(\tilde{\mathbf{x}}^*,\tilde{\mathbf{x}}_{min}))}$$

where $\tilde{\mathbf{x}}_{min}$ is the basin of attraction that minimizes the cost of exiting the boundaries of $\tilde{\mathbf{x}}^*$. It then follows that $\lim_{n\to\infty}\mathscr{M}(\Omega^*_\varepsilon) = \infty$.

5.2. Contagion rate

The expected waiting time and metastability measures above are concerned with inter-metastable sets transitions, and hence the medium-run behavior of the process $(\mathbf{X}, P_\varepsilon)$. In this subsection, we characterize the short-run behavior of $(\mathbf{X}, P_\varepsilon)$, which is the dynamics within the basins of attraction. Once the process enters the boundaries of the basin of attraction of a given metastable set, it acquires a quasi-stationary distribution over the state space of the given basin of attraction. The quasi-stationary distribution attained places most weight on the corresponding metastable set. We define the contagion rate within the basin of attraction as the rate at which $(\mathbf{X}, P_\varepsilon)$ converges to the quasi-stationary distribution it attains within the given basin of attraction. Formally,

Definition 4: *Let Ω_ε be any metastable set of $(\mathbf{X}, P_\varepsilon)$ with the respective basin of attraction $\tilde{\mathbf{x}}$. Let $\nu_{\tilde{\mathbf{x}}}$ be the quasi-stationary distribution $(\mathbf{X}, P_\varepsilon)$ attains within $\tilde{\mathbf{x}}$. Then the contagion rate $\gamma_{\tilde{\mathbf{x}}}$ within $\tilde{\mathbf{x}}$ is defined as*

$$\gamma_{\tilde{\mathbf{x}}} = \limsup_{t\to T_{\tilde{\mathbf{x}}}} \left|\left|P^t_\varepsilon \mathbf{q}_{\tilde{\mathbf{x}}} - \nu_{\tilde{\mathbf{x}}}\right|\right|^{\frac{1}{t}}$$

where $T_{\tilde{\mathbf{x}}}$ is the period at which $(\mathbf{X}, P_\varepsilon)$ attains $\nu_{\tilde{\mathbf{x}}}$ once its in $\tilde{\mathbf{x}}$ and $\mathbf{q}_{\tilde{\mathbf{x}}}$ is the PMF that places most weight on the initial state of $(\mathbf{X}, P_\varepsilon)$ in $\tilde{\mathbf{x}}$. Though the definition above applies to all basins of attraction of $(\mathbf{X}, P_\varepsilon)$, we are specially interested in the contagion rate within the basin of attraction of the long-run $\boldsymbol{\varepsilon}$-stable set. If the $\boldsymbol{\varepsilon}$-stable set is that in which all players play (coordinate on) the same action, then the contagion rate within its basin of attraction measures the rate at which the corresponding action spreads across the population once its *threshold* has been reached. By threshold we mean the minimum number of agents required to initially play the action before it diffuses to the rest of the population.

We show in the theorem below that $\gamma_{\tilde{\mathbf{x}}}$ can be deduced from the spectral properties of the transition matrix P_ε and of normalized adjacency matrix of the interaction network. That is the matrix formed by normalizing each players influence by their degree.

Theorem 3: *Let $(\mathbf{X}, P_\varepsilon)$ be a model of learning with mistakes and $\tilde{\mathbf{x}}^*$ be the basin of attraction of its long-run $\boldsymbol{\varepsilon}$-stable set. Let $1 > \lambda_2 \geq \cdots \geq \lambda_{min}$ and $1 \geq \eta_2 \geq \cdots \geq \eta_n$ be the eigenvalue spectra of P_ε and the normalized adjacency matrix of the network respectively.*

(i) If L is the number of metastable sets of $(\mathbf{X}, P_\varepsilon)$, then $\gamma_{\tilde{\mathbf{x}}^*} \leq |\lambda_{L+1}|$

(ii) If the payoff structure is such that there do not exists metastable sets induced by the network, then $\gamma_{\tilde{\mathbf{x}}^*} \leq |\eta_2|$

Proof. See Appendix A.6 □

Theorem 3 states that provided the payoff structure is such that the network does not induce metastable sets, then the contagion rate can be deduced from the second largest eigenvalue (or the spectral gap) of the normalized adjacency matrix. The second eigenvalue of an adjacency matrix is a well studied property of networks, normally characterized in terms of its *conductance* measure. Generally, network topologies that are sparsely connected (such as the one-dimensional cyclic structure) and those that are near-completely decomposable have a higher second largest eigenvalue than identical size networks that are densely connected. This argument follows directly from the principal of *interlacing eigenvalues.*

Theorems 2 and 3 can together be employed to characterize strategic diffusion in networks. The specific cases in the literature are Morris (2000), Lee *et al.* (2003), Montanari and Saberi (2010) and Young (2011). For example the finding in Morris (2000), Montanari and Saberi (2010) and Young (2011) that strategic diffusion is faster if the network structure is made up of cohesive subgroups, can be derived as a corollary of Theorems 2 and 3. The following example illustrates this argument.

Example 4: Consider two extreme cases in which n players are arranged in a cirlce such that each is connected to two other players (hereafter cyclic interaction structure) and that in which they interact globally (that is each player interacts with every other player). Let agents play a coordination games with payoff structure in Table 1. The direct implication is that there are two metastable sets $\vec{B}$ and $\vec{A}$ with respective basins of attraction $\tilde{B}$ and $\tilde{A}$. Assume that the learning process is at $\tilde{B}$. It follows directly from (22) that under cyclic interactions, $d_c(\tilde{B}, \tilde{A}) = \frac{2}{3n}$ and that under global interactions is $d_g(\tilde{B}, \tilde{A}) = \frac{1}{3}$. Assume that the dynamics is governed by (6) with $\beta = 0$ (that is, mistakes are uniformly and randomly distributed), which implies that the risk dominant action $\vec{A}$ is the $\boldsymbol{\varepsilon}$-stable state. Then the cost of the transition $\tilde{B} \to \tilde{A}$ under cyclic and global interactions are respectively $c_c(\tilde{B}, \tilde{A}) = -\frac{2}{3n}\ln(\frac{\varepsilon}{2})$ and $c_g(\tilde{B}, \tilde{A}) = -\frac{1}{3}\ln(\frac{\varepsilon}{2})$

From Theorem 2, we thus have the respective expected waiting times to be:

$$\text{(27a)} \qquad \mathbb{E}T_c(\vec{B}, \tilde{A}) \geq Ke^{-\frac{2}{3}\ln(\frac{\varepsilon}{2})} = \left(\frac{\varepsilon}{2}\right)^{-\frac{2}{3}}$$

$$\text{(27b)} \qquad \mathbb{E}T_g(\vec{B}, \tilde{A}) \geq Ke^{-n\frac{1}{3}\ln(\frac{\varepsilon}{2})} = \left(\frac{\varepsilon}{2}\right)^{-\frac{n}{3}}$$

Implying that $\mathbb{E}T_c(\vec{B}, \tilde{A}) > \mathbb{E}T_g(\vec{B}, \tilde{A})$ for any value of $n \geq 3$. Theorem 3 on the other hand directly implies that once the process has entered the boundaries of $\tilde{A}$, the contagion rate $\gamma^c_{\tilde{A}}$

under cyclic interactions is greater then $\gamma^g_{\tilde{A}}$ under global interactions for any $n \geq 4$. This directly implies that the network topology with optimal rate of strategic diffusion lies between these to extreme cases. Thats is, it should consists of subgroups that are small enough to favor lower expected waiting times, but large enough with high connectivity to favor a higher contagion rate.

The expressions for the expected waiting time in (27a) and (27b) are also consistent with the findings in Ellison (1993) and others, that under cyclic interactions the waiting time is independent of the population size and it is not the case for global interaction.

6. Conclusion

Stochastic evolutionary modeling has been and still is an important approach in game theory. It is most appealing in its remarkable ability to select among multiple equilibria and as a means of modeling boundedly-rational strategic players. This paper developed a general framework for noisy stochastic evolutionary dynamics with the objective of circumventing some of the main limitations and criticisms surrounding such models. The first main contribution of this paper has been to define the concept of epsilon-stability as a more robust solution concept than the commonly used notion of stochastic stability. We then provided an efficient algorithm for computing epsilon-stable sets that is based on fairly fundamental measures.

The second main contribution is the derivation of bounds for expected waiting time, metastability and contagion rate. The expressions we provided for the expected waiting time and metastability can in addition (to being measures for characterizing the short-run and medium-run behavior of learning dynamics) be independently employed as measures for equilibrium selection in some specific settings of stochastic evolutionary dynamics. In conclusion, we hope that the characterization we have provided in this paper will fix some of the skepticism toward stochastic evolutionary models as a mechanism for equilibrium selection, and consequently allow for more of their application to modeling social and economic behavior.

A. Appendix

A.1. Proof of Lemma 1

Note that $\pi_\varepsilon(\tilde{\mathbf{x}}) = \sum_{\mathbf{x}\in\tilde{\mathbf{x}}} \pi_\varepsilon(\mathbf{x})$. Let $\#\mathbf{X}$ and $\#\tilde{\mathbf{X}}$ be the cardinalities of $\mathbf{X}$ and $\tilde{\mathbf{X}}$ respectively. Define an event matrix $\mathcal{E}$ as an $\#\mathbf{X} \times \#\tilde{\mathbf{X}}$ matrix whose entries take on a value one if a state $\mathbf{x} \in \mathbf{X}$ belongs to $\tilde{\mathbf{x}} \in \tilde{\mathbf{X}}$ and zero otherwise. Denote by $\mathcal{E}_{\tilde{\mathbf{x}}}$ for the $\tilde{\mathbf{x}}^{\text{th}}$ column of $\mathcal{E}$. It then follows that $P_\varepsilon \mathcal{E} = \mathcal{E}\tilde{P}_\varepsilon$, and that

$$\pi_\varepsilon(\tilde{\mathbf{x}}) = \pi_\varepsilon \mathcal{E}_{\tilde{\mathbf{x}}} \quad \forall\, \tilde{\mathbf{x}} \in \tilde{\mathbf{X}}$$

Consequently, $\pi_\varepsilon \mathcal{E} = \pi_\varepsilon P_\varepsilon \mathcal{E} = \pi_\varepsilon \mathcal{E} \tilde{P}_\varepsilon$. Implying that $\pi_\varepsilon \mathcal{E}$ is the stationary distribution of $\tilde{P}_\varepsilon$, hence $\tilde{\pi}_\varepsilon = \pi_\varepsilon \mathcal{E}$.

A.2. *Proof of Proposition 1*

Consider the subsets $\tilde{\mathbf{x}}, \tilde{\mathbf{y}} \in \tilde{\mathbf{X}}$ such that $\tilde{\mathbf{x}} \cap \tilde{\mathbf{y}} = \emptyset$, and let $U_{\tilde{\mathbf{x}}\tilde{\mathbf{y}}} = \tilde{\mathbf{x}} \cup \tilde{\mathbf{y}}$ and $U^c_{\tilde{\mathbf{x}}\tilde{\mathbf{y}}}$ its complement. From *irreducibility* of $(\tilde{\mathbf{X}}, \tilde{P}_\varepsilon)$ it follows that

$$\pi_\varepsilon(\tilde{\mathbf{x}})\tilde{P}_\varepsilon(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) + \pi_\varepsilon(\tilde{\mathbf{y}})\tilde{P}_\varepsilon(\tilde{\mathbf{y}}, \tilde{\mathbf{x}}) + \pi_\varepsilon(U^c_{\tilde{\mathbf{x}}\tilde{\mathbf{y}}})\tilde{P}_\varepsilon(U^c_{\tilde{\mathbf{x}}\tilde{\mathbf{y}}}, \tilde{\mathbf{x}}) = \pi_\varepsilon(\tilde{\mathbf{x}})$$

$$\pi_\varepsilon(\tilde{\mathbf{x}})(1 - \tilde{P}_\varepsilon(\tilde{\mathbf{x}}, \tilde{\mathbf{x}})) = \pi_\varepsilon(\tilde{\mathbf{y}})\tilde{P}_\varepsilon(\tilde{\mathbf{y}}, \tilde{\mathbf{x}}) + \pi_\varepsilon(U^c_{\tilde{\mathbf{x}}\tilde{\mathbf{y}}})\tilde{P}_\varepsilon(U^c_{\tilde{\mathbf{x}}\tilde{\mathbf{y}}}, \tilde{\mathbf{x}})$$

$$\frac{\pi_\varepsilon(\tilde{\mathbf{y}})}{\pi_\varepsilon(\tilde{\mathbf{x}})} = \frac{1 - \tilde{P}_\varepsilon(\tilde{\mathbf{x}}, \tilde{\mathbf{x}})}{\tilde{P}_\varepsilon(\tilde{\mathbf{y}}, \tilde{\mathbf{x}})} - \frac{\pi_\varepsilon(U^c_{\tilde{\mathbf{x}}\tilde{\mathbf{y}}})}{\pi_\varepsilon(\tilde{\mathbf{x}})} \frac{\tilde{P}_\varepsilon(U^c_{\tilde{\mathbf{x}}\tilde{\mathbf{y}}}, \tilde{\mathbf{x}})}{\tilde{P}_\varepsilon(\tilde{\mathbf{y}}, \tilde{\mathbf{x}})}$$

By substituting for $1 - \tilde{P}_\varepsilon(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) = \tilde{P}_\varepsilon(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}^c)$, we then have

$$\frac{\pi_\varepsilon(\tilde{\mathbf{y}})}{\pi_\varepsilon(\tilde{\mathbf{x}})} \leq \frac{\tilde{P}_\varepsilon(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}^c)}{\tilde{P}_\varepsilon(\tilde{\mathbf{y}}, \tilde{\mathbf{x}})}$$

From the definition of $(\tilde{\mathbf{X}}, \tilde{P}_\varepsilon)$, we have

$$\begin{aligned} \tilde{P}_\varepsilon(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}^c) &= \frac{1}{\pi_\varepsilon(\tilde{\mathbf{x}})} \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} \sum_{\mathbf{y} \in \tilde{\mathbf{x}}^c} \pi_\varepsilon(\mathbf{x}) P_\varepsilon(\mathbf{x}, \mathbf{y}) \\ &\leq \frac{1}{\pi_\varepsilon(\tilde{\mathbf{x}})} \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} \sum_{\mathbf{y} \in \tilde{\mathbf{x}}^c} \pi_\varepsilon(\mathbf{x}) \max_{\mathbf{x} \in \tilde{\mathbf{x}}} P_\varepsilon(\mathbf{x}, \mathbf{y}) \\ &= \sum_{\mathbf{y} \in \tilde{\mathbf{x}}^c} \max_{\mathbf{x} \in \tilde{\mathbf{x}}} P_\varepsilon(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{x} \in \tilde{\mathbf{x}}} P_\varepsilon(\mathbf{x}, \tilde{\mathbf{x}}^c) \end{aligned}$$

Similarly,

$$\begin{aligned} \tilde{P}_\varepsilon(\tilde{\mathbf{y}}, \tilde{\mathbf{x}}) &= \frac{1}{\pi_\varepsilon(\tilde{\mathbf{y}})} \sum_{\mathbf{y} \in \tilde{\mathbf{y}}} \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} \pi_\varepsilon(\mathbf{y}) P_\varepsilon(\mathbf{y}, \mathbf{x}) \\ &\geq \frac{1}{\pi_\varepsilon(\tilde{\mathbf{y}})} \sum_{\mathbf{y} \in \tilde{\mathbf{y}}} \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} \pi_\varepsilon(\mathbf{y}) \min_{\mathbf{y} \in \tilde{\mathbf{y}}} P_\varepsilon(\mathbf{y}, \mathbf{x}) \\ &= \sum_{\mathbf{x} \in \tilde{\mathbf{x}}} \min_{\mathbf{y} \in \tilde{\mathbf{y}}} P_\varepsilon(\mathbf{y}, \mathbf{x}) = \min_{\mathbf{y} \in \tilde{\mathbf{y}}} P_\varepsilon(\mathbf{y}, , \tilde{\mathbf{x}}) \end{aligned}$$

which completes the proof.

A.3. *Proof of Theorem 1*

(i) The proof of Theorem 1 involves deriving bounds for the probabilities on the right hand side of (9). We adopt the notation $\Omega^i_\varepsilon \in \mathbf{\Omega}_\varepsilon$ for the typical metastable set of $(\mathbf{X}, P_\varepsilon)$, and $\tilde{\mathbf{x}}_i \in \tilde{\mathbf{X}}$ for its corresponding basin of attraction. We also write $\tilde{\mathbf{x}}^c$ for the complement of $\tilde{\mathbf{x}}$. Through out the proof, we write $n(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$ for the number of agents who must switch their action by mistake for the transition from $\mathbf{x} \in \Omega^i_\varepsilon$ into the boundaries of $\tilde{\mathbf{x}}_j$ to occur in a single time step. From the definition of the transition probabilities of the collapsed process $(\tilde{\mathbf{X}}, \tilde{P}_\varepsilon)$ we have

$$\begin{aligned}
\max_{\mathbf{x}\in\tilde{\mathbf{x}}_i} P_\varepsilon(\mathbf{x}, \tilde{\mathbf{x}}_i^c) &= P_\varepsilon(\mathbf{x}\in\Omega_\varepsilon^i, \tilde{\mathbf{x}}_i^c) = \sum_{\mathbf{y}\in\tilde{\mathbf{x}}_i^c} P_\varepsilon(\mathbf{x}\in\Omega_\varepsilon^i, \mathbf{y}) \\
&\leq \sum_{\mathbf{y}\in\tilde{\mathbf{x}}^c} \max_{\mathbf{y}\in\tilde{\mathbf{x}}^c} P_\varepsilon(\mathbf{x}\in\Omega_\varepsilon^i, \mathbf{y}) = K_1 \max_{\mathbf{y}\in\tilde{\mathbf{y}}\in\tilde{\mathbf{x}}^c} P_\varepsilon(\mathbf{x}\in\Omega_\varepsilon^i, \mathbf{y}) \\
&= K_1 \max_{\tilde{\mathbf{x}}_j\neq\tilde{\mathbf{x}}_i} \mathbb{P}\Big(n(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \geq d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n\Big) \qquad \text{(A.1)}
\end{aligned}$$

Where K_1 is some positive constant proportional to the cardinality of $\tilde{\mathbf{x}}^c$. Similarly, for $\Omega_\varepsilon^j \in \tilde{\mathbf{x}}_j \in \tilde{\mathbf{x}}^c$ and $\mathbf{x}\in\tilde{\mathbf{x}}$,

$$\begin{aligned}
\min_{\mathbf{y}\in\tilde{\mathbf{y}}} P_\varepsilon(\mathbf{y}, \tilde{\mathbf{x}}) &\approx P_\varepsilon(\mathbf{y}\in\Omega_\varepsilon^j, \tilde{\mathbf{x}}) = \sum_{\mathbf{x}\in\tilde{\mathbf{x}}} P_\varepsilon(\mathbf{y}\in\Omega_\varepsilon^j, \mathbf{x}) \\
&\approx K_2 \max_{\mathbf{x}\in\tilde{\mathbf{x}}} P_\varepsilon(\mathbf{y}\in\Omega_\varepsilon^j, \mathbf{x}) \\
&\geq K_2 \mathbb{P}\Big(n(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_i) \geq d(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_i)n\Big). \qquad \text{(A.2)}
\end{aligned}$$

where K_2 is some positive constant.We thus seek to provide a bound on $\mathbb{P}(n(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \geq d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n)$ for every pair $(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \in \tilde{\mathbf{X}}$.

The derivation follows a combinatorial argument, where we first transform individual transition probabilities into Boolean random variables. That is, let $\mathbb{P}_i(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) = \mathbb{P}_i(x_{t+1} = y|\mathbf{x}_t = \mathbf{x})$ for $y \in \mathbf{y} \in \Omega_\varepsilon^j \in \tilde{\mathbf{x}}_j$ and $y \notin \mathbf{x} \in \Omega_\varepsilon^i \in \tilde{\mathbf{x}}_i$. Define a parameter $p \in [0, 1]$ such that if $\mathbb{P}_i(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \geq p$ agent i chooses $y \in \mathbf{y}$ or else he does not. This leads to a random variable denoted by I_i, which is equal to one if i chooses y and zero otherwise. Let $I = (I_1, \cdots, I_n)$ be the realization of I_i for all $i \in N$. From the definition of $n(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$, we then rephrase our problem as the case of bounding $\mathbb{P}\Big(\sum_{i=1}^n I_i \geq d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n\Big)$.

Now, consider the problem of a binomial independent sampling over the vector I, $Bin(n, \sigma)$, such that with probability σ, I_i is picked and with $1-\sigma$ it is not. Denote the n-dimensional vector generated by $Bin(n, \sigma)$ by $\mathbf{u} = (u_1, \cdots, u_n)$, where $\mathbb{P}(u_i = 1) = \sigma$ and $\mathbb{P}(u_i = 0) = 1-\sigma$. We can then regard the problem of bounding $\mathbb{P}\Big(\sum_{i=1}^n I_i \geq d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n\Big)$ as determining the probability of "efficiently" finding a subset $S \subseteq N$ of at least $d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n$ players all of whom simultaneously switch to play y. Define an event $\forall_{i\in S} I_i = 1$; that is all members of S choose y, and consequently $\mathbb{P}(\forall_{i\in S} I_i = 1)$ is the probability that all $i \in S$ choose y. We can then define the following conditional relation,

$$\mathbb{P}\left(\sum_{i=1}^n I_i \geq d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n\right) \leq \frac{\mathbb{E}\Big[\forall_{i\in S} I_i = 1\Big]}{\mathbb{E}\Big[\forall_{i\in S} I_i = 1 | \sum_{i=1}^n I_i \geq d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n\Big]}, \qquad \text{(A.3)}$$

where the expectations are taken over the vector $\mathbf{u}$. Since the elements of $\mathbf{u}$ are a result of independent sampling, we have

$$\mathbb{E}[\forall_{i\in S} I_i = 1] \leq \sum_{S\subseteq N} \left(\sigma^{\#S}(1-\sigma)^{n-\#S} \prod_{i\in S} \mathbb{P}_i(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)\right) = \mathbb{E}_{\forall u_i\in\mathbf{u}}\left[\prod_{i=1}^n (\mathbb{P}_i(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)|u_i)\right],$$

where the first inequality follows from the fact that $\mathbb{P}(\forall_{i\in S} I_i = 1) \leq \prod_{i\in S} \mathbb{P}_i(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$. It follows that,

$$\mathbb{E}\Big[\forall_{i\in S} I_i = 1\Big] \leq \prod_{i=1}^{n} \mathbb{E}_{u_i}\Big[\mathbb{P}_i(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)|u_i\Big] = \prod_{i=1}^{n} \Big(\sigma\mathbb{P}_i(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) + 1 - \sigma\Big) \tag{A.4}$$

If we define $\mathbb{P}_A(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}_i(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$ as the arithmetic average of all $\mathbb{P}_i(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$, then from the convex relation between the logarithms of the arithmetic and geometric means,

$$\frac{1}{n}\sum_{i=1}^{n} \ln\Big(\sigma\mathbb{P}_i(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) + 1 - \sigma\Big) \leq \ln\Big(\sigma\mathbb{P}_A(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) + 1 - \sigma\Big)$$

$$\mathbb{E}\Big[\forall_{i\in S} I_i = 1\Big] \leq \prod_{i=1}^{n} \Big(\sigma\mathbb{P}_i(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) + 1 - \sigma\Big) \leq \Big(\sigma\mathbb{P}_A(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) + 1 - \sigma\Big)^n \tag{A.5}$$

To obtain the bound for $\mathbb{E}\Big[\forall_{i\in S} I_i = 1 | \sum_{i=1}^{n} I_i \geq d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n\Big]$, recall that $1-\sigma$ is the probability that $u_i = 0$. We also note that if at least $d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n$ of the elements of $\mathbf{u}$ are ones, then there are at most $n - d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n$ zeros, that is at most $n - d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n$ agents are not in set S. It follows that

$$\mathbb{E}\left[\forall_{i\in S} I_i = 1 | \sum_{i=1}^{n} I_i \geq d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n\right] \geq (1-\sigma)^{(1-d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j))n} \tag{A.6}$$

Equations (A.5) together with (A.6) when substituted into (A.3) yield,

$$\mathbb{P}\left(\sum_{i=1}^{n} I_i \geq d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n\right) \leq \left(\frac{\Big(\sigma\mathbb{P}_A(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) + 1 - \sigma\Big)}{(1-\sigma)^{(1-d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j))}}\right)^n, \tag{A.7}$$

If we choose σ that optimizes the quantity $g = \frac{\big(\sigma\mathbb{P}_A(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) + 1 - \sigma\big)}{(1-\sigma)^{(1-d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j))}}$ (by equating the derivative to zero and solving for σ) and substituting back gives

$$\mathbb{P}\left(\sum_{i=1}^{n} I_i \geq d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n\right) \leq \left(\left(\frac{\mathbb{P}_A(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)}{d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)}\right)^{d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)} \left(\frac{1-\mathbb{P}_A(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)}{1-d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)}\right)^{1-d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)}\right)^n \tag{A.8}$$

Re-expressing (A.8) in exponential form results to

$$\mathbb{P}(n(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \geq d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n) \leq e^{n\left[d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)\ln\left(\frac{\mathbb{P}_A(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)}{d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)}\right) + (1-d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j))\ln\left(\frac{1-\mathbb{P}_A(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)}{1-d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)}\right)\right]} \tag{A.9}$$

Note that each $\mathbb{P}_i(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) = \mathscr{P}_i(x_{t+1} = y|\mathbf{x}_t = \mathbf{x} \in \Omega_\varepsilon^i) \ll 1$, hence $\ln(1 - \mathbb{P}_A(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)) \ll \ln(1 - d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j))$ such that

$$\mathbb{P}(n(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \geq d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n) \leq e^{-n[f(d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)) - d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)\ln(\mathbb{P}_A(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j))]} \tag{A.10}$$

where $f(d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)) = d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)\ln\left(\frac{d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)}{1-d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)}\right) + \ln\left(d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)\right)$. Recall that the cost of the transition $\tilde{\mathbf{x}}_i \to \tilde{\mathbf{x}}_j$ is

$$c(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) = -d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)\ln\left(\mathbb{P}_A(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)\right).$$

It then follows from the definition of the resistance $\mathscr{R}(\tilde{\mathbf{x}}_i)$ of $\tilde{\mathbf{x}}_i$ that

$$\max_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} \mathbb{P}(n(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \geq d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)n) \leq \max_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} e^{-n[f(d(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j))+c(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j)]} = e^{-n[f(d(\tilde{\mathbf{x}}_i))+\mathscr{R}(\tilde{\mathbf{x}}_i)]} \tag{A.11}$$

where $d(\tilde{\mathbf{x}}_i) = \arg\min_{d(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j)} c(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$.

To obtain a lower bound for $\mathbb{P}(n(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_i) \geq d(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_i)n)$ from the steps above, we assume that there exists a positive constant K_3 for which the geometric mean is greater than the arithmetic mean and likewise for the inequality in (A.3) and (A.6) (the rest of the proof follows), such that for some positive constant K_4,

$$\mathbb{P}(n(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_i) \geq d(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_i)n) \geq K_4 e^{-n[f(d(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_i))+c(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_i)]}$$

Now, consider a typical path $H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i} = (\tilde{\mathbf{x}}_j, \cdots, \tilde{\mathbf{x}}_\kappa, \cdots, \tilde{\mathbf{x}}_i)$ from $\tilde{\mathbf{x}}_j$ to $\tilde{\mathbf{x}}_i$. By substituting for the expressions in (A.1) and (A.2), we obtain the quotient $\frac{\pi_\varepsilon(\tilde{\mathbf{x}}_j)}{\pi_\varepsilon(\tilde{\mathbf{x}}_i)}$ to be

$$\frac{\pi_\varepsilon(\tilde{\mathbf{x}}_j)}{\pi_\varepsilon(\tilde{\mathbf{x}}_i)} \leq \frac{\max_{\mathbf{x}_i \in \tilde{\mathbf{x}}_i} P_\varepsilon(\mathbf{x}_i, \tilde{\mathbf{x}}_i^c)}{\min_{\mathbf{x}_j \in \tilde{\mathbf{x}}_j} P_\varepsilon(\mathbf{x}_j, \tilde{\mathbf{x}}_j)} \tag{A.12}$$

$$\leq K e^{n[(c(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_i)-\mathscr{R}(\tilde{\mathbf{x}}_i))+(f(d(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_i))-f(d(\tilde{\mathbf{x}}_i)))]} \tag{A.13}$$

By applying the chain rule on the quotients, we have

$$\frac{\pi_\varepsilon(\tilde{\mathbf{x}}_j)}{\pi_\varepsilon(\tilde{\mathbf{x}}_i)} = \frac{\pi_\varepsilon(\tilde{\mathbf{x}}_j)}{\pi_\varepsilon(\tilde{\mathbf{x}}_{j+1})} \cdots \frac{\pi_\varepsilon(\tilde{\mathbf{x}}_\kappa)}{\pi_\varepsilon(\tilde{\mathbf{x}}_{\kappa+1})} \cdots \frac{\pi_\varepsilon(\tilde{\mathbf{x}}_{i-1})}{\pi_\varepsilon(\tilde{\mathbf{x}}_i)}$$

$$\leq K e^{n\left[\sum_{\kappa=i}^{j-1}(c(\tilde{\mathbf{x}}_\kappa,\tilde{\mathbf{x}}_{\kappa+1})-\mathscr{R}(\tilde{\mathbf{x}}_{\kappa+1}))+\sum_{\kappa=i}^{j-1}(f(d(\tilde{\mathbf{x}}_\kappa,\tilde{\mathbf{x}}_{\kappa+1}))-f(d(\tilde{\mathbf{x}}_{\kappa+1})))\right]} \tag{A.14}$$

$$= KF\left(H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i}\right) e^{n\left[\sum_{\kappa=i}^{j-1}(c(\tilde{\mathbf{x}}_\kappa,\tilde{\mathbf{x}}_{\kappa+1})-\mathscr{R}(\tilde{\mathbf{x}}_{\kappa+1}))\right]} \tag{A.15}$$

where $F\left(H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i}\right) = e^{n\left[\sum_{\kappa=i}^{j-1}(f(d(\tilde{\mathbf{x}}_\kappa,\tilde{\mathbf{x}}_{\kappa+1}))-f(d(\tilde{\mathbf{x}}_{\kappa+1})))\right]}$ and

$$\phi(H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i}) = \sum_{\kappa=i}^{j-1} (c(\tilde{\mathbf{x}}_\kappa, \tilde{\mathbf{x}}_{\kappa+1}) - \mathscr{R}(\tilde{\mathbf{x}}_{\kappa+1}))$$

is the path potential of $H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i}$. Implying that $\pi_\varepsilon(\tilde{\mathbf{x}}_i) > \pi_\varepsilon(\tilde{\mathbf{x}}_j)$ if $\phi(H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i}) < 0$.We thus have $\phi(H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i}) < 0$ as the necessary condition for $\pi_\varepsilon(\tilde{\mathbf{x}}_i) > \pi_\varepsilon(\tilde{\mathbf{x}}_j)$.

Consider all paths $\mathscr{H}_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i}$ from $\tilde{\mathbf{x}}_j \to \tilde{\mathbf{x}}_i$ and let $H^*_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i} = \arg\min_{H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i} \in \mathscr{H}_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i}} \phi(H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i})$. If $\phi(H^*_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i}) < 0$ then for all $H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i} \neq H^*_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i}$ we have that $\phi(H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i}) < 0$.

Similarly, let $\tilde{\mathbf{x}}^*_j = \arg\max_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} \phi(H^*_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i})$. If $\phi(H^*_{\tilde{\mathbf{x}}^*_j\tilde{\mathbf{x}}_i}) < 0$ then for all $H^*_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i} \neq H^*_{\tilde{\mathbf{x}}^*_j\tilde{\mathbf{x}}_i}$ we have that $\phi(H^*_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_i}) < 0$. Implying that a metastable set for which $\phi(H^*_{\tilde{\mathbf{x}}^*_j\tilde{\mathbf{x}}_i}) < 0$ belongs to the union set containing the $\boldsymbol{\varepsilon}$-stable set.

(ii) The proof of Theorem 1 *(ii)* makes use of the combinatorial methods of Freidlin and Wentzell (1984, Chapter 6, Lemma 3.1). That is, if we define a number $q_\varepsilon(\mathbf{x})$ for every state $\mathbf{x} \in \mathbf{X}$ with its associated set of $\mathbf{x}$-trees $\mathscr{T}_\mathbf{x}$ as

$$q_\varepsilon(\mathbf{x}) = \sum_{\tau \in \mathscr{T}_\mathbf{x}} \prod_{(\mathbf{y},\mathbf{z}) \in \tau} P_\varepsilon(\mathbf{t}, \mathbf{z}) \tag{A.16}$$

Then the stationary distribution π_ε of P_ε can be equivalently expressed as

$$\pi_\varepsilon(\mathbf{x}) = \frac{q_\varepsilon(\mathbf{x})}{\sum_{\mathbf{y}\in\mathbf{X}} q_\varepsilon(\mathbf{y})} \quad \text{for each } \mathbf{x}\in\mathbf{X} \tag{A.17}$$

In a similar manner, we can define an equivalent number $q_\varepsilon(\tilde{\mathbf{x}}_i)$ for a typical state $\tilde{\mathbf{x}}_i \in \tilde{\mathbf{X}}$ as

$$q_\varepsilon(\tilde{\mathbf{x}}_i) = \sum_{\tau\in\mathscr{T}_{\tilde{\mathbf{x}}_i}} \prod_{(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_\kappa)\in\tau} \mathbb{P}(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_\kappa) \tag{A.18}$$

where τ is a typical $\tilde{\mathbf{x}}_i$-tree on the graph $\mathscr{W}$, such that $\mathbb{P}(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_\kappa)$ is the transition probability associated the minimum cost path from $\tilde{\mathbf{x}}_j$ to $\tilde{\mathbf{x}}_\kappa$. Since the minimum cost path is also the maximum probability path, we then have that

$$\begin{aligned}
\mathbb{P}(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_\kappa) &= \max_{H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa}\in\mathscr{H}_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa}} \prod_{(\tilde{\mathbf{x}}_l,\tilde{\mathbf{x}}_m)\in H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa}} \tilde{P}_\varepsilon(\tilde{\mathbf{x}}_l,\tilde{\mathbf{x}}_m)\\
&\le K_1 \max_{H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa}\in\mathscr{H}_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa}} \prod_{(\tilde{\mathbf{x}}_l,\tilde{\mathbf{x}}_m)\in H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa}} \mathbb{P}(n(\tilde{\mathbf{x}}_l,\tilde{\mathbf{x}}_m)\ge d(\tilde{\mathbf{x}}_l,\tilde{\mathbf{x}}_m)n)\\
&\le K_1 \max_{H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa}\in\mathscr{H}_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa}} \prod_{(\tilde{\mathbf{x}}_l,\tilde{\mathbf{x}}_m)\in H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa}} e^{-n[f(d(\tilde{\mathbf{x}}_l,\tilde{\mathbf{x}}_m))+c(\tilde{\mathbf{x}}_l,\tilde{\mathbf{x}}_m)]}\\
&= K_1 \max_{H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa}\in\mathscr{H}_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa}} e^{-n\left[\sum_{(\tilde{\mathbf{x}}_l,\tilde{\mathbf{x}}_m)\in H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa}} f(d(\tilde{\mathbf{x}}_l,\tilde{\mathbf{x}}_m))+\sum_{(\tilde{\mathbf{x}}_l,\tilde{\mathbf{x}}_m)\in H_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa}} c(\tilde{\mathbf{x}}_l,\tilde{\mathbf{x}}_m)\right]}\\
&= K_1 e^{-n\left[S(H^*_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa})+w(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_\kappa)\right]} \qquad \text{(A.19)}
\end{aligned}$$

where $S(H^*_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa}) = \sum_{(\tilde{\mathbf{x}}_l,\tilde{\mathbf{x}}_m)\in H^*_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa}} f(d(\tilde{\mathbf{x}}_l,\tilde{\mathbf{x}}_m))$ and $H^*_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa}$ is the minimum cost path. Substituting for $\mathbb{P}(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_\kappa)$ in (A.18) yields

$$\begin{aligned}
q_\varepsilon(\tilde{\mathbf{x}}_i) &= K_1 \sum_{\tau\in\mathscr{T}_{\tilde{\mathbf{x}}_i}} \prod_{(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_\kappa)\in\tau} e^{-n\left[S(H^*_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa})+w(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_\kappa)\right]} && \text{(A.20)}\\
&= K_1 \sum_{\tau\in\mathscr{T}_{\tilde{\mathbf{x}}_i}} \psi(\tau) e^{-n\left[\sum_{(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_\kappa)\in\tau} w(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_\kappa)\right]} && \text{(A.21)}\\
&\le K_5 \max_{\tau\in\mathscr{T}_{\tilde{\mathbf{x}}_i}} \psi(\tau) e^{-n\left[\sum_{(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_\kappa)\in\tau} w(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_\kappa)\right]} && \text{(A.22)}\\
&= K_5 \psi(\tau^*) e^{-n[\mathscr{S}(\tilde{\mathbf{x}}_i)]} && \text{(A.23)}
\end{aligned}$$

where K_5 is some constant, $\psi(\tau) = e^{-n\left[\sum_{(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_\kappa)\in\tau} S(H^*_{\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_\kappa})\right]}$ and the last equality follows from the definition of stochastic potential. That is

$$\mathscr{S}(\tilde{\mathbf{x}}_i) = \min_{\tau\in\mathscr{T}_{\tilde{\mathbf{x}}_i}} \sum_{(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_\kappa)\in\tau} w(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_\kappa).$$

Since $q_\varepsilon(\tilde{\mathbf{x}}_i) \propto \pi_\varepsilon(\tilde{\mathbf{x}}_i)$, it follows that $\tilde{\mathbf{x}}^* = \arg\max_{\tilde{\mathbf{x}}\in\tilde{\mathbf{X}}} \pi_\varepsilon(\tilde{\mathbf{x}}) \equiv \arg\max_{\tilde{\mathbf{x}}\in\tilde{\mathbf{X}}} q_\varepsilon(\tilde{\mathbf{x}})$ is that whose stochastic potential is such that $\mathscr{S}(\tilde{\mathbf{x}}^*) \le \mathscr{S}(\tilde{\mathbf{x}})$ for all $\tilde{\mathbf{x}} \ne \tilde{\mathbf{x}}^*$. That is, $\mathscr{S}(\tilde{\mathbf{x}}^*)$ is the minimum stochastic potential.

A.4. *Proof of Proposition 2*

(i) Recall that for any basin of attraction $\tilde{\mathbf{x}}_i \in \tilde{\mathbf{X}}$, $\mathscr{R}(\tilde{\mathbf{x}}_i) = \min_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} \{c(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)\}$, and $\mathscr{C}\mathscr{R}(\tilde{\mathbf{x}}_i) = \max_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} \{c(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_i)\}$. If $\boldsymbol{\varepsilon} = (\varepsilon, \cdots, \varepsilon)$, then

$$\begin{aligned} \mathscr{R}(\tilde{\mathbf{x}}_i) &= \min_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} \{c(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)\} \\ &= \min_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} \{-d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \ln (\mathbb{P}_A(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j | G))\} \\ &= \min_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} \left\{-d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \ln \left(\varepsilon \frac{1}{n} \sum_{i=1}^{n} \mathcal{P}_i(y | \mathbf{x} \in \Omega_{\varepsilon}^i) \right) \right\} \end{aligned} \tag{A.24}$$

where Ω_{ε}^i is the metastable set of basin of attraction $\tilde{\mathbf{x}}_i$, and $y \in \mathbf{y} \in \Omega_{\varepsilon}^j \in \tilde{\mathbf{x}}_j$, $y \notin \mathbf{x} \in \Omega_{\varepsilon}^i$ Similarly, for an $x \in \mathbf{x} \in \Omega_{\varepsilon}^i$,

$$\mathscr{C}\mathscr{R}(\tilde{\mathbf{x}}_i) = \max_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} \left\{-d(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_i) \ln \left(\varepsilon \frac{1}{n} \sum_{i=1}^{n} \mathcal{P}_i(x | \mathbf{y} \in \Omega_{\varepsilon}^j) \right) \right\} \tag{A.25}$$

If the mistakes probability mass function is bounded, that is $0 < \mathcal{P}_i(x|\mathbf{x}) < 1 \; \forall \; i \in N$, $\forall \; x \in X$ and $\forall \; \mathbf{x} \in \mathbf{X}$, then so must be the averages $\frac{1}{n} \sum_{i=1}^{n} \mathcal{P}_i(.|\mathbf{y})$. Implying that there must exist an ε' close to zero for which the quantities $\ln \left(\varepsilon \frac{1}{n} \sum_{i=1}^{n} \mathcal{P}_i(.|\mathbf{y}) \right)$ are very much close to -1, such that

$$\mathscr{R}(\tilde{\mathbf{x}}_i) \sim \min_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \tag{A.26}$$

$$\mathscr{C}\mathscr{R}(\tilde{\mathbf{x}}_i) \sim \max_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} d(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_i) \tag{A.27}$$

In which case $d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$ is a sufficient measure of the cost of transition $\tilde{\mathbf{x}}_i \to \tilde{\mathbf{x}}_j$. These are precisely the measures employed in the computation of the long-run stochastically stable set. For example $\min_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$ and $\max_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} d(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_i)$ are the radius and coradius respectively in Ellison (2000). Implying that $\pi_{\varepsilon} \sim \lim_{\varepsilon \to 0} \pi_{\varepsilon}$.

(ii) When $\mathcal{P}_i(x|\mathbf{x}) = \frac{1}{\#X} \; \forall \; i \in N$, $\forall \; x \in X$ and $\forall \; \mathbf{x} \in \mathbf{X}$, then

$$\mathscr{R}(\tilde{\mathbf{x}}_i) = -\ln \left(\frac{\varepsilon}{\#X} \right) \min_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \tag{A.28}$$

$$\mathscr{C}\mathscr{R}(\tilde{\mathbf{x}}_i) = -\ln \left(\frac{\varepsilon}{\#X} \right) \max_{\tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}_i} d(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_i) \tag{A.29}$$

In which case $d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$ is a sufficient measure of the cost of transition $\tilde{\mathbf{x}}_i \to \tilde{\mathbf{x}}_j$, an hence $\pi_{\varepsilon} \sim \lim_{\varepsilon \to 0} \pi_{\varepsilon}$.

A.5. *Proof of Theorem 2*

From the definition of $\mathbb{E}T(\Omega_{\varepsilon}^i, \tilde{\mathbf{x}}_j)$, we have that

$$\mathbb{E}T(\Omega_{\varepsilon}^i, \tilde{\mathbf{x}}_j) = \frac{1}{\mathbb{P}(n(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \geq d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) n)} \tag{A.30}$$

It then follows from (A.10) that

$$\mathbb{E}T(\Omega^i_\varepsilon, \tilde{\mathbf{x}}_j) \geq e^{n[f(d(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j))+c(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j)]} \tag{A.31}$$

In the case of metastability, we have that

$$\mathscr{M}(\Omega^i_\varepsilon) = \max_{\tilde{\mathbf{x}}_j\neq\tilde{\mathbf{x}}_i} \mathbb{E}T(\Omega^i_\varepsilon, \tilde{\mathbf{x}}_j) \geq \max_{\tilde{\mathbf{x}}_j\neq\tilde{\mathbf{x}}_i} e^{n[f(d(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j))+c(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j)]} = e^{n[f(d(\tilde{\mathbf{x}}_i))+\mathscr{R}(\tilde{\mathbf{x}}_i)]} \tag{A.32}$$

A.6. Proof of Theorem 3

(i) The proof makes use of the spectral properties and near-complete decomposability of transition matrix P_ε. Under the mutationless process $(\mathbf{X}, P)$, the transition matrix is completely decomposable into the form

$$P = \begin{pmatrix} M_1^* & & & & \\ & \ddots & & & \\ & & M_l^* & & \\ & & & \ddots & \\ & & & & M_L^* \end{pmatrix}$$

where M_l^* for $l = 1, \cdots, L$ is a block matrix describing the transitions within each basin of attraction under $(\mathbf{X}, P)$. The rest of the undisplayed elements are zeros and L is the number of limit sets. All leading eigenvalues of the block matrices are ones.

When each $\varepsilon_i(\mathbf{x}) > 0$ (but small) and mistakes probability mass functions are bounded, the transition matrix P_ε is near-completely decomposable into L “loosely” connected block matrices that we denote by M_l for $l = 1, \cdots, L$. That is we can write $P_\varepsilon = P + \epsilon P^*$, where ϵ is a small real number and P^* is an arbitrary $\#\mathbf{X}$ by $\#\mathbf{X}$ matrix. A more detailed exposition on the notion of near-complete decomposability can be found in Simon and Ando (1961). For ϵ small enough, the leading eigenvalues of the diagonal block matrices of P_ε are close to one. Let λ_{i_l} denote the ith eigenvalue of the lth diagonal block matrix, such that $(\lambda_{1_1}, \lambda_{1_2}, \cdots, \lambda_{1_L})$ are the largest eigenvalues in blocks 1 to L, and $(\lambda_{2_1}, \lambda_{2_2}, \cdots, \lambda_{2_L})$ are the respective second largest eigenvalues. Index by n_l as the number of columns in diagonal block l such that the eigenvalue spectrum $\rho(P_\varepsilon)$ of P_ε can be written as $\rho(P_\varepsilon) = (\lambda_{1_1}, \lambda_{2_1}, \cdots, \lambda_{n_{1_1}}, \cdots, \lambda_{1_2}, \cdots, \lambda_{1_l}, \cdots, \lambda_{n_{l_l}}, \cdots, \lambda_{1_L}, \cdots, \lambda_{n_{L_L}})$. The spectral decomposition of $(\mathbf{X}, P_\varepsilon)$ is then given by

$$\begin{aligned} \mathbf{q}_0 P^t_\varepsilon = {} & \mathbf{q}_0\mathbf{r}_{1_1}\mathbf{z}'_{1_1} + \sum_{j=2}^{n_1} \lambda^t_{j_1}\mathbf{q}_0\mathbf{r}_{j_1}\mathbf{z}'_{j_1} + \lambda^t_{1_2}\mathbf{q}_0\mathbf{r}_{1_2}\mathbf{z}'_{1_2} + \sum_{j=2}^{n_2} \lambda^t_{j_2}\mathbf{q}_0\mathbf{r}_{j_2}\mathbf{z}'_{j_2} \\ & + \cdots + \lambda^t_{1_L}\mathbf{q}_0\mathbf{r}_{1_L}\mathbf{z}'_{1_L} + \sum_{j=2}^{n_L} \lambda^t_{j_L}\mathbf{q}_0\mathbf{r}_{j_L}\mathbf{z}'_{j_L} \end{aligned} \tag{A.33}$$

where the “prime” implies the transpose, and $\mathbf{r}_{j_l}$ and $\mathbf{z}_{j_l}$ are the right and left eigenvectors of λ_{j_l}. We would like to show that the second largest eigenvalue λ_{2_l} of M_l describes the rate at which $(\mathbf{X}, P_\varepsilon)$ converges to its quasi-stationary distribution within $\tilde{\mathbf{x}}_l$. Let $\mathbf{x}_l$ be the initial state of $(\mathbf{X}, P_\varepsilon)$ in $\tilde{\mathbf{x}}_l$ and $\mathbf{q}_{\tilde{\mathbf{x}}_l}$ be the $\#\mathbf{X}$-dimensional vector of zeros except a one at the point corresponding to the state $\mathbf{x}_l$. Let t_l be the period at which $(\mathbf{X}, P_\varepsilon)$ is in the state $\mathbf{x}_l$ and $T_{\tilde{\mathbf{x}}_l}$ the period at which it exits $\tilde{\mathbf{x}}_l$ (or equivalently the period at which it attains the quasi-stationary distribution ν_l). Then for $t_l \leq t \leq T_{\tilde{\mathbf{x}}_l}$ and all $l = 1, \cdots, L$,

$$\mathbf{q}_t = \lambda_{1_l}^t \mathbf{q}_{\tilde{\mathbf{x}}_l} \mathbf{r}_{1_l} \mathbf{z}'_{1_l} + \sum_{j=2}^{n_l} \lambda_{j_l}^t \mathbf{q}_{\tilde{\mathbf{x}}_l} \mathbf{r}_{j_l} \mathbf{z}'_{j_l}$$

$$\nu_l = \lim_{t \to T_{\tilde{\mathbf{x}}_l}} \mathbf{q}_t = \lambda_{1_l}^{T_{\tilde{\mathbf{x}}_l}} \mathbf{q}_{\tilde{\mathbf{x}}_l} \mathbf{r}_{1_l} \mathbf{z}'_{1_l} \approx \lambda_{1_l}^t \mathbf{q}_{\tilde{\mathbf{x}}_l} \mathbf{r}_{1_l} \mathbf{z}'_{1_l}$$

where the approximation holds form the fact that λ_{1_l} is close to one for all l. It then follows that

$$\gamma_{\tilde{\mathbf{x}}_l} = \limsup_{t \to T_{\tilde{\mathbf{x}}_l}} \left|\left| P_\varepsilon^t \mathbf{q}_{\tilde{\mathbf{x}}_l} - \nu_{\tilde{\mathbf{x}}_l} \right|\right|^{\frac{1}{t}}$$

$$\approx |\lambda_{2_l}| \limsup_{t \to T_{\tilde{\mathbf{x}}_l}} \left|\left| \mathbf{q}_{\tilde{\mathbf{x}}_l} \mathbf{r}_{2_l} \mathbf{z}'_{2_l} + \sum_{j=3}^{n_l} \left(\frac{\lambda_{j_l}}{\lambda_{2_l}} \right)^t \mathbf{q}_{\tilde{\mathbf{x}}_l} \mathbf{r}_{j_l} \mathbf{z}'_{j_l} \right|\right|^{\frac{1}{t}} \approx |\lambda_{2_l}|$$

Implying that the contagion rate within $\tilde{\mathbf{x}}^*$ is $\gamma_{\tilde{\mathbf{x}}^*} = \gamma_{\tilde{\mathbf{x}}_1} \approx |\lambda_{2_1}| \leq |\lambda_{L+1}|$.

(ii) To prove the second part of the theorem, we consider the linearization of $(\mathbf{X}, P_\varepsilon)$ of the form

$$\mathbf{q}_t \Psi = \mathbf{q}_0 P_\varepsilon^t \Psi = \mathbf{q}_0 \Psi \Pi_\varepsilon^t \tag{A.34}$$

where Ψ is the event matrix derived by stacking into rows all possible realizations of states of $(\mathbf{X}, P_\varepsilon)$ written in the *basis vector* form. The choice basis vector for each player $i \in N$ is a vector of zeros except a one in a position corresponding to the action i is playing. For example for a binary action set $X = \{A, B\}$, a vector $(1, 0)$ implies that i is playing action A and $(0, 1)$ implies that i is playing action B. In the case of two players and binary action set, there are four possible realization such that

$$\Psi = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \tag{A.35}$$

where the first row corresponds to the state in which both players play action A, and so forth. Let the action set $X = (x_1, \cdots, x_m)$ be the same for all $i \in N$. Then for n players, Π_ε is an $nm \times nm$ matrix defined by $\Pi_\varepsilon = \mathscr{A}' \otimes \Sigma$, where $\otimes$ is a Kronecker product, $\mathscr{A}'$ is the transpose of the normalized adjacency matrix $\mathscr{A}$ and Σ is the action-transition matrix defined as follows. For any directly connected pair of players, let $\mathbb{P}(x_j|x_i)$ be the probability that a given player plays action $x_j \in X$ in the next period given that his opponent is playing $x_i \in X$ in the current period. Then Σ is given by

$$\Sigma = \begin{pmatrix} \mathbb{P}(x_1|x_1) & \mathbb{P}(x_2|x_1) & \cdots & \mathbb{P}(x_m|x_1) \\ \mathbb{P}(x_1|x_2) & \mathbb{P}(x_2|x_2) & \cdots & \mathbb{P}(x_m|x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{P}(x_1|x_m) & \mathbb{P}(x_2|x_m) & \cdots & \mathbb{P}(x_m|x_m) \end{pmatrix} \tag{A.36}$$

A detailed exposition on the validity of (A.34) can be found in Asavathiratham (2001, Chapter 5). The following lemma follows directly from (A.34) and the definition of Π_ε above.

Lemma 2: *Let $\rho(\Pi_\varepsilon) = \tilde{\lambda}_1, \cdots, \tilde{\lambda}_{nm}$, $\rho(\mathscr{A}) = (\eta_1, \cdots, \eta_m)$ and $\rho(\Sigma) = (\delta_1, \cdots, \delta_n)$ denote the eigenvalue spectra of Π_ε, $\mathscr{A}$ and Σ respectively.*

- *(a) If λ_1 and $\tilde{\lambda}_1$ are the unique largest eigenvalues of P_ε and Π_ε respectively, then $\lambda_1 = \tilde{\lambda}_1 = 1$.*
- *(b) $\rho(\Pi_\varepsilon) = (\delta_i \eta_j)$ $\forall \delta_i \in \rho(\Sigma)$, $\eta_j \in \rho(\mathscr{A})$.*

Proof. Multiplying (A.34) by the right eigenvector $\mathbf{r}_i$ of P_ε, we have $P_\varepsilon \Psi \mathbf{r}_1 = \Psi \Pi_\varepsilon \mathbf{r}_1$. Since P_ε is a stochastic matrix, $\lambda_1 = 1$, which implies that $P_\varepsilon \Psi \mathbf{r} = \Psi \mathbf{r}_1$, which is true if and only if $\Pi_\varepsilon \mathbf{r}_1 = \mathbf{r}_1$. That is $\tilde{\lambda}_1 = \lambda_1 = 1$. For the proof of Lemma 2 (*b*) see Horn and Johnson (1990, page 245, Theorem 4.2.12). □

It then follows that if the payoff structure is such that the network does not induce metastable sets (which also automatically implies that the network must be connected), then for small enough values of $\varepsilon_i(\mathbf{x})$ $\forall$ $i \in N$ and $\mathbf{x} \in \mathbf{X}$, $\lambda_{L+1} \leq \tilde{\lambda}_{L+1} = \eta_2 \delta_1 = \eta_2$.

Acknowledgments

This study has benefited from comments and suggestions of Julio Davila, Arkadi Predtetchinsk, Antoine Mandel, François Lafond, Hippolyte d'Albis, Giorgio Triulzi and Robin Cowan. We also gratefully acknowledge the fruitful discussions with seminar participants at The Centre d'Economie de La Sorbonne and Paris School of Economics, University of Cergy Pontoise THEMA, Maastricht Lecture Series in Economics and Cournot seminars at Bureau d'Economie Théorique et Appliquée (BETA), University of Strasbourg. This research was supported by UNU-MERIT. The usual disclaimer applies.

References

Akerlof, G. A. (1997). Social distance and social decisions. *Econometrica*, **65** (5), 1005–1027.

Aldous, D. and Fill, J. A. (1994). Reversible markov chains.

— and — (1999). Reversible Markov chains and random walks on graphs.

Alós-Ferrer, C. and Weidenholzer, S. (2008). Contagion and efficiency. *Journal of Economic Theory*, **143** (1), 251–274.

Asavathiratham, C. (2001). *The influence model: a tractable representation for the dynamics of networked Markov chains*. Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science, Ph.D Thesis.

Bergin, J. and Lipman, B. L. (1996). Evolution with state-dependent mutations. *Econometrica*, **64** (4), 943–56.

Binmore, K. and Samuelson, L. (1997). Muddling through: Noisy equilibrium selection. *Journal of Economic Theory*, **74** (2), 235–265.

Blume, L. E. (1995). The statistical mechanics of best-response strategy revision. *Games and Economic Behavior*, **11** (2), 111–145.

— (2003). How noise matters. *Games and Economic Behavior*, **44** (2), 251–271.

Brock, W. A. and Durlauf, S. N. (2001). Discrete choice with social interactions. *Review of Economic Studies*, **68** (2), 235–60.

Ellison, G. (1993). Learning, local interaction, and coordination. *Econometrica*, **61** (5), 1047–1071.

— (2000). Basins of attraction, long-run stochastic stability, and the speed of step-by-step evolution. *Review of Economic Studies*, **67** (1), 17–45.

Foster, D. and Young, P. (1990). Stochastic evolutionary game dynamics. *Theoret. Population Biol.*

Freidlin, M. and Wentzell, A. D. (1984). *Random perturbations of dynamical systems.* New York: Springer Verlag.

Glaeser, E. L. and Scheinkman, J. A. (2001). *Non-Market Interactions.* Tech. rep.

Horn, R. A. and Johnson, C. R. (1990). *Matrix Analysis.* Cambridge University Press.

Huck, S., Kübler, D. and Weibull, J. (2012). Social norms and economic incentives in firms. *Journal of Economic Behavior & Organization*, **83** (2), 173 – 185.

Kandori, M., Mailath, G. J. and Rob, R. (1993). Learning, mutation, and long run equilibria in games. *Econometrica*, **61** (1), 29–56.

— and Rob, R. (1998). Bandwagon effects and long run technology choice. *Games and Economic Behavior*, **22** (1), 30–60.

Lee, I. H., Szeidl, A. and Valentinyi, A. (2003). Contagion and state dependent mutations. *The B.E. Journal of Theoretical Economics*, **3** (1), 1–29.

Maruta, T. (2002). Binary games with state dependent stochastic choice. *Journal of Economic Theory*, **103** (2), 351–376.

Montanari, A. and Saberi, A. (2010). The spread of innovations in social networks. *Proceedings of the National Academy of Sciences.*

Morris, S. (2000). Contagion. *Review of Economic Studies*, **67** (1), 57–78.

Nöldeke, G. and Samuelson, L. (1997). A dynamic model of equilibrium selection in signaling markets. *Journal of Economic Theory*, **73** (1), 118–156.

Simon, H. and Ando, A. (1961). Aggregation of variables in dynamic systems. *Econometrica*, **29** (2), 111–138.

van Damme, E. and Weibull, J. W. (2002). Evolution in games with endogenous mistake probabilities. *Journal of Economic Theory*, **106** (2), 296–315.

Vega-Redondo, F. (1997). The evolution of walrasian behavior. *Econometrica*, **65** (2), 375–384.

Young, H. P. (1993). The evolution of conventions. *Econometrica*, **61** (1), 57–84.

Young, P. H. (2011). The dynamics of social innovation. *Proceedings of the National Academy of Sciences*, **108** (4), 21285–21291.