GALATASARAY UNIVERSITY INSTITUTE OF SOCIAL SCIENCES
ECONOMICS DEPARTMENT

Coevolution of reciprocity and trust

MASTER'S THESIS

Özlem YAŞAR

Thesis Supervisor: Assoc. Prof. Bilge Öztürk Göktuna

August 2022

Preface

I am extremely grateful to my supervisor, Assoc. Prof. Bilge Ozturk Goktuna helped and guided me dedicatedly, even when everyone lost their motivation during the pandemic. I also thank all my professors from Galatasaray University for their valuable contributions during my courses.

I also am thankful to my parents and siblings, who always offered me their help during my life. I wouldn't be where I am if not for them. I am also so grateful to my friend Emre Kaymak for always offering his help.

Finally, I want to dedicate this thesis to my husband and thank him for always being patient and supportive. I am so lucky to have him in my life.

Contents

List of Figures

List of Tables

Abstract

Trust is a crucial concept in society. Many social interactions require trust to a degree. This trust is often the trust in the opponent's probability of complying with social norms. A trustworthy individual abides by the rules of the contract she and the trustee make. More interestingly, if not many, some people do not betray the trustee even when the trustee cannot get her revenge, and it would be only for the benefit of the trusted. If a materialist agent only considers her own material payoff, then we cannot explain this behavior. To illustrate this phenomenon, many economists study reciprocal behaviors, behaviors that cause agents to act trustworthy by forfeiting some of their payoffs. Different reciprocal behaviors are suggested to explain how trustworthiness emerges from various observed encounters.

This paper analyzes the differences between reciprocal behaviors in an evolutionary game theory framework. These reciprocal behaviors ensure that people act trustworthy in social interactions. Otherwise, classical game theory suggests that a materialist agent would only play what is best for her and would not care about others' well-being. However, we observe that this is not the case. The motivations behind reciprocal behaviors are a topic of discussion; however, it is easy to argue that some people work for the well-being of others. For example, we cannot explain some parents' behaviors by an expected future payoff; moreover, there are volunteers of non-profit organizations who make efforts for the development of society. These behaviors cannot be considered materialistic. They are motivated by at least a kind of responsibility, if not complete altruism.

In the game theory literature, economists defined many kinds of reciprocal behaviors. Some experiments suggest that trustworthiness may come from unconditional other-regarding preferences. In contrast, others suggest that it may be due to some degree of responsiveness to the trusting behavior of the opposite party. Cox et al. (2016) indicate that the main drive behind trustworthiness may be vulnerability-responsiveness. Vulnerability-responsiveness is a phenomenon that emerges in an agent when someone else makes himself vulnerable to her by trusting. Our paper studies Cox et al. (2016)'s experiments to find how vulnerability-responsiveness performs against pure altruism and selfishness in an evolutionary context to understand it.

We create a 2-step game similar to the famous investment game. In the investment game, the first player decides to send all of her money to the second player or not. The money is multiplied by 3 when it arrives at the second player. Then, the second player decides to send half her money or not. Even though the socially optimum scenario is choosing to send for both players, the best response for the second player is not to send as it does not provide any gain. However, experiments observe that a positive amount of experimentees choose to send the money back. Therefore, there must exist a motivation for agents to send the money back. In our paper, we create a two-step prisoner's dilemma game where the first player chooses between cooperating and defecting. The second player instead chooses between cooperating conditionally, defecting unconditionally, and cooperating unconditionally. The first player always chooses to cooperate with a probability based on her trust level. However, we assign three strategies to the second-player: (1) altruism; (2) vulnerability-responsiveness; and (3) selfishness. Altruist agents benefit from their opponents' payoffs. On the other hand, vulnerability-responsive agents care about not letting their opponents harm just because their opponents trusted them. Finally, selfish agents think only about themselves. This game allows us to understand under which conditions altruism or vulnerability-responsiveness prevails over pure selfish acts.

We use the evolutionary game theory framework to analyze the genetic fitness of these strategies over time. A set of differential equations that are commonly referred to as replicator equations is created and solved to find the strategy with the best genetic fitness. Replicator equations work by finding how each strategy performs against the mean of all strategies. A strategy that serves better than the average increases over time, while a strategy that performs worse declines. Solutions to these equations denote equilibrium points. Each equilibrium indicates possible stability. To determine if these equilibrium points are stable or not, we find which of these equilibrium points causes the eigenvalues of the Jacobian matrix to be negative. However, not all games have to have a set of evolutionary stable strategies. Instead, some games like hawk and dove may end with an endless cycle similar to a prey-predator relationship. We show that there exist multiple equilibria in our model that may end up with the co-existence of two strategies or complete domination of a strategy.

In the following chapter, we conduct a numerical analysis. To do this, we enumerate our parameters to find a numerical solution. This enumeration ensures that we understand how our model behaves under different initial conditions, and we can understand how each strategy responds to a difference in various parameters. Furthermore, enumeration is necessary to visualize the results of the analytical model. Since our model has three strategies, we need to have a three-dimensioned space to plot our results; however, a plot in the 3D space is often considered questionable and unnecessary. Therefore, we use a 2-simplex and convert our results from cartesian to barycentric coordinates. This way, we are able to show our phase diagrams in a more concise and understandable two-dimensioned space. These phase diagrams allow us to plot our numerical results.

Finally, we create an agent-based model. Agent-based models are computational models that work through multiple iterations based on a simple set of rules. Our steps for making our agent-based model are the following: first, initial agents are created. Then, each agent randomly matches each turn to play the two-step prisoner's dilemma game. One of the agents is randomly chosen to play first. The first player decides to trust her opponent or not. Then, the second player plays her best response. For example, the best response for an altruist agent is to cooperate unconditionally, while a selfish agent always defects unconditionally. After a turn ends, each agent looks at a random agent and decides to imitate him with a probability based on the payoff difference between the two agents. Thus, we are able to imitate replicator dynamics using a computational method. Moreover, an agent-based model can be stochastic more conveniently than replicator equations. Due to these differences, it is not extremely rare to have different results in an agent-based model than in an equation-based model.

In an agent-based model, it is essential to run multiple simulations to see how stochasticity affects the model. However, it is equally important to examine each individual run to understand the dynamics of the model. We carefully analyze multiple runs individually to prevent run-specific details from being lost. Moreover, we study the average dynamics of the agent-based model as well. To do this, we run the simulation under a set of parameters that vary between runs and take the average of each simulation. We also run the simulation multiple times for the same parameter combination to take stochasticity into account. We show both averaged and individual results. Averaged results provide information about the overall direction of the model, while individual results show how a simulation evolves over time.

Our results in all three chapters indicate multiple equilibria. Depending on the parameters, being selfish may be beneficial under some circumstances, while selfish agents and altruist agents may co-exist as well. We can observe this co-existence in all analytical, numerical, and agent-based models. It is also consistent with what we can observe in our environment as well. Furthermore, parameter-dependence is also consistent with the literature because we know that trust is inherited between generations. Therefore, we know that initial conditions affect the economic outcomes of a country.

Our paper contributes to the literature in 3 ways. First, this is the first paper to create a coevolutionary model between trust and reciprocal preferences to our knowledge. We show that an increase in the trust attitude may not always increase reciprocal preferences. This result is important to the literature since it describes a cyclic relationship between high trust and low trustworthiness. Second, this is the first model which analyzes the vulnerability-responsiveness in a theoretical framework to our knowledge. Finally, we create a framework to use replicator equations in two-stage games. In replicator dynamics literature, having a symmetrical one-stage game is almost always prevalent. Our approach shows that evolutionary game theory can be used to study two-stage interactions as well.

A logical next step from this paper could be the study of this model in a social network framework. This way, an economist can understand how the concept of social trust's evolution affects a network's trustworthiness. Moreover, this could prove useful when one studies the formation of a social network. We also provide thresholds for parameters to be used in an experimental study. These parameters can be tested in an experiment to see if agents' behaviors will change significantly. Furthermore, our paper could provide a framework for a coevolutionary centipede game that is asymmetrical. Trust and trustworthiness concepts are naturally intertwined. However, the correlation between them does not have to be positive. High trust in a society can also lead to a malicious agent having a higher profit.

We suggest that this phenomenon could be why distrust is a natural response on many occasions.

Résumé

La confiance est un concept crucial dans la société. De nombreuses interactions sociales nécessitent la confiance dans une certaine mesure. Cette confiance est souvent la confiance dans la probabilité de l'adversaire de se conformer aux normes sociales. Nous définissons cette probabilité comme la fiabilité. Une personne digne de confiance respecte les règles du contrat qu'elle et le fiduciaire concluent. Plus intéressant, sinon beaucoup, certaines personnes ne trahissent pas l'administrateur même lorsque l'administrateur ne peut pas se venger, et ce ne serait que pour le bénéfice de la confiance. Si un agent matérialiste ne considère que son propre gain matériel, alors nous ne pouvons pas expliquer ce comportement. Pour illustrer ce phénomène, de nombreux économistes étudient les comportements réciproques, des comportements qui amènent les agents à agir de manière digne de confiance en renonçant à certains de leurs gains. Différents comportements réciproques sont suggérés pour expliquer comment la fiabilité émerge des diverses rencontres observées.

Cet article analyse les différences entre les comportements réciproques dans un cadre de théorie des jeux évolutifs. Ces comportements réciproques garantissent que les gens agissent de manière digne de confiance dans les interactions sociales. Sinon, la théorie classique des jeux suggère qu'un agent matérialiste ne jouerait que ce qui est le mieux pour lui et ne se soucierait pas du bien-être des autres. Cependant, nous constatons que ce n'est pas le cas. Les motivations derrière les comportements réciproques sont un sujet de discussion ; cependant, il est facile d'affirmer que certaines personnes travaillent pour le bien-être des autres. Par exemple, nous ne pouvons pas expliquer les comportements de certains parents par un gain futur attendu ; de plus, il y a des volontaires d'organisations à but non lucratif qui font des efforts pour le développement de la société. Ces comportements ne peuvent être considérés comme matérialistes. Ils sont motivés par au moins une sorte de responsabilité, sinon par un altruisme complet.

Dans la littérature sur la théorie des jeux, les économistes ont défini de nombreux types de comportements réciproques. Certaines expériences suggèrent que la fiabilité peut provenir de préférences inconditionnelles envers les autres. En revanche, d'autres suggèrent que cela peut être dû à un certain degré de réactivité au comportement de confiance de la partie adverse. Cox et al. (2016) indiquent

que le principal moteur de la fiabilité peut être la sensibilité à la vulnérabilité. La vulnérabilité-réactivité est un phénomène qui émerge chez un agent lorsque quelqu'un d'autre se rend vulnérable à lui en lui faisant confiance. Notre article étudie Cox et al. (2016) pour découvrir comment la sensibilité à la vulnérabilité se comporte contre l'altruisme pur et l'égoïsme dans un contexte évolutif pour le comprendre.

Nous créons un jeu en 2 étapes similaire au célèbre jeu d'investissement. Dans le jeu d'investissement, le premier joueur décide d'envoyer ou non tout son argent au deuxième joueur. L'argent est multiplié par 3 lorsqu'il arrive au deuxième joueur. Ensuite, le deuxième joueur décide d'envoyer ou non la moitié de son argent. Même si le scénario socialement optimal choisit d'envoyer pour les deux joueurs, la meilleure réponse pour le deuxième joueur est de ne pas envoyer car cela ne rapporte aucun gain. Cependant, les expériences observent qu'un nombre positif de personnes expérimentées choisissent de renvoyer l'argent. Par conséquent, il doit exister une motivation pour que les agents renvoient l'argent. Dans notre article, nous créons un jeu de dilemme du prisonnier en deux étapes où le premier joueur choisit entre coopérer et faire défection. Le deuxième joueur choisit à la place entre coopérer conditionnellement, faire défection inconditionnellement et coopérer inconditionnellement. Le premier joueur choisit toujours de coopérer avec une probabilité basée sur son niveau de confiance. Cependant, nous attribuons trois stratégies au second joueur : (1) l'altruisme ; (2) vulnérabilité-réactivité ; et (3) l'égoïsme. Les agents altruistes profitent des gains de leurs adversaires. D'un autre côté, les agents sensibles à la vulnérabilité se soucient de ne pas laisser leurs adversaires nuire simplement parce que leurs adversaires leur font confiance. Enfin, les agents égoïstes ne pensent qu'à eux-mêmes. Ce jeu permet de comprendre dans quelles conditions l'altruisme ou la réceptivité à la vulnérabilité l'emporte sur les actes purement égoïstes.

Nous utilisons le cadre de la théorie des jeux évolutifs pour analyser l'aptitude génétique de ces stratégies au fil du temps. Un ensemble d'équations différentielles communément appelées équations réplicatrices est créé et résolu pour trouver la stratégie avec la meilleure aptitude génétique. Les équations du réplicateur fonctionnent en trouvant les performances de chaque stratégie par rapport à la moyenne de toutes les stratégies. Une stratégie qui sert mieux que la moyenne augmente avec le temps, tandis qu'une stratégie qui fonctionne moins bien diminue. Les solutions à ces équations dénotent des points d'équilibre. Chaque équilibre indique une stabilité possible. Pour déterminer si ces points d'équilibre sont stables ou non, nous trouvons lequel de ces points d'équilibre rend les valeurs propres de la matrice jacobienne négatives. Cependant, tous les jeux ne doivent pas nécessairement avoir un ensemble de stratégies évolutives stables. Au lieu de cela, certains jeux comme le faucon et la colombe peuvent se terminer par un cycle sans fin semblable à une relation proie-prédateur. Nous montrons qu'il existe plusieurs équilibres dans notre modèle qui peuvent aboutir à coexistence de deux stratégies ou domination complète d'une stratégie.

Dans le chapitre suivant, nous procédons à une analyse numérique. Pour ce faire, nous énumérons nos paramètres pour trouver une solution numérique. Cette énumération garantit que nous comprenons comment notre modèle se comporte dans différentes conditions initiales, et nous pouvons comprendre comment chaque stratégie répond à une différence dans divers paramètres. De plus, l'énumération est nécessaire pour visualiser les résultats du modèle analytique. Étant donné que notre modèle a trois stratégies, nous avons besoin d'un espace tridimensionnel pour tracer nos résultats ; cependant, un tracé dans l'espace 3D est souvent considéré comme discutable et inutile. Par conséquent, nous utilisons un 2-simplex et convertissons nos résultats des coordonnées cartésiennes en coordonnées barycentriques. De cette façon, nous sommes en mesure de montrer nos diagrammes de phase dans un espace bidimensionnel plus concis et compréhensible. Ces diagrammes de phase nous permettent de tracer nos résultats numériques.

Enfin, nous créons un modèle basé sur les agents. Les modèles à base d'agents sont des modèles de calcul qui fonctionnent à travers plusieurs itérations basées sur un simple ensemble de règles. Nos étapes pour créer notre modèle à base d'agents sont les suivantes : d'abord, les agents initiaux sont créés. Ensuite, chaque agent correspond au hasard à chaque tour pour jouer au jeu du dilemme du prisonnier en deux étapes. L'un des agents est choisi au hasard pour jouer en premier. Le premier joueur décide de faire confiance ou non à son adversaire. Ensuite, le deuxième joueur joue sa meilleure réponse. Par exemple, la meilleure réponse pour un agent altruiste est de coopérer inconditionnellement, alors qu'un agent égoïste fait toujours défaut inconditionnellement. Après la fin d'un tour, chaque agent regarde un agent au hasard et décide de l'imiter avec une probabilité basée sur la différence de gain entre les deux agents. Ainsi, nous sommes capables d'imiter la dynamique du réplicateur en utilisant une méthode de calcul. De plus, un modèle à base d'agents peut être stochastique plus facilement que les équations de réplicateur. En raison de ces différences, il n'est pas extrêmement rare d'avoir des résultats différents dans un modèle à base d'agents et dans un modèle à base d'équations.

Dans un modèle à base d'agents, il est essentiel d'exécuter plusieurs simulations pour voir comment la stochasticité affecte le modèle. Cependant, il est tout aussi important d'examiner chaque exécution individuelle pour comprendre la dynamique du modèle. Nous analysons soigneusement plusieurs exécutions individuellement pour éviter que les détails spécifiques à l'exécution ne soient perdus. De plus, nous étudions également la dynamique moyenne du modèle à base d'agents. Pour ce faire, nous exécutons la simulation sous un ensemble de paramètres qui varient entre les exécutions et prenons la moyenne de chaque simulation. Nous exécutons également la simulation plusieurs fois pour la même combinaison de paramètres afin de prendre en compte la stochasticité. Nous montrons les résultats moyens et individuels. Les résultats moyens fournissent des informations sur la direction générale du modèle, tandis que les résultats individuels montrent comment une simulation évolue dans le temps.

Nos résultats dans les trois chapitres indiquent des équilibres multiples. Selon les paramètres, être égoïste peut être bénéfique dans certaines circonstances, tandis que des agents égoïstes et des agents altruistes peuvent également coexister. Nous pouvons observer cette coexistence dans tous les modèles analytiques, numériques et multi-agents. Cela correspond également à ce que nous pouvons observer dans notre environnement également. De plus, la dépendance des paramètres est également cohérente avec la littérature car nous savons que la confiance est héritée entre les générations. Par conséquent, nous savons que les conditions initiales affectent les résultats économiques d'un pays.

Notre article contribue à la littérature de 2 manières. Premièrement, il s'agit du premier article à créer un modèle coévolutif entre la confiance et les préférences réciproques à notre connaissance. Nous montrons qu'une augmentation de l'attitude de confiance n'augmente pas toujours les préférences réciproques. Ce résultat est important pour la littérature car il décrit une relation cyclique entre une confiance élevée et une faible fiabilité. Deuxièmement, nous créons un cadre pour utiliser des équations réplicatrices dans des jeux en deux étapes. Dans la littérature sur la dynamique des réplicateurs, avoir un jeu symétrique à une étape est presque toujours répandu. Notre approche montre que la théorie des jeux évolutifs peut également être utilisée pour étudier les interactions en deux étapes.

Une prochaine étape logique de cet article pourrait être l'étude de ce modèle dans un cadre de réseau social. De cette façon, un économiste peut comprendre comment le concept d'évolution de la confiance sociale affecte la fiabilité d'un réseau. De plus, cela pourrait s'avérer utile lorsqu'on étudie la formation d'un réseau social. Nous fournissons également des seuils pour les paramètres à utiliser dans une étude expérimentale. Ces paramètres peuvent être testés dans une expérience pour voir si les comportements des agents vont changer de manière significative. De plus, notre article pourrait fournir un cadre pour un jeu de mille-pattes coévolutif asymétrique.

Les concepts de confiance et de fiabilité sont naturellement liés. Cependant, la corrélation entre eux ne doit pas nécessairement être positive. Une confiance élevée dans une société peut également conduire un agent malveillant à réaliser un profit plus élevé. Nous suggérons que ce phénomène pourrait être la raison pour laquelle la méfiance est une réponse naturelle à de nombreuses occasions.

Özet

Güven, toplumlar için çok önemli bir kavramdır. Birçok sosyal etkileşim bir dereceye kadar güven gerektirir. Bu güven genellikle rakibin sosyal normlara uyma olasılığına duyulan güvendir. Bu olasılığı güvenilirlik olarak tanımlıyoruz. Güvenilir bir kişi, kendisi ve ona güvenen kişi arasındaki sözleşmenin kurallarına uyar. Daha da ilginci, çok olmasa da bazı kimseler, kendi faydalarına olsa da güvenen kişinin intikamını alma şansı olmasa bile bu kişiye ihanet etmezler. İnsanların materyalist olduğu ve sadece sadece kendi maddi getirisini düşündükleri varsayımı ile bu davranış açıklanamaz. Bu fenomeni göstermek için birçok ekonomist karşılıklılık davranışlarını, ajanların getirilerinin bir kısmından vazgeçerek güvenilir davranmalarına neden olan davranışları incelemektedirler. Güvenilirliğin nasıl ortaya çıktığını açıklamak için farklı karşılıklılık davranışları önerilmektedir.

Bu tez, evrimsel oyun teorisi çerçevesinde karşılıklılık davranışları arasındaki farklılıkları analiz etmektedir. Bu davranışlar, insanların sosyal etkileşimlerde güvenilir davranmasını sağlar. Aksi takdirde, klasik oyun teorisi, materyalist bir ajanın yalnızca kendisi için en iyi olanı oynayacağını ve başkalarının iyiliğini umursamayacağını öne sürer. Ancak durumun böyle olmadığını gözlemliyoruz. Karşılıklılık davranışların arkasındaki motivasyonlar bir tartışma konusudur ; ancak bazı insanların başkalarının iyiliği için çalıştığını iddia etmek kolaydır. Örneğin, bazı ebeveynlerin davranışlarını gelecekte beklenen bir getiri ile açıklayamayız ; üstelik toplumun gelişmesi için çaba sarf eden kar amacı gütmeyen kuruluşlarda çalışan gönüllüler de var. Bu davranışlar materyalist olarak kabul edilemez. Tam bir fedakarlık olmasa da, en azından bir tür sorumlulukla motive olurlar.

Oyun teorisi literatüründe, ekonomistler birçok türde karşılıklılık davranışı tanımladılar. Bazı deneyler, güvenilirliğin koşulsuz olarak diğer kişileri önceleme tercihinden gelebileceğini öne sürüyor. Buna karşılık, diğerleri bunun karşı tarafın güven davranışına bir dereceye kadar tepki vermesinden kaynaklanabileceğini öne sürüyor. Cox et al. (2016), güvenilirliğin arkasındaki ana itici gücün, güvenen kişinin kendini karşısındakinin ihanet etmesine açık hale getirmesinden ve güvenilir olan kişinin karşısındakinin bu zaafiyetine duyarlı olmasından kaynaklanabileceğini belirtmektedir. Zaafiyete karşı duyarlılık, bir başkasına güvenerek kendini ona karşı savunmasız hale getirdiğinde ortaya çıkan bir olgudur. Bu

çalışmamızda Cox et al. (2016)'nın deneysel çalışmasından yola çıkarak zaafiyete karşı gösterilen duyarlılığın saf özgecilik ve bencilliğe karşı nasıl performans gösterdiğini evrimsel bir bağlamda inceliyoruz.

Ünlü yatırım oyununa benzer 2 adımlı bir oyun oluşturuyoruz. Yatırım oyununda birinci oyuncu elindeki paranın tamamını ikinci oyuncuya gönderip göndermemeye karar verir. İkinci oyuncuya para 3 ile çarpılarak ulaşır. Ardından, ikinci oyuncu elde ettiği paranın yarısını geri gönderip göndermemeye karar verir. Sosyal olarak en uygun senaryo her iki oyuncu için de göndermeyi seçmekse de, ikinci oyuncu için en iyi strateji, herhangi bir kazanç sağlamadığı için hiçbir şey göndermemektir. Ancak deneyler, olumlu miktarda katılımcının parayı geri göndermeyi seçtiğini gözlemlemektedir. Bu nedenle, kişilerin parayı geri göndermeleri için bir motivasyon olmalıdır. Makalemizde, ilk oyuncunun işbirliği yapmak ve yapmamak arasında seçim yaptığı iki aşamalı bir mahkum ikilemi oyunu oluşturuyoruz. İkinci oyuncu bunun yerine koşullu işbirliği, koşulsuz işbirliği ve koşulsuz ihanet arasında seçim yapar. İlk oyuncu her zaman güven düzeyine bağlı olarak bir olasılıkla işbirliği yapmayı seçer. Ancak ikinci oyuncuya üç strateji atarız : (1) diğerkamlık; (2) zaafiyete yanıt verme; ve (3) bencillik. Diğerkam ajanlar, rakiplerinin getirilerinden pozitif etkilenirler. Öte yandan zaafiyete duyarlı ajanlar rakipleri onlara güvendiği için rakiplerinin zarar görmesine izin vermemeye özen gösterirler. Son olarak, bencil ajanlar sadece kendi faydalarını düşünürler. Bu oyun, hangi koşullar altında diğerkamlık veya zaafiyete duyarlılığının saf bencil eylemlere üstün geldiğini anlamamızı sağlar.

Bu stratejilerin zaman içindeki genetik uygumluluğunu analiz etmek için evrimsel oyun teorisi çerçevesini kullanıyoruz. En iyi genetik uyuma sahip stratejiyi bulmak için yaygın olarak ”çoğaltıcı denklemler” olarak adlandırılan bir dizi diferansiyel denklem oluşturulur ve çözülür. Çoğalıcı denklemler, her stratejinin tüm stratejilerin ortalamasına karşı nasıl performans gösterdiğini bularak çalışır. Ortalamadan daha iyi fayda getiren bir strateji zamanla artarken, daha kötü performans gösteren bir strateji azalır. Bu denklemlerin sıfıra eşit olduğu yani stratejilerin zamana bağlı artış veya azalış göstermediği durumda bulunan çözümler denge noktalarını verir. Bu denge noktalarının durağan olup olmadığını belirlemek için, bu denge noktalarından hangilerinin Jacobian matrisinin özdeğerlerinin negatif olduğunu buluruz. Eğer bir denge noktası için bulunan özdeğerlerin tamamının reel kısımları negatifse bu o denge noktasının durağan olduğunu gösterir. Ancak, tüm oyunların bir dizi durağan stratejiye sahip olması gerekmez. Bunun yerine, şahin ve güvercin gibi bazı oyunlar, av-yırtıcı ilişkisine benzer sonsuz bir döngüye sahip olabilir. Modelimizde, iki stratejinin bir arada bulunması veya bir stratejinin tam hakimiyeti ile sonuçlanabilecek çoklu dengelerin olduğunu gösteriyoruz.

Bir sonraki bölümde, sayısal bir analiz yapıyoruz. Bunun için parametrelere değer atayarak değişkenlere karşılık gelen değerleri buluyoruz. Bu analiz, modelimizin farklı başlangıç koşulları altında nasıl davrandığını anlamamızı ve her bir stratejinin çeşitli parametrelerdeki bir farklılığa nasıl tepki verdiğini anlamamızı sağlar. Ayrıca, analitik modelin sonuçlarını görselleştirmek için parametrelere değer vermemiz gereklidir. Modelimizin üç stratejisi olduğundan, sonuçlarımızı çizmek için üç boyutlu bir alana ihtiyacımız var ; bununla birlikte, 3B uzaydaki bir çizim genellikle şüpheli ve gereksiz olarak kabul edilir. Bu nedenle, 2-simplex kullanıyoruz ve sonuçlarımızı kartezyen koordinatlardan barycentric koordinatlara dönüştürüyoruz. Bu sayede faz diyagramlarımızı daha anlaşılır iki boyutlu bir uzayda gösterebiliyoruz. Bu faz diyagramları, sayısal sonuçlarımızı çizmemize izin verir.

Son olarak, ajan temelli bir model oluşturuyoruz. Ajan temelli modeller, basit bir dizi kurala dayalı olarak birden çok yinelemeyle çalışan hesaplama modelleridir. Ajan temelli modelimizi oluşturma adımlarımız şunlardır : İlk olarak ajanlar oluşturulur. Ardından, her ajan iki aşamalı mahkumun ikilemi oyununu oynamak için her turda rastgele eşleşir. Ajanlardan biri rastgele ilk oynamak için seçilir. İlk oyuncu rakibine güvenip güvenmemeye karar verir. Ardından, ikinci oyuncu en iyi tepkisini oynar. Örneğin, özgeci bir kişi için en iyi tepki, koşulsuz olarak işbirliği yapmaktır, bencil bir ajan için ise her zaman koşulsuz olarak ihanet etmektir. Karşılıklı oyun sona erdikten sonra, her ajan rastgele bir ajana bakar ve iki ajan arasındaki getiri farkına dayalı bir olasılıkla onu taklit etmeye karar verir. Böyle bir hesaplama yöntemi kullanarak çoğaltıcı dinamiklerini taklit edebiliyoruz. Ayrıca, ajan temelli bir model, çoğaltıcı denklemlerden daha uygun şekilde stokastik olabilir. Bu farklılıklar nedeniyle, ajan temelli bir modelde denklem temelli bir modelden farklı sonuçlara sahip olması çok nadir değildir.

Ajan temelli bir modelde, stokastikliğin modeli nasıl etkilediğini görmek için birden çok simülasyon çalıştırmak esastır. Bununla birlikte, modelin dinamiklerini anlamak için her bir koşulu incelemek eşit derecede önemlidir. Simülasyona özel ayrıntıların kaybolmasını önlemek için birden çok simülasyonu ayrı ayrı dikkatlice analiz ederiz. Ayrıca, ajan temelli modelin ortalama dinamiklerini de incelenmektedir. Bunu yapmak için simülasyonu, çalıştırmalar arasında değişen ve her simülasyonun ortalamasını alan bir dizi parametre altında çalıştırırız. Ayrıca, stokastikliği hesaba katmak için aynı parametre kombinasyonu için simülasyonu birden çok kez çalıştırırız. Bu sayede hem ortalama hem de bireysel sonuçları göstermek mümkün olmaktadır. Ortalama sonuçlar, modelin genel yönü hakkında bilgi sağlarken, bireysel sonuçlar bir simülasyonun zaman içinde nasıl geliştiğini gösterir.

Her üç bölümdeki sonuçlarımız çoklu dengeleri gösterir. Parametrelere bağlı olarak, bazı durumlarda bencil olmak faydalı olabilirken, bencil ajanlar ve fedakar ajanlar da bir arada var olabilir. Bu birlikteliği tüm analitik, sayısal ve ajan temelli modellerde gözlemleyebiliriz. Aynı zamanda çevremizde gözlemleyebildiklerimizle de tutarlıdır. Ayrıca, güvenin nesiller arasında miras kaldığını bildiğimiz için parametre bağımlılığı literatürle de uyumludur. Dolayısıyla başlangıç koşullarının bir ülkenin ekonomik sonuçlarını etkilediğini biliyoruz.

Makalemiz literatüre 2 şekilde katkıda bulunmaktadır. Birincisi, bu, bildiğimiz kadarıyla güven ve karşılıklı tercihler arasında ortak evrimsel bir model yaratan ilk makaledir. Güven tutumundaki bir artışın karşılıklılık tercihlerini her zaman artırmayabileceğini gösteriyoruz. Bu sonuç, yüksek güven ve düşük güvenilirlik arasındaki döngüsel bir ilişkiyi tanımladığı için literatür için önemlidir. İkinci olarak, iki aşamalı oyunlarda çoğalıcı denklemleri kullanmak için bir çerçeve oluşturuyoruz. Çoğalıcı dinamiği literatüründe, neredeyse her zaman simetrik tek aşamalı bir oyun kullanılmaktadır. Yaklaşımımız, evrimsel oyun teorisinin iki aşamalı etkileşimleri incelemek için de kullanılabileceğini gösteriyor.

Bu makaleden sonraki mantıklı bir adım, bu modelin bir sosyal ağ çerçevesinde incelenmesi olabilir. Bu şekilde, bir ekonomist sosyal güvenin evrimi kavramının bir ağın güvenilirliğini nasıl etkilediğini anlayabilir. Ayrıca, bir sosyal ağın oluşumunu incelerken bu yararlı olabilir. Ayrıca deneysel bir çalışmada kullanılacak parametreler için eşikler de sağlıyoruz. Bu parametreler, ajanların davranışlarının önemli ölçüde değişip değişmeyeceğini görmek için bir deneyde test edilebilir. Ayrıca, makalemiz asimetrik olan ortak evrimsel bir kırkayak oyunu için bir çerçeve sağlayabilir.

Güven ve güvenilirlik kavramları doğal olarak iç içedir. Ancak, aralarındaki korelasyonun pozitif olması gerekmez. Bir toplumdaki yüksek güven, kötü niyetli bir ajanın daha yüksek kâr elde etmesine de yol açabilir. Bu fenomenin, güvensizliğin birçok durumda doğal bir tepki olmasının nedeni olabileceğini düşünüyoruz.

1 INTRODUCTION

Trust lies in the very heart of the community. Not only it does allow one to have a healthy environment, but also it creates a foundation for every economic interaction. Lending money, paying taxes, and even accepting fiat money are all based on the trust that others will follow the social norms and reciprocate in the future. Two companies not trusting each other will be very reluctant to work together. It is a safe assumption that a community where no one trusts each other will be pruned to elimination.

Let us consider a basic Prisoner's Dilemma game where betraying yields the highest payoff. If this is the case, how is it not the prevailing behavior from evolution? Also, what prevents us from betraying each other? To understand how trust can emerge in a game theoretical framework, Axelrod and Hamilton (1981) organized a competition of different models. In these models, their students tried to come up with a model that allows cooperation to win against defection. One of the winning strategies was the tit-for-tat strategy, where players would maximize their utilities by considering the long-term benefits of cooperating while punishing the defectors. However, this would indicate that every player is highly rational and strategical. This assumption would eliminate the cases where one could act with complete altruism. If altruism exists in real life, how can it survive? One of the explanations for this question is given by Becker (1976), where he has shown that altruism can survive in the evolution in an analytical framework. Therefore, we can assume that both altruism and selfish rationality can exist in a society of cooperation and coordination. In this context, *trust* can be defined as the belief that other parties will comply with social norms.

In this paper, we study the dynamics behind trust and reciprocity. Trust is modeled as the belief that others will cooperate in an interaction, while reciprocity is someone's response to others' cooperation decisions. We modeled a two-stage sequential prisoner's dilemma game to show how trust and reciprocal behaviors affect each other's evolution. Moreover, we assumed three types of reciprocal behaviors; altruism, vulnerability-responsiveness, and selfishness. Altruism indicates that an individual gets a positive payoff from her opponent's payoff. Vulnerability-responsive agents feel obligated to reciprocate if their opponents become vulnerable against them by cooperating, while selfish agents only care about their personal payoffs.

The first mover's decision on the game depends on their trust level. They decide to cooperate blindly or defect blindly; blindly means they take their decision without having any information about their opponent's type. For the second movers, there are three strategies; cooperate unconditionally, cooperate conditionally, and defect unconditionally. Unconditionally implies that these actions are taken without considering the opponent's action; however, cooperate conditionally indicates these agents only cooperate if their opponents have cooperated.

Trust and trustworthiness seem similar at the beginning. However, World Values Survey results show how different these two concepts are. This survey has been conducted for more than 40 years and is one of the most established surveys. Moreover, WVS conducts surveys with thousands of individuals from each country. Trust is measured with the question: "Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?" This question is considered accepted by the literature (Alesina & La Ferrara, 2002), even though there are papers that discuss its validity (Glaeser et al., 2000).

The question of trustworthiness is not so simple. Slemrod and Katuščák (2005) uses the question:" Please tell me whether you think lying in your own interest can always be justified, never be justified, or something in between." as an indicator of trustworthiness. Even though this question is not a generally accepted measurement of trustworthiness, we can say it is a valid candidate. However, this question restricts us to using the second wave of the World Values Survey, as it has not been asked later.

Table 1.1 shows how different trust and trustworthiness behave on average. We can observe that people mostly claim that there is a necessity to be careful when interacting with other people; however, they do not justify lying overall. Therefore, we can safely suggest that the behavior difference between first-movers and second-movers in our model can be justified using empirical data.

	n	mean	sd	min	max
Trust	22930	0.30	0.46	0	1
Trustworthiness	22125	8.13	2.54	1	10

Table 1.1: Trust and trustworthiness in the second wave of World Values Survey.

We analyze our model first with a population-based dynamic by differential equations to describe the global changes over time. The replicator dynamics for the reciprocal behaviors' evolution and an update function for trust change are used to define the equation-based model. Then, we adopted an agent-based model for the robustness check and received approximate results.

The population-based model (PBM) has some advantages and disadvantages compared to agent-based modeling (ABM). In the population-based dynamic, defining each individual's characteristics is not possible. Therefore, we use representative agents to define the model's characteristics. In our model, we use the mean trust level of the population and examine its change over time instead of assigning a different trust level to each agent. Moreover, the reciprocity behaviors are also presented by the mean level, e.g., altruists' altruism level and vulnerability responsive agents' response level, and update speeds. ABMs are considered more realistic since they allow the modeling of complex systems by varying the attributes of agents Bosse et al. (2012). Additionally, determining the results for micro-level interaction is possible with ABM, while the PBMs give the results for the global level. For example, even if the population mean is the same as PBM for trust, we can analyze how having a segregated population with very low and high trust affects the other dynamics or the speed of the population reaching the steady state. On the other hand, population-based dynamics also have some advantages over agent-based modeling; for example, simulation is much faster than ABMs, since the ABM has to run for many agents to receive similar results.

The rest of the paper is structured as follows; Chapter 2 reviews the literature on social capital, trust, reciprocity, population-based, and agent-based dynamics. Chapter 3 describes the analytical model by giving assumptions, attributes of agents, and the game setting. Then, we analyze the steady-state points of the model and examine the local stability of these equilibrium points. Chapter 4 presents the model's numerical simulation for the population-based model. Afterward, we will adapt our theoretical model to an agent-based model to check if they are consistent. Finally, in chapter 5, we will conclude our study with suggestions for future works using this framework.

2 LITERATURE REVIEW

Different countries' development and growth patterns have been studied for a long time. The disparity in economic performance is not fully understood even today. The mainstream growth models such as Solow and Walras models focus on standard factors of production (human capital, physical capital, and technology). However, these theories are criticized for not explaining overall reasons for development, differentiation across similar societies with similar production factors, and unwanted outcomes such as inequality, social conflicts, and negative externality on the environment. As a result of numerous critiques, new economic theories embedding socio-cultural factors have been enhanced (Bhandari & Yasunobu, 2009). Eventually, social capital became one of the most critical elements of economic models.

The root of social capital literature can be traced back to the pioneers of the concept Bourdieu (1986) and Coleman (1988). Moreover, the famous book by R. D. Putnam et al. (2000), "Bowling Alone: The Collapse and Revival of American Community", drew attention to the subject and created the popular definition of social capital currently used (Poder, 2011): "a collection of networks, norms, and social trust that promote coordination and cooperation for common interests" (R. D. Putnam et al., 2000). Social capital is criticized for not having the features of classical capital (Bhandari & Yasunobu, 2009). However, Robison et al. (2002) object to these arguments by showing its capital-like characteristics, such as its capability to create other capital forms, to be invested for creating revenue, worn by usage, require investment for maintenance, used for transformation of outputs in a production function based on its performance and duration, flexibility, substitutability. After Putnam brought into focus that social capital is a critical asset for societies to work effectively and showed how it declined over time in the US, social scientists and policymakers started giving much more attention to determining the components and the ways to increase it. Besides the benefits of social capital on economic performance, its effects on education, health, happiness, well-being, life satisfaction, inequality, and crime have been thoroughly analyzed by different fields [1].

The assessment of social capital is crucial to promote the driven effects and prevent depreciation, but it is problematic because of its tangled structure, and there is no consensus about it. Studies differ according to the proxy they used; for example, some empirical studies only use network density to evaluate social capital, some assess only trust measures, while others combine network density and proxies for relevant norms (Krishna & Shrader, 1999). Survey data on trust, cooperation, participation, and social interaction is used as a proxy to determine the change of social capital and the comparison across countries, as pointed out in Alesina and Giuliano (2015). For example, R. Putnam (2001) used the data covering social interaction and participation in US organizations such as clubs, churches, and associations to see the change in civic engagement.

Game-theoretical approach has also adopted social capital to explain over-cooperation instead of predicting a rational individual's optimal choice. Paldam (2000) described the social capital in game theory as people's excess desire for cooperation in prisoner's dilemma game. As we mentioned before, because of the complex structure of social capital, there are also multiple approaches in game theory to handling social capital. In this study, we integrate two essential elements of social capital: trust and reciprocity. Since these two factors are critical determinants of economic behaviour and change due to economic interaction, understanding how they influence each other is crucial.

2.1 Trust and reciprocal preferences

Social trust represents the belief of an individual to the trustworthiness of the other members of the society (Zak & Knack, 2001). The impact of social trust has been found crucial as the prospects of cooperation are perceived higher by trusting in-

[1]For detailed literature on social capital and health, see Ehsan et al., n.d.; for social capital and education, see Mishra, 2020; for social capital and SME, see Gamage et al., 2020.

dividuals, and they happen to be less concerned with exploitation (Rathbun, 2011)[2]. Trusting individuals are more likely to engage in costly cooperative actions and declare to be more satisfied with their life (Sønderskov, 2011). High trust participants in Uslaner (2002) were more likely to empathize with others, and this attitude allows them to build more expansive moral communities and create cooperative bonds across racial and ethnic boundaries (Rathbun, 2011).

Social trust affects information transmission, cooperation, and the enforcement of sanctions within a society (Hagen and Choe (1998); Kwon and Arenius (2010); Leonardi et al. (2001)). Fukuyama (2001) proposed two dimensions of social trust: level of trust (the strength of cooperative norms) and radius of trust (the circle of people among whom cooperative norms are operative). Living in a small village where everyone knows each other is not the same experience as living in a big metropolitan city where daily interactions are mostly between individuals who have the option of not seeing the other part for the rest of their lives. Therefore, average trust values are much higher in rural areas (Hardin, 1999).

Reciprocity is the manifestation of social fairness. The fairness model of Rabin (1993) is the pioneer in introducing reciprocity preference to the game theory literature (Isoni & Sugden, 2019). In this literature, *reciprocity* is defined as an agent's incentive to respond to kindness (unkindness) with reward (punishment). The investment game of Berg et al. (1995) provides an experimental setup to observe trust and reciprocity. In this two-player game, player A can send any amount out of his money to player B who will triple this amount upon reception. Then player B, in return will decide how much of this tripled amount to send back to player A. The Nash equilibrium is player A sending 0 to player B, but in the experiments both players send more on average. Charness and Rabin (2002) bring a new perspective to reciprocity by designing games to examine the driven effects of social preferences and shows that subjects had more interest in increasing social welfare by reducing the payoff difference. The reciprocity of subjects is observed by the punishment of unfair action to others or unwillingness to sacrifice when others do not sacrifice.

[2]For trust and economic activity, see Guiso et al. (2006), Knack and Keefer (1997) and La Porta et al. (1997).

Even if reciprocity is mainly studied in a game-theoretical framework, studies based on surveys also exist. For example, Dohmen et al. (2008) examined the determinants of trust and reciprocity by using “German Socio-Economic Panel” data. In this study, reciprocity is classified as positive and negative reciprocity. Positive reciprocity indicates the agent’s incentive to respond with kindness towards kind actions. Conversely, negative reciprocity shows a person’s willingness to reciprocate mistreatment when she is mistreated. They concluded that trust and positive reciprocity are weakly correlated, yet trust and negative reciprocity are negatively correlated. They also showed that personality traits have an impact on trust and reciprocity. Guzmán et al. (2020) provides a game-theoretic model for trust and reciprocity by combining personality attributes, tests the model by doing a survey and compares the results through experiments. Their findings are consistent with Dohmen et al. (2008), personality affects agents’ reciprocity preferences and trust bias.

Altruism is another critical driven effect of cooperative behavior. Fehr and Fischbacher (2003) explain how people become distinct from the other animals in cooperative behavior and what are the supportive factor of this phenomenon. Even if most commonly assumed in economic theories that agents are selfish and choose the option that maximizes their payoffs, experimental studies show that agents also consider the other agents’ payoffs besides their own. One of the most famous examples to this phenomenon is the behavior of some parents, where they sacrifice a portion (if not all) of their utility for their children (Becker, 1976). Even though Simon (1993) suggests that altruism may be enforced through social norms, Fehr and Schmidt (2006) suggest that it is safe to assume that other-regarding preferences exist to some degree. Levine (1998) claims that agents pay regard to their opponents’ altruism or spitefulness while taking action for public good and ultimatum bargaining games. In other words, agents consider how their opponents treat others besides how they are treated. People have less willing to be kind to a hostile person, while they are more prepared to act kindly to a benevolent person.

Sethi and Somanathan (2001) creates an evolutionary model for reciprocity, described as an element of interdependent preferences. They encountered that the materialist cannot survive in evolutionary dynamics if a population has reciprocal preferences. Danielson (2002) creates evolutionary agent-based models to test Levine (1998) and Sethi and Somanathan (2001)'s predictions for different sequential games. Even though he found out that the results are consistent with their conclusions about the existence of reciprocity is essential to stabilize cooperation, reciprocation does not create optimal results necessarily. This is because the existence of enough spiteful agents would cause reciprocal agents to not cooperate; thus, reducing overall payoff. [3]

Sobel (2005) argues that reciprocity can be seen as an intrinsic behavior in opposition to traditionally assumed that it is a reputation-creation mechanism for selfish agents. Carrasco et al. (2018) adds reciprocal preferences to Levine's model, and they differentiate the intrinsic and behavioral preferences. They showed that agents could act selfishly even if they are intrinsically altruists and vice versa. From this point of view, we assume that agents' intrinsic preferences change over time depending on their experience.

Investment game, established by Berg et al. (1995), requires an agent to send an amount of money to the opposite player, who may choose to send an amount back or not. Therefore, first player has to *trust* her opponent, since there will not be a third round. Here, the behavior of the second agent is defined as *trustworthiness*. Trust is measured both empirically and experimentally (see Glaeser et al. (2000) for a comparison of two approaches). On the other hand, trustworthiness usually is measured solely through experiments, Cox et al. (2016) being the most related to this paper. They show that vulnerability-responsiveness is a solid drive in trustworthiness; that is, an agent may feel responsible to another agent because they showed a vulnerability to them. Since we assume that an altruist agent may benefit from the first-mover opponent's utility, our altruist agents show vulnerability-responsiveness in the sense that they are afraid of punishing someone for trusting them. They feel rewarded when they do not punish their opponents for showing vulnerability.

2.2 Social networks

Social networks are the building block of social capital. Studying social networks allows us to understand the behavioral differences between an individual and a community (Krause et al., 2007). There are many emergent behaviors caused by the structure

[3] For a detailed survey about the evolutionary dynamic of reciprocity, see Sethi and Somanathan (2003).

or even the existence of social networks. One important example can be the segregation model of Schelling (1971). Schelling has shown how agents with a significantly low amount of rules can create crucial macro-behaviors in this agent-based model.

Information sharing in a network allows members to have a reputation that forces them to act cooperatively. Lippert and Spagnolo (2011) describes the strength of a network with the capability of enforcing agreements where members share information, and they have the potential to punish agents who do not fit into society. Hardin (1999) tells a story of a fictional village where everybody has to act cooperatively to survive. Since everybody needs each other in this small village, everyone is afraid of being ostracized, preventing them from betraying others.

Other emergent results of social networks are reciprocity and altruism. Leider et al. (2009) prove that social networks allow for building a higher level of reciprocity. In a two waves game setting with players from Harvard's two students' dormitories, the authors conducted a network elicitation task to reveal degrees of friendships between students. Then, it is followed by an allocation game (a modified version of the dictator game in the first wave, the original helping game in the second wave), revealing the altruistic attitudes of students.

Social networks and trust have bilateral relationships. In other words, social networks allow trust formation, while trust facilitates building networks. An example of this is Di Cagno and Sciubba (2010) combination of network formation and trust formation model. They analyzed two different settings: (i) a trust game followed by a network generation game where players are anonymous; and (ii) a network generation game followed by a trust game where people know other people's fake identities throughout the whole game. These two settings allowed the authors to see how networks affect trust by creating a simple comparison of trust games where people know each other in one and do not in the other.

Networks are dynamic structures. It is impossible to think of a community where nothing changes over time. Since migration is a shock to society, the social capital transformation of immigrants and locals can be observed by comparing the host and home countries' dynamics. For example, Algan and Cahuc, 2010 use the trust information of immigrants in the US to calculate the relationship between trust and GDP.

They claim that immigrants carry their trust attitudes to the host country and inherit through generations. Dohmen et al. (2012) also finds proof of this assumption and shows a correlation between the trust of one generation and their children in Germany.

The literature mainly demonstrates that the more ethnically segregated communities have lower trust mean, and since immigration increases the segregation, it decreases the social capital level of the country [4]. However, Dinesen (2012) find that it is possible that the immigrants who came from low trust countries can exceed the mean trust level of original inhabitants. Lütz et al. (2021) use an evolutionary-game theoretical model to analyze the effect of immigration on cooperation in a society. Their model supports that immigrants can increase the level of cooperation between natives if there are altruistic immigrants as well as self-centered ones.

Moreover, the trust level of the immigrants is generally lower than the host countries' mean in the studies; however, a phenomenon of today's world is that well-educated people leave their countries and move to developed countries. Since there is a strong correlation between higher education and trust (Alesina & La Ferrara, 2002), highly educated individuals could have a higher trust level than their home countries' means, maybe even their host countries' means. In addition, trust change is studied as a single factor, yet the reciprocal behavior between immigrants and locals could cause a decrease in trust level. For instance, a local may not reciprocate to an immigrant because they can think that their reputation will not get hurt. (Douds & Wu, 2018) demonstrated that racial discrimination has a negative impact on the trust level, and racial segregation and trust are positively correlated. From this point, it is possible that immigration does not necessarily harm social capital only because immigrants' own actions and the host countries miss the opportunity to enrich their social and human capital. Therefore, we expand our model to show how reciprocal behaviors affect trust modification and vice versa, where individuals face or do not face a consequence of their non-cooperative actions.

[4] For a detailed meta-analysis about ethnic diversity and trust, see Dinesen et al. (2020).

2.3 Population-based dynamics and agent-based modelling

Aggregation is one of the most common forms of creating a population-based dynamic model. This models are based on averaging individual behaviors, and then estimating the change on the average behavior (Brown et al., 2016). These models give the opportunity to make an estimation easier compared to individual-based dynamics; especially when there are too many types of behaviors. It can be used for so many fields related to economics.

In economics, one of the introductory topics can be aggregate supply and aggregate demand topics. Here, an economist does not consider any company or any specific consumer; instead, the economist considers the economy as a whole. This is a very crucial part of the mainstream economics (Dutt, 2006). This is a very convenient example of population-based dynamics, as it does not consider each agent individually in calculations.

Evolutionary game theory allows for a number of applications of population-based dynamics including auction theory (Phelps et al., 2005). These models could be defined as computation-based population dynamics. In these models, an economist use computational simulations to derive results about parameters or macro results. However, agents are not individual beings; instead, the populations strategically interact with each other in the computation.

Maybe the most important application of population-based dynamics is related to replicator equations. Replicator equations are considered as population-based dynamics, even though they still include different types of agents. This is because it is possible to categorize strategies in replicator equations as a population of many individuals. These models were not only extremely useful in social sciences; but they have also been in use in computational sciences as well. For example, Hennes et al. (2019) use replicator equations to develop a neural network in computational sciences. Moreover, Takeuchi and Hogeweg (2012) use these models to study a topic relevant to life sciences as well. Therefore, it is safe to say that these models are quite crucial. Replicator equations can be simulated analytically or numerically. Numerical solutions tend to rely on computational iterations, while analytical solutions do not.

Agent-based models work by creating a set of rules for each type of agent. Then, a computational simulation is run by allowing each agent to act each turn using the previously defined sets of rules. Therefore, one can find macro results of micro-behaviors (Schelling, 1971). These micro-behaviors can be as simple as ”move until you are happy” (where happiness can be a simple parameter); while they can be as complex as creating a fully developed machine-learning algorithm for stock trading as well. Then, the modeler studies the results created by the interactions of different agents who either have the same or different sets of rules.

Compared to population-based dynamics, agent-based models are better at capturing the important changes caused by various types (Bosse et al., 2012). They may also be easier in a sense that they may require less analytical expressions, since they rely more on the computations. However, when there are a lot of different agent-types; population-based dynamics could be preferred, especially if the modeler knows the average behavior in the population (or at least in a certain category). Nonetheless, assistance of computations create a possibility for an economist to structure models that cannot be predicted otherwise (Helbing, 2012).

3 THE MODEL

3.1 Economic environment

We have a population of agents denoted by $N = \{1, 2, ..., n\}$, at $t = 0$ where t is discrete time with $t = 0, 1, ...$ All agents are matched randomly in pairs and assigned as first and second player each time. Then, they play a sequential prisoner's dilemma game by choosing to cooperate or defect. We use the game theoretical setup provided by Guzmán et al. (2020) where the first player chooses either to "cooperate blindly (CB)" or "defect blindly (DB)". Without knowing the first player's action, the second player decides to "cooperate unconditionally (CU)", "cooperate conditionally (CC)", or "defect unconditionally (DU)". By unconditionally, we mean that the player chooses to cooperate or defect without caring about the opponent's action and by conditionally, we mean that the player will choose to cooperate only if her opponent cooperates. The extensive form of the game[5] is given in Figure 3.1.

For a player i at time t, $\pi_{i,t}(s_i, s_j)$ is her material payoff from the game where (s_i, s_j) is the outcome of the game round with the action taken by player i, s_i and the action taken by her opponent player j, s_j.

In our economy, each agent $i \in N$ has two characteristics: i. ***trust*** denoted as $b_i \in [0, 1]$ reflecting the subjective probabilistic belief of agent i that other agents will cooperate. ii. ***trustworthiness*** indicating the motivation of agent i's response to her opponent's trust. Based on the terminology in Cox et al. (2016), we assume that there are three types of agents: *altruists* having unconditional other regarding preferences, *vulnerability-responsive* agents who are afraid of punishing people for trusting them, and feel rewarded when they do not punish their opponents for showing vulnerability

[5] The payoffs satisfy $\pi_{dc} > \pi_{cc} > \pi_{dd} > \pi_{cd}$.

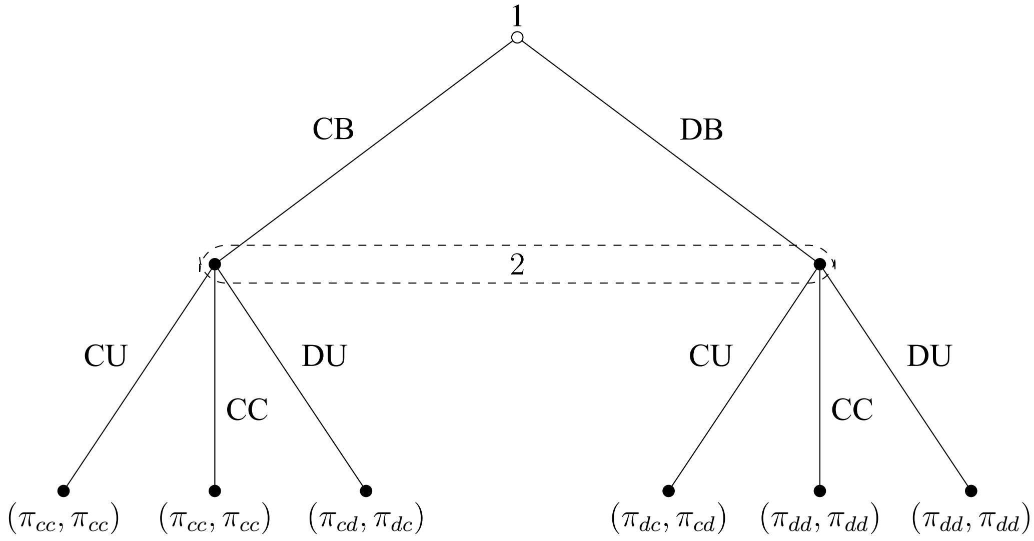


Figure 3.1: The sequential prisoners' dilemma

and *selfish* agents interested only on their material payoffs. We will denote these types by $r \in \{a, s, v\}$ where a stands for altruistic agents, s for selfish agents and v for vulnerability-responsive agents.

Combining the characteristics of the agents that we have mentioned above, we assume that the first player will select her strategy based on her trust $b_{i,t}$, which is the co-operation probability of opponents, for agent i at time t. Then, according to her intrinsic preference coefficient $\theta_i^r \in [0, 1]$, the second player will choose her strategy. Agents have different perspectives when they are in the first and second mover positions. We suggest that an agent may not trust her opponent even though she is trustworthy. For example, Glaeser et al. (2000) found that trustworthy agents do not have to trust their opponents. Therefore, an altruist agent, even though she always acts trustworthy in the second turn, may not trust her opponent to do the same. Moreover, a spiteful agent can always defect in the second turn as it makes the most profit; however, if she knew that her opponent is a reciprocator when she is the first-mover, she would have to play cooperate in the first turn to get the highest payoff. In conclusion, we suggest that there are enough reasons for agents to play different strategies in two different positions.

For the strategy profile (s_i, s_j), these intrinsic preferences can be represented by the following utility function:

$$u_{i,t}(s_i, s_j) = \pi_{i,t}(s_i, s_j) + \theta_i^r \pi_{j,t}(s_j, s_i) \tag{1}$$

with $\theta_i^r \in [0,1]$. Obviously θ_i^r will be 0 for a selfish agent and $\theta_i^r > 0$ for an altruist agent. Without loss of generality, we will take $\theta_i^a = 1$. For vulnerability-responsive agents, there are fairness concerns and the utility will be responsive to the harm they can cause by defecting, therefore we suppose that their preferences may be represented by

$$\begin{aligned} u_{i,t}(s_i, s_j) &= \pi_{i,t}(s_i, s_j) + \theta_i^v \left(\pi_{j,t}(s_i, s_j) - \mathbb{E}[\pi_{j,t}(s_i, s_j)]\right)\big|_{s_j=C} \\ &= \pi_{i,t}(s_i, s_j) + \theta_i^v \left(\pi_{j,t}(s_i, s_j) - \mathbb{E}b_t \pi_{cc} - (1 - \mathbb{E}b_t)\pi_{cd}\right) \end{aligned} \tag{2}$$

where θ_i^v is the degree of responsiveness of an agent i to the opponent's vulnerability and $\mathbb{E}b_t$ is the mean trust level. We will suppose that the degree of responsiveness to opponent's utility θ_i^r will be the same for each type of agent and every agent i.e. $\theta_i^r = \theta^r$.

The vulnerability-responsive agents have other-regarding preferences, and this concern for others' well-being is conditional on the impact that they can have on others. This impact can be positive if the agent is being cooperative in response to the opponent's vulnerability and negative if the agent is betraying the opponent's vulnerability. Both of these effects force a vulnerability-responsive agent to cooperate in return for a cooperative act. In other words, responsiveness is in question for a vulnerability-responsive agent i if and only if the opponent j cooperated first i.e. $\theta_i^v > 0$ if and only if $s_j = C$. In equation 2, the difference between a cooperative opponent's realized payoff $\pi_{j,t}(s_i, s_j)$ and the expected payoff $\mathbb{E}[\pi_{j,t}(s_i, s_j)]$ measures the impact an agent i can have on a cooperative first mover. The expected payoff is determined by the mean cooperative action and the payoffs associated to the outcome of the game.

3.2 Equilibrium with rational agents

The first mover has imperfect information because the type of the second mover is private information. Let $p^r_{i,t}$ be the probability assigned by the agent i to being matched with an opponent of type r. As the first players choose their strategy according to their belief that others will cooperate, they choose to cooperate blindly (CB) with probability $1 - p^s_{i,t}$, and defect blindly (DB) with probability of $p^s_{i,t}$. As far as the decision of second movers is concerned, each type will choose an alternative based on the expected cooperative outcome. For selfish agents, DU is a dominant strategy and for the other types, the choice of strategy is conditional on their degree of responsiveness and payoffs of the game. We can give the following result where altruistic agents will choose CU, and vulnerability-responsive agents CC.

Proposition 1. *Selfish agents will play DU at the equilibrium and it is strictly dominant for altruist agents to play CU, vulnerability-responsive agents CC if and only if* $\frac{\pi_{dc}-\pi_{cc}}{\pi_{cc}-\pi_{cd}} < \theta^v < 1$, $\theta^a > \frac{\pi_{dd}-\pi_{cd}}{\pi_{dc}-\pi_{dd}}$ *and* $\theta^a > \frac{\pi_{dc}-\pi_{cc}}{\pi_{cc}-\pi_{cd}}$. *For the first mover, if* $p^v_{i,t} > p^a_{i,t}\frac{-\pi_{cc}+\pi_{dc}}{\pi_{cc}-\pi_{dd}} + p^s_{i,t}\frac{-\pi_{cd}+\pi_{dd}}{\pi_{cc}-\pi_{dd}}$, *then CB is the best response and otherwise the agent will choose DB.*

The choice of strategies for first and second movers is given in detail in Appendix A. We see that for higher degrees of responsiveness, vulnerability-responsive agents would cooperate conditionally i.e. $\theta^v \in [\frac{\pi_{dc}-\pi_{cc}}{\pi_{cc}-\pi_{cd}}, 1]$.

3.3 Evolution of attitudes and beliefs

3.3.1 Formation of trust

We have defined that trust is a probabilistic belief and if agents do not have complete information on their environment, we suppose that they are forming this belief based on their intuition or personal observations. Therefore, this probability is not constant and is updated frequently with new information. This belief or opinion may be calculated out of the positive and negative experiences.

The first players choose their strategies based on their trust. At the end of each period, they will observe the actual response of their opponent and may time to time update their belief. Based on their observations, they can update their belief based the following rule. First, they calculate their average payoffs for k consecutive plays and if the average is greater than the payoff of the non-cooperative outcome π_{dd}, they will increase their subjective probability by the adjustment parameter $\alpha > 0$ which is common for all players. On the contrary, if the average is less than π_{dd}, their belief that others are likely to cooperate will decrease by the same adjustment parameter α. Formally, trust $b_{i,t+1}$, the belief of agent i that other agents are likely to cooperate at time $t+1$ based on previous period observations of payoffs changes as follows:

$$\frac{b_{i,t+1} - b_{i,t}}{b_{i,t}} = \alpha \left(\frac{\sum_{l=1}^{k} u_{i,t-l+1}}{k+1} - \pi_{dd} \right) \tag{3}$$

This is a general form of the adjustment dynamics where agent i updates her belief based on the average of the observations for k periods. Since trust is a probabilistic belief in our model, it takes a value between 0 and 1. We will first suppose that the agent uses only current period observation to make adjustments[6].

$$\frac{b_{i,t+1} - b_{i,t}}{b_{i,t}} = \alpha \left(u_{i,t} - \pi_{dd} \right) \tag{4}$$

This will result in the following continuous time dynamics:

$$\frac{db_i}{dt} = \dot{b_i} = \alpha b_{i,t} \left(u_{i,t} - \pi_{dd} \right) \tag{5}$$

3.3.2 Evolution of reciprocal attitude

Proposition 1 states that selfish rational agents will defect each period and the cooperative outcome will not be observed at the equilibrium. However, in real life, we observe a diversity of behaviors and unconditional and conditional other regarding preferences are some of these. Here, instead of supposing that agents are rational and maximise their payoff, we will suppose that initially, agents in the population have different other regarding attitudes summarised by utility functions given by equations 1-2. So an agent in the population may be selfish, altruistic or vulnerability-responsive

[6] See Jaffry and Treur (2013) for a similar formulation.

and when she chooses her strategy as a second mover, she acts accordingly. Let $u^r_{i,t}$ be the utility of intrinsic preference type r. The expected utilities of each type are

$$u^a_t = (1+\theta^a)\bar{b}_t\pi_{cc} + (1-\bar{b}_t)(\pi_{cd} + \theta^a\pi_{dc}) \tag{6}$$

$$u^s_t = \bar{b}_t\pi_{dc} + (1-\bar{b}_t)\pi_{dd} \tag{7}$$

$$u^v_t = \bar{b}_t(1+\theta^v - \bar{b}_t\theta^v)\pi_{cc} + (1-\bar{b}_t)(\pi_{dd} - \bar{b}_t\theta^v\pi_{cd}) \tag{8}$$

where $\bar{b}_t$ is the mean trust level and given by $\bar{b}_t = \frac{\sum_{i=1}^n b_{i,t}}{n}$. We assume that agents have information on the average trust in the population. In other words, each agent has a subjective probabilistic belief that another agent will cooperate if they interact and furthermore, they have a signal about the mean level of trust. We argue that living in a community provides enough information about trust to agents. It can be passed not only through familial bonds but also through links with other social circles. Fukuyama (1995), for example, argues that low trust and high trust societies differ in behavior. We believe that it is possible to perceive this difference as an extrapolation of the average trust level by agents.

We suppose that at time t, σ^a_t is ratio of altruistic agents, σ^s_t is the ratio of selfish agents and therefore the ratio of vulnerability-responsive agents σ^v_t will be $\sigma^v_t = 1 - \sigma^a_t - \sigma^s_t$. Then, the average utility of the second movers in the population at time t, $\bar{u}^2_t$ can be expressed as

$$\bar{u}^2_t = \sigma^a_t u^a_t + \sigma^s_t u^s_t + \sigma^v_t u^v_t = \sigma^a_t u^a_t + \sigma^s_t u^s_t + (1-\sigma^a_t - \sigma^s_t)u^v_t. \tag{9}$$

Second movers in the population may revise their attitude after observing their payoffs. The revision process can be described as follows. Upon encounters with randomly chosen agents in the population, they compare the performance of their behaviors with this agent's behavior and adopt it if they find that it is better performing. Through time, the share of agents of different behavioural types changes given their imitation behavior. The imitation dynamics describes the change in the share of a behavioral trait in the population.

More specifically, at time t, the overall frequency of pairwise interactions between agents of different types r and $p \neq r$ is $\sigma_t^r \times \sigma_t^p$ which is the product of fractions of agents of preference type r and $p \neq r$ respectively. When an agent meets randomly another agent, they communicate each other their intrinsic preference and the respective payoff for this preference. Depending on the difference of their payoffs and possibly other factors, agents either go on having the same preference or update their preference to the one of the agent they are interacting. The imitation dynamics that we have described, reduces to the replicator equation (Ritzberger & Weibull, 1995).

In case of altruistic and selfish agents, given the respective utilities provided in Equations 6-7, the change in their share can be written as follows:

$$\begin{aligned}
\sigma_{t+1}^a &= \sigma_t^a + \beta\sigma_t^a(u_t^a - \bar{u}_t^2) && (10)\\
&= \sigma_t^a + \beta\sigma_t^a\left((1-\sigma_t^a)(u_t^a - u_t^v) - \sigma_t^s(u_t^s - u_t^v)\right)\\
\sigma_{t+1}^s &= \sigma_t^s + \beta\sigma_t^s(u_t^s - \bar{u}_t^2) && (11)\\
&= \sigma_t^s + \beta\sigma_t^s\left((1-\sigma_t^s)(u_t^s - u_t^v) - \sigma_t^a(u_t^a - u_t^v)\right)
\end{aligned}$$

The share of the vulnerability-responsive is $\sigma_t^v = 1 - \sigma_t^a - \sigma_t^s$ and the change in this share follows equations 10-11. The continuous time versions of the previous dynamics are:

$$\begin{aligned}
\frac{d\sigma^a}{dt} &= \dot{\sigma^a} = \beta\sigma^a\left((1-\sigma^a)\left(\pi_{cd} + \bar{b}_t(1-\theta^v + \bar{b}_t\theta^v)(\pi_{cc} - \pi_{cd})\right.\right. && (12)\\
&\left.\left. + (1-\bar{b}_t)(\pi_{dc} - \pi_{dd})\right) - \bar{b}_t\sigma^s\left(\pi_{dc} - \pi_{cc} - (1-\bar{b}_t)\theta^v(\pi_{cc} - \pi_{cd})\right)\right)\\
\frac{d\sigma^s}{dt} &= \dot{\sigma^s} = \beta\sigma^s\left(\bar{b}_t(1-\sigma^s)\left(\pi_{dc} - \pi_{cc} - (1-\bar{b}_t)\theta^v(\pi_{cc} - \pi_{cd})\right)\right. && (13)\\
&\left. - \sigma^a\left(\pi_{cd} + \bar{b}_t(1-\theta^v + \bar{b}_t\theta^v)(\pi_{cc} - \pi_{cd}) + (1-\bar{b}_t)(\pi_{dc} - \pi_{dd})\right)\right)
\end{aligned}$$

As the first players choose their strategy according to their belief that others will cooperate, they choose to cooperate blindly (CB) with probability $b_{i,t}$, and defect blindly (DB) with probability of $1 - b_{i,t}$. Then the expected utility of first players is:

$$u_{i,t} = \begin{cases} (1-\sigma_t^s)\pi_{cc} + \sigma_t^s\pi_{cd} & \text{with probability} \quad b_{i,t} \\ \sigma_t^a\pi_{dc} + (1-\sigma_t^a)\pi_{dd} & \text{with probability} \quad 1 - b_{i,t} \end{cases} \tag{14}$$

The utility of a first player when she cooperates is $(1-\sigma_t^s)\pi_{cc}+\sigma_t^s\pi_{cd}$ as with probability $(1-\sigma_t^s)$, she will meet an altruistic agent who always cooperates or a vulnerability-responsive who cooperates when the opponent cooperates and there will be a cooperative outcome. On the other hand, with probability σ_t^s, she will meet a selfish agent who behaves in accordance with Proposition 1 and her opponent will defect. The utility of a first player when she defects is $\sigma_t^a\pi_{dc}+(1-\sigma_t^a)\pi_{dd}$ as with probability σ_t^a, she will meet an altruistic agent who always cooperates and receives the highest utility. On the other hand, with probability $1-\sigma_t^a$, she will meet a selfish agent who behaves in accordance with Proposition 1 or a vulnerability-responsive who defects when the opponent defects and the Nash equilibrium will be the outcome. The trust dynamics that we have defined in equation 5 can be written using the utility function for the first movers given by equation 14.

3.4 Equilibrium distribution of behaviours

Now, we will study the asymptotic properties of the above mentioned dynamics. Through time, the behaviours and attitudes in the population evolve through two types of dynamics. The first is the adjustment dynamics whereby each agent updates her belief that others in the population are likely to cooperate which is described with equation 5. We describe the change in the trust level at the individual level. Since all first players have the same utility function and behaviour toward second players, the trust changes at the population level in parallel with the changes at the individual level[7] as follows:

$$\dot{\bar{b}} = \alpha\bar{b}_t\left(\bar{u}_t^1-\pi_{dd}\right) = \alpha\bar{b}_t\left(\bar{b}_t(1-\sigma_t^s)(\pi_{cc}-\pi_{cd})+(1-\bar{b}_t)\sigma_t^a(\pi_{dc}-\pi_{dd})\right) \tag{15}$$

where $\dot{\bar{b}}$ represents the change of trust level at the population level and $\bar{u}_t^1$ is the expected utility for first movers given by the following equation:

$$\bar{u}_t^1 = \bar{b}_t\left(((1-\sigma_t^s)\pi_{cc}+\sigma_t^s\pi_{cd})+(1-\bar{b}_t)\left(\sigma_t^a\pi_{dc}+(1-\sigma_t^a)\pi_{dd}\right)\right). \tag{16}$$

In the following section where we present numerical results, we use an agent-based

[7] See (Jaffry & Treur, 2013) for a similar approach

model, the discussion is left to that section. The second is the imitation dynamics through which agents modify their behaviours in pairwise interactions presented by equations 12 and 13.

It is interesting to study the equilibria or rest points of this dynamic system as at the equilibrium, the share of behaviours or attitudes in the population will not change. There is an equilibrium when either the entire population has the same behaviour or a mixture of behaviours are observed. The equilibrium is reached when $\dot{\bar{b}} = 0, \dot{\sigma^a} = 0, \dot{\sigma^s} = 0$. The system of equations have 6 equilibria $(\bar{b}^*, \sigma^{a*}, \sigma^{s*}, \sigma^{v*})$. Following are situations where the entire population has the same altruistic attitude.

- $(0, 1, 0, 0)$: This equilibrium corresponds to one agent cooperating and the other defecting outcome where only altruistic behaviours survive with no trust in the society. Although second movers are altruistic, without trust first movers will defect.

- $\left(\frac{\pi_{dc}-\pi_{dd}}{\pi_{dc}-\pi_{cc}}, 1, 0, 0\right)$: This is an equilibrium with altruist agents with $\frac{\pi_{dc}-\pi_{dd}}{\pi_{dc}-\pi_{cc}}$ level of trust. Note that $\bar{b}^*$ is in the unit interval but $\frac{\pi_{dc}-\pi_{dd}}{\pi_{dc}-\pi_{cc}} > 1$. We will evaluate the system at the boundary 1. With all altruistic agents and full trust, the imitation dynamics are at equilibrium. The trust dynamics on the other hand has the value $\alpha(\pi_{cc} - \pi_{dd}) > 0$. At the boundary, the movement is upwards but as the dynamics are bounded at 1 then there will not be any change. As a result, there is a rest point at $(1, 1, 0, 0)$. There will be a cooperative outcome with probability 1 as altruist second movers cooperate and first movers will cooperate with certainty.

Following equilibria depict situations where vulnerability-responsive and selfish agents co-exist.

- $(0, 0, \sigma^{s*}, 1 - \sigma^{s*})$: This solution corresponds to the non-cooperative outcome where only selfish and vulnerability-responsive behaviours are viable with no trust in the society. Any combination of these behaviours are possible with $\sigma^{s*} \in [0, 1]$. We have in this range of behavioural mixture, only selfish agents with no trust case bringing us to the traditional defecting rational agents in the prisoners' dilemma scenario.

- $\left(1 - \frac{\pi_{dc}-\pi_{cc}}{\theta^v(\pi_{cc}-\pi_{cd})}, 0, \frac{\pi_{cc}-\pi_{dd}}{\pi_{cc}-\pi_{cd}}, \frac{\pi_{dd}-\pi_{cd}}{\pi_{cc}-\pi_{cd}}\right)$: If we define $\pi_{dc} - \pi_{cc}$ as the value of the temptation, it is easy to see that an increase in the temptation reduces the trust at this rest point. Moreover, we can observe that an increase in the cooperation payoff leads to selfish agents dominating even though trust approaches to 1. This is because malicious agents can exploit trustee agents under the new equilibrium.

Following equilibria depicts a situation where altruist and selfish agents co-exist. This co-existence causes vulnerability-responsive agents to be removed from the system.

- $(\frac{\pi_{dd}-\pi_{dc}-\pi_{cd}}{2(\pi_{cc}-\pi_{dc})+\pi_{dd}-\pi_{cd}}, 1-\sigma^{s**}, \sigma^{s**}, 0)$ where $\sigma^{s**} = \frac{\pi_{cd}(\pi_{cc}-\pi_{dd})+(\pi_{cc}+\pi_{dd}-\pi_{dc})(\pi_{dd}-\pi_{dc})}{\pi_{cd}(\pi_{cc}-\pi_{cd})+(\pi_{dd}-\pi_{dc})(\pi_{cc}+\pi_{cd}-\pi_{dc})} \in [0,1]$: If $\frac{\pi_{dc}}{2} < \pi_{cc} < \pi_{dc} - \frac{\pi_{dd}+\pi_{cd}}{2}$ then $\bar{b}^{**}$ is in the unit interval. This situation is particularly interesting since one might expect vulnerability-responsive agents to perform better than the other types as they are more flexible. However, under this equilibrium, altruist agents benefit more in a defective situation since their altruistic attitude allows them to gain utility even when their own share is relatively low. Therefore, when trust does not exist, a vulnerability-responsive agent cannot be separated from a selfish agent. Moreover, when both types face a cooperative first-mover, selfish agents may gain a higher utility if the criteria depicted is met.

Finally, in our last equilibrium altruist and vulnerability-responsive agents co-exist.

- $(M, \sigma^{a*}, 0, 1-\sigma^{a*})$ where $\sigma^{a*} = \frac{M(\pi_{cc}-\pi_{dd})(2(\pi_{dc}-\pi_{cc})-(\pi_{dd}-\pi cd)-(\pi_{dc}+\pi_{cd}-\pi_{dd}))}{(1+M)(\pi_{dc}\pi_{dd})(\theta^v(\pi_{cd}\pi_{cc})+2\pi_{dc}-\pi_{cc}-\pi_{dd}+\pi_{cd})-M\pi_{cc}}$[8]: Contrary to our intuitions, it is a rare occurrence to have altruist and vulnerability-responsive agents to co-exist. The reason behind is our assumption of altruist agents receiving utility from their opponents, while vulnerability-responsive agents do not. However, if the condition is met, this difference is offset when the first-mover defects in the first turn.

3.5 Local stability analysis

Now, we will study the stability of the rest points of our model. Here, we refer to asymptotic stability or Liapunov stability (See Hirsch and Smale (1974)). To study

[8] M is given by the solution of the quadratic equation $(\theta_v(\pi_{cc}-\pi_{cd})M^2 + (\pi_{cc}+\pi_{cd}-\theta^v(\pi_{cc}) - \pi_{dc}-\pi_{cd}+\pi_{dd})M + \pi_{dc}+\pi_{cd}-\pi_{dd}) = 0$

if our equilibrium points are stable or not, we have to derive the Jacobian matrix. This matrix consists of gradients of our differential equations. Then, we find the eigenvalues of the matrix given an equilibrium point. If all of the real parts of the eigenvalues are negative at that point, we define that point as stable. Appendix C provides with the study of the Jacobian matrix at each equilibrium.

We find that the first altruistic population with no trust is unstable. The third steady state with combinations of selfish and vulnerability-responsive agents with no trust is also unstable. One of these combinations is all the population consisting of selfish agents with no trust and this result represents the traditional prediction in game theory resulting in non-cooperative outcome. This situation is unstable under the current scenario. The unstability has a cyclic nature of trust and selfishness. When people do not trust each other, a vulnerability-responsive agent weakly dominates a selfish agent in an evolutionary game theory setting. If we consider defect-defect scenario, it is easy to see that both selfish agents and vulnerability-responsive agents receive the same utility. However, if there is a deviation from no trust; two vulnerability responsive agents cooperate against each other; however, two selfish agents will get one temptation and one sucker's payoff at total. Since trustworthy vulnerability-responsive agent will receive extra payoff from not betraying her opponent, we can expect that the total payoff can be higher. However, this will lead to an increase in trust, which will benefit selfish agents. Therefore, the average trust will decrease again. The fourth state with some level of trust and a particular combination of selfish and vulnerability-responsive agents is unstable as well. This state can actually be considered as the continuation of the previous state, which means it will eventually return back to the third state in a cyclic nature. Any kind of increase in trust will always cause selfish agents to have an increased payoff; therefore, a reduction in trust.

The second altruistic population with full belief on the cooperative response is stable[9]. The cooperative outcome is stable if altruistic motivations by the trustworthy agents is backed by a certain level of trust by the trustees. We have seen that without trust altruist population will not survive and with full trust there will not be a steady state of altruistic agents.

[9]As $\pi_{dd} < \pi_{cc} < \pi_{dc}$ $0 \leq \frac{\pi_{dc}-\pi_{dd}}{\pi_{dc}-\pi_{cc}} \leq 1$

Proposition 2. *The only altruistic agents with full belief on cooperative response becomes stable if* $\frac{\pi_{dc}}{2} < \pi_{cc} < \frac{\pi_{dc}+\pi_{dd}}{2}$.

We see that there can be a cooperative outcome with all agents choosing to behave altruistically with full trust on their partners if the cooperative outcome is sufficiently high.

The fifth state where a variety of behaviours is observed with altruist and selfish agents remaining in the population $(\frac{\pi_{dd}-\pi_{dc}-\pi_{cd}}{2(\pi_{cc}-\pi_{dc})+\pi_{dd}-\pi_{cd}}, 1-\sigma^{s**}, \sigma^{s**}, 0)$ where $\sigma^{s**} = \frac{\pi_{cd}(\pi_{cc}-\pi_{dd})+(\pi_{cc}+\pi_{dd}-\pi_{dc})(\pi_{dd}-\pi_{dc})}{\pi_{cd}(\pi_{cc}-\pi_{cd})+(\pi_{dd}-\pi_{dc})(\pi_{cc}+\pi_{cd}-\pi_{dc})}$. If $\pi_{cc} < \frac{\pi_{dc}}{2}$ and $\pi_{cc} < \pi_{dc} - \frac{\pi_{dd}-\pi_{cd}}{2}$ then $\bar{b}^{**}$ is in the unit interval.

For simplification and a better illustration, we will take $\pi_{cd} = 0$ as the lowest level of utility without loss of generality. We find the following result for the stability of this mixed state.

Proposition 3. *The population distribution of behaviours can be stable at a mixture of altruist and selfish agents with* $\frac{\pi_{dd}-\pi_{dc}-\pi_{cd}}{2(\pi_{cc}-\pi_{dc})+\pi_{dd}-\pi_{cd}}$ *level of trust in cooperative action if* $\theta^v > \frac{(\pi_{cc}-\pi_{dd})(2\pi_{dc}-2\pi_{cc}-\pi_{dd})}{\pi_{cc}(2\pi_{cc}-\pi_{dc})}$ *and* $\frac{\beta}{\alpha} < \frac{\pi_{dd}(-2\pi_{dc}+2\pi_{cc}+\pi_{dd})}{4(\pi_{dc}-\pi_{cc})(\pi_{cc}+\pi_{dd}-\pi_{dc})}$.

For instance, for the set of parameters $\pi_{cd} = 0, \pi_{dd} = 1, \pi_{cc} = 8.3, \pi_{dc} = 13.55, \beta = 0.02, \alpha = 0.005, \theta^v = 0.8$; this steady state is stable. We see that unlike the previous stability conditions, this condition involves the relative speed of trust and imitation dynamics. In the next section, we will illustrate numerically this result.

4 NUMERICAL SIMULATIONS

Numerical simulations differ from analytical simulations in a very important way. It is possible to say that numerical simulations are often computational calculations of different initial values of analytical models. When an analytical model is created, it is a generalized model for a parameter space under certain assumptions. Then, numerical simulations attempt to throw many parameters to simulational tools to understand how does the analytical model respond to various parameters with many iterations.

4.1 Population-based model

Algorithm 1 Phase diagrams

Create a 2-simplex triangle.
for $x = 0, 1, 2, \ldots, 100$ **do** ▷ x, y, z coordinates for initial values
    for $y = 0, 1, 2, \ldots, 100$ **do**
        for *1-x-y* $= 0, 1, 2, \ldots, 100$ **do** ▷ To ensure that the sum is 1.
            Find the gradient for x, y, and z.
            Draw the gradient in the 2-simplex.
        end for
    end for
end for

Algorithm 1 shows how phase diagrams are drawn. Most important obstacle to overcome here is to convert Cartesian coordinates to barycentric coordinates. Even though it is still possible to plot our phase diagrams in a virtual 3-dimensional space, a 2-simplex allows us to show our results in 2-dimensional plane which is significantly more convenient and understandable. A 2-simplex is essentially a virtual representation of a 3-dimensional space by using a triangle's center of mass as the reference point.

Each corner symbolize a point in 3-dimensional space where only one of the dimensions is one and the other two are zero.

On contrary to the previous section, we enumerate our parameters to computationally calculate gradients. These parameters are shown in Table 4.1. We should mention that our results are mostly prone to different parameter combinations, although one should be careful when assigning parameters since there are constraints for parameters as well. For example, our biggest constraint is caused by prisoner's dilemma game: $\pi_{dc} > \pi_{cc} > \pi_{dd} > \pi_{cd}$.

We should also mention that we have tried various parameters that is between our constraints for behavioral rules as well. We have found that as long as these parameters are valid for our rule set, then our results will be in the same direction. However, even a slight overcome in any of the thresholds causes massive discords in the model. The parameters described in Table 4.1 are the final parameters we have used in our paper. From these parameters, only the trust value is being changed in the following results. The rest of the parameters will be considered as constant.

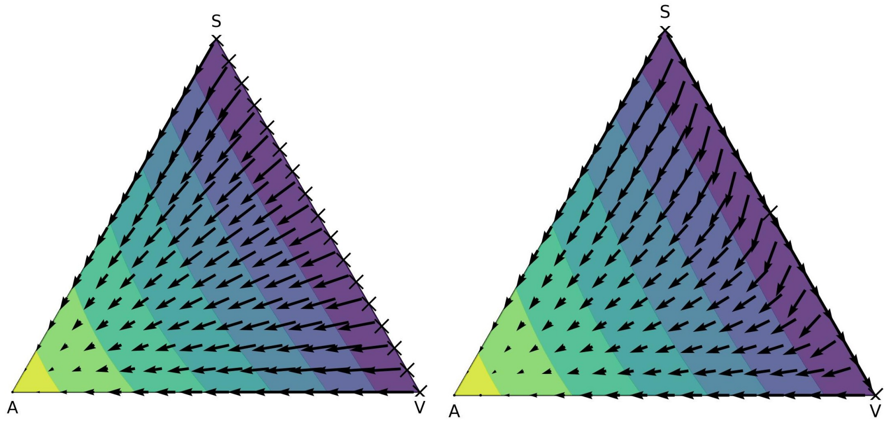
Parameter	Value range	Description
π_{cc}	1	Cooperation payoff.
π_{dc}	2	Temptation payoff.
π_{dd}	-1	Loser payoff.
π_{cd}	-2	Sucker payoff.
β	0.1	Evolution speed of reciprocal preferences.
θ^v	1	Reciprocal parameter of vulnerability-responsive agents.
$\bar{b}_t$	(0, 1)	Trust (assumed constant in phase diagrams).

Table 4.1: Parameters of phase diagrams.

4.1.1 A low level of trust

Figures 4.1a and 4.1b show the phase diagrams for a case where mean trust is extremely low. These phase diagrams show how the power dynamics between various states in a population happen throughout the time. However, one should not forget that the graph itself is statical and do not represent time-relationship. Each arrow shows a

Figure 4.1: Simplex for a low level of trust



(a) $\bar{b}_t = 0$

(b) $\bar{b}_t = 0.1$

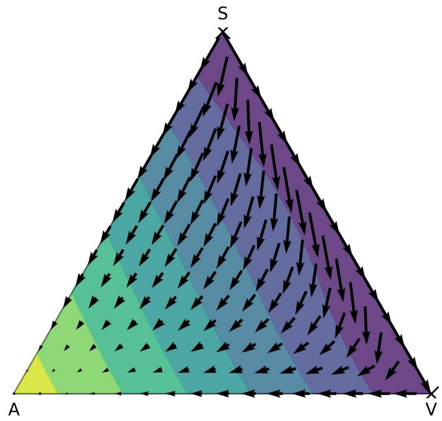
Phase Diagrams; A refers to altruist agents, V refers to vulnerability-responsive agents while S refers to selfish agents.

possible way of direction between the first point of the arrow and the end point of the arrow. Therefore, each coordinate in the simplex can be considered as a possible initial time.

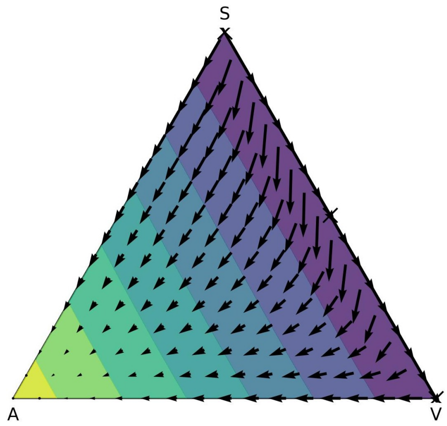
We observe that altruist agents prevail over other strategies in such a situation. We would expect that selfish agents would prevail in an atmosphere where trust is extremely low; therefore, this result is quite counter-intuitive. However, it makes sense as altruist agents care about others even when their opponents betray them. Moreover, these results are static from the average trust perspective in the model. In the dynamic model, we would observe an increase in trust, leading to an increase in selfish agents.

It is possible to observe a significant change of direction even when the trust increases a little bit. This is because a positive ratio of cooperation in the first turn causes a significant increase in the payoffs of vulnerable-responsive agents as they play cooperate conditionally. Therefore, there is a slight difference between Figures 4.1a and 4.1b in the bottom right corner. A slightly higher trust level causes vulnerability-responsive agents to gain some power compared to selfish agents.

Figure 4.2: Simplex for a medium level of trust



(a) $\bar{b}_t = 0.3$



(b) $\bar{b}_t = 0.6$

Phase Diagrams; A refers to altruist agents, V refers to vulnerability-responsive agents while S refers to selfish agents.

4.1.2 A medium level of trust

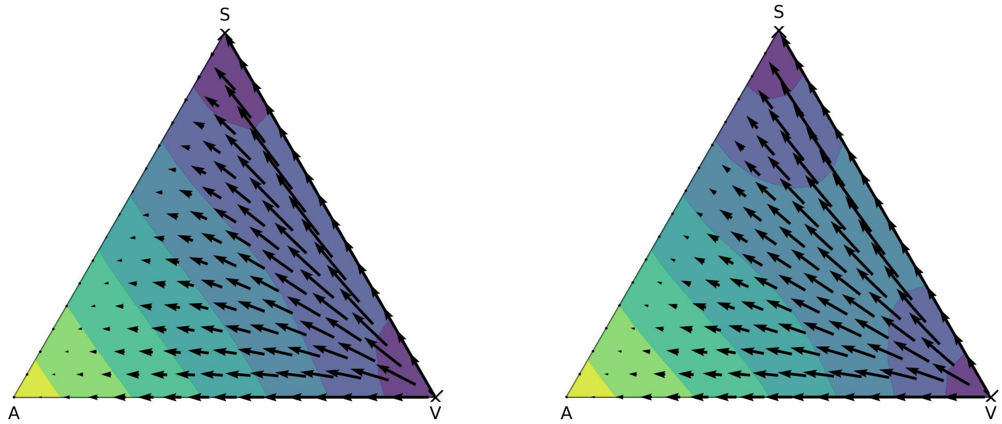
Figures 4.2a and 4.2b show the phase diagrams when trust level is mediocre in a population. The most important difference is the fact that how vulnerability-responsive agents became stronger in the community, even though altruist agents still win the evolutionary game. We should emphasise that a dynamic model in such a case resolves with an increased amount of trust; ergo, selfish agents gain more power and trust declines once more.

Moreover, we have realized that there difference between trust levels of 0.4 and 0.6 were not significant. Hence, we have decided to use 0.3 for an example to medium level of trust as well as 0.6. However, it is still surprising that Figure 4.2b resembles first two figures more than Figure 4.2a.

4.1.3 A high level of trust

Figures 4.3a and 4.3b show the scenario where trust is extremely high in the community. It is quite easy to suggest that the difference between low-trust and high-trust cases are visible to the naked eye. First, now selfish agents have a chance to steadily exist in a population. Moreover, altruist agents and selfish agents can co-exist. This result was also confirmed in our analytical models. However, our simulations

Figure 4.3: Simplex for a high level of trust



(a) $\bar{b}_t = 0.9$ (b) $\bar{b}_t = 1$

Phase Diagrams; A refers to altruist agents, V refers to vulnerability-responsive agents while S refers to selfish agents.

suggest that this is the only clear co-existence within the current parameter range.

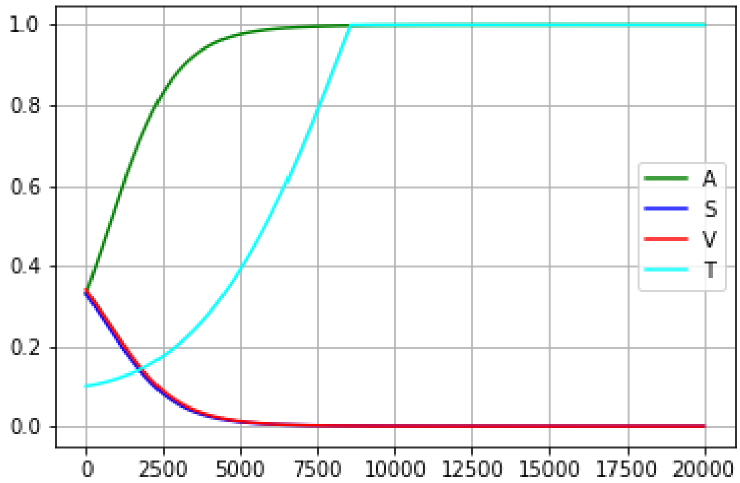
It is also important to note the color difference between various plots. Here, colors indicate the convergence speed of the equilibrium. Therefore, altruism is more powerful when it comes to a point where altruism exists in the model, while selfish agents do not benefit from other selfish agents. We believe that this distinction is quite characteristic. The more altruistic a population have, the more others benefit from altruism. On the opposite way, the more selfish agents a model have, the more the model is harmed from further selfish agents.

4.1.4 Representation of equilibrium points

We present some numerical simulations of our theoretical model in this section. We found some equilibrium points in Section 3. In order to visualize these points, we run simulations for different initial parameter combinations.

Figure 4.4 is a representation of a stable point in which altruism dominates other types. The trust level reaches one, similar to the previous model's result. The inequalities between payoffs are satisfied in the same way as the previous model assumed for stability.

Figure 4.4: Altruist agents dominate other types

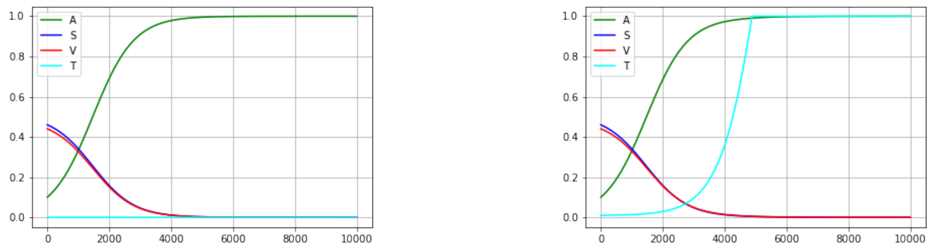


Results for $\sigma^a = 0.33, \sigma^s = 0.33, \sigma^v = 0.34, \bar{b}_t = 0.1, \theta^a = 1, \theta^v = 0.8, \alpha = 0.02, \beta = 0.05, \pi_{cd} = 0, \pi_{dd} = 1.1, \pi_{cc} = 2, \pi_{dc} = 3$. A refers to altruist agents, V refers to vulnerability-responsive agents while S refers to selfish agents.

Figure 4.5a displays altruist agents dominating the population even when they start with the lowest share and the trust level in the population is 0. However, figure 4.5b shows that even a slight deviation of trust from 0 causes trust level to go to 1 as in the figure 4.4. This result illustrates that the situation with no trust and entirely altruistic agents is not stable.

Figure 4.6 shows an equilibrium point with selfish and vulnerability-responsive agents co-exist. This co-existence has a cyclic nature. The proportion of selfish agents

Figure 4.5: Instability of altruism with no trust

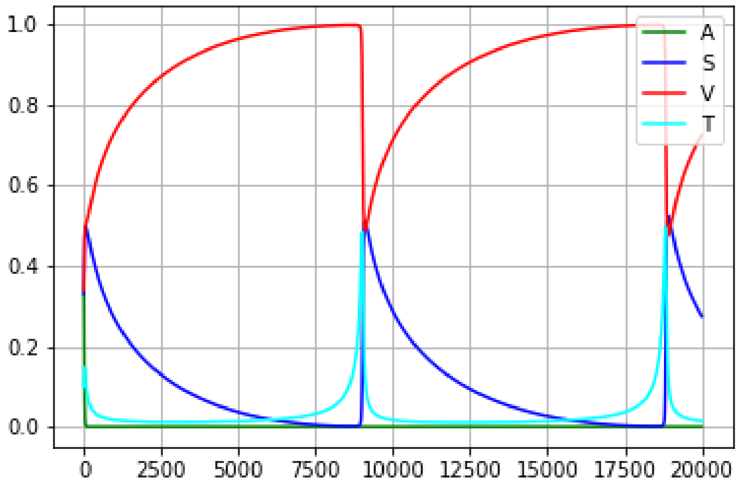


(a) $\bar{b}_t = 0$ (b) $\bar{b}_t = 0.01$

Results for $\sigma^a = 0.1, \sigma^s = 0.45, \sigma^v = 0.45, \bar{b}_t = 0.1, \theta^a = 1, \theta^v = 0.8, \alpha = 0.02, \beta = 0.05, \pi_{cd} = 0, \pi_{dd} = 1.1, \pi_{cc} = 2, \pi_{dc} = 3.$

decreases initially. Therefore, trust increases since vulnerable-responsive agents cooperate when faced with a cooperative action as opposed to selfishes. However, the rise of trust creates a suitable environment for selfish agents. Finally, this situation causes a cycle.

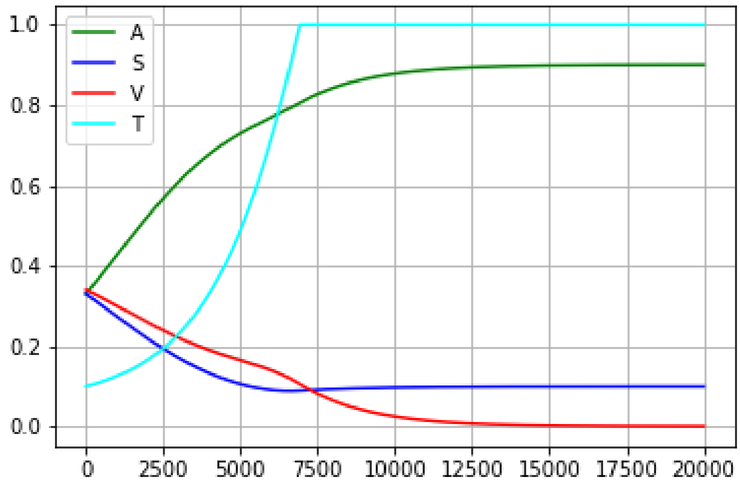
Figure 4.6: Selfish and vulnerability-responsive agents co-exist with a cyclic relation



Results for $\sigma^a = 0.33, \sigma^s = 0.33, \sigma^v = 0.34, \bar{b}_t = 0.1, \theta^a = 1, \theta^v = 0.8, \alpha = 0.02, \beta = 0.05, \pi_{cd} = -5, \pi_{dd} = 1, \pi_{cc} = 2, \pi_{dc} = 5.3.$

Figure 4.7 shows the stability of an equilibrium point where altruist and selfish agents co-exist. For this equilibrium, the initial population of types is equal. Equilibrium can reach complete trust, even with an initially low trust level. Selfish agents cannot prevent the trust from reaching the full level even though they always defect.

Figure 4.7: Altruism and selfishness exist together at equilibrium with complete trust



Results for $\sigma^a = 0.33, \sigma^s = 0.33, \sigma^v = 0.34, \bar{b}_t = 0.1, \theta^a = 1, \theta^v = 0.8, \alpha = 0.02, \beta = 0.05, \pi_{cd} = -2, \pi_{dd} = -1, \pi_{cc} = 1, \pi_{dc} = 2.$

4.2 Agent-based model

An agent-based model (ABM) is different from an equation-based model in a few crucial ways. First, an ABM is about the macro result of a micro-behavior. Therefore, the emphasis in the model is on the behavioral rules of the agent. Since it derives a numerical solution, results are more prone to parameters than an analytical solution. Moreover, the model is stochastic due to behaviors being based on probabilities. Hence, the simulation is run multiple times with different initial seeds and take the average of different runs. However, one has to be careful when taking the average of multiple runs because it may erase some important results. For example, let us consider two different runs where strategy A and B wins in each run, consecutively. In both runs, let one strategy peak first and then go to zero. An average of two runs would be two horizontal lines with no significant importance.

4.2.1 Formation of trust and evolution of reciprocal attitudes

To update trust, we use equation 5 since it allows agents to update their trusts using their previous encounters. However, for the strategy of second-turn, we use a different methodology. Each player looks for another random agent and decides to imitate his type with a proportion based on the payoff difference between hers and his. If the difference is high, favoring the opponent, there is a higher chance of imitation.

$$S_{i,t+1} = \frac{1}{1 + e^{(u_{i,t} - u_{j,t})/\mu}} \tag{17}$$

where $S_{i,t+1}$ is the probability of the agent i replicating the agent j. $u_{i,t}$ is the individual i's utility and $u_{j,t}$ is the opponent's utility at time t, $\mu > 0$ is the mutation rate. Here if μ approaches 0, an agent i with a higher utility never imitates her opponent, while if the opponent's payoff is higher even by a minimal amount, she always imitates. On the other hand, if μ approaches infinity, the imitation probability is $\frac{1}{2}$ for each agent, no matter the payoff difference. Equation 17 is a modified version of the Fermi equation. This equation allows for the probability to be a Sigmoid function.

4.2.2 Algorithm of the model

Algorithm 2 Algorithm for the agent-based model

```
for a parameter k = [range of the parameter k] do
    for a parameter l = [range of the parameter l] do    ▷ Try different parameter combinations.
        for seed = 0, 1, ..., 20 do    ▷ For each parameter combination, reduce stochasticity.
            Create an agent i with random type and random trust value.
            for turn = 0, 1, 2, ..., 200 do
                Agents randomly match and play the game
                Agents observe their payoffs and update their trust attitudes.
            end for
        end for
    end for
end for
```

Algorithm 2 shows how ABM is created. We run the simulations for 240 agents and 500 turns. In each run, all agents are created with a random trust level based on a Beta distribution and a reciprocal behavior type based on the initial population distribution of reciprocal behaviors. Since $\frac{a}{a+b}$ gives the mean of a Beta distribution, we choose a and b according to our desired initial mean trust level. In each turn, half of the population is assigned as player one and the other half as player two. Then, they match randomly and play a prisoner's dilemma game. The first players decide to cooperate or defect with a probability based on their trust level. The second players decide their actions according to their types. Altruist agents always choose to cooperate unconditionally, and selfish agents defect unconditionally. However, vulnerability-responsive agents cooperate if the first player's decision is cooperation; otherwise, they defect. After the game is completed and utilities are calculated, the first movers update their trust according to Equation 5. The second movers choose a random person among the second players and compare their payoff with them. They replicate their type with a probability given by the Fermi function described in Equation 17.

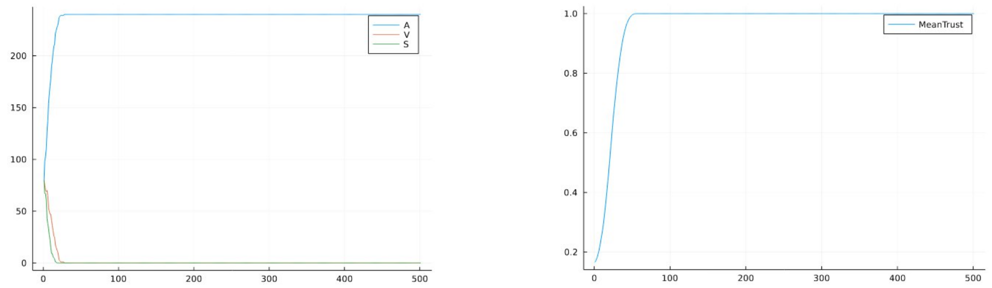
4.2.3 Results of the agent-based model

We choose random results for the specific parameters instead of taking the average results of multiple simulations. Taking average causes losing individual characteristics since some of the simulations end up at 0 for one type because of randomization. Moreover, taking the average allows only for determining when the equilibrium is reached. Since we aim to show the equilibrium can also be achieved with ABM using the same parameters, we must use individual results.

Figure 4.8 shows that altruist agents dominate the population, and trust reaches 1 when we use the same initial parameters as population-based simulation. These results are a clear indication of how robust our analytical models are from Figure 4.5b. It is easy to see that our agent-based model has achieved the same result as our analytical model, even when there is a significant difference in how both models behave.

Figure 4.9 shows that vulnerability-responsive and selfish agents co-exist with a cyclic relation as in the PBM; however, trust goes to a stable point in the ABM. This equilibrium point shows the most significant difference between ABM and PBM. Figure 4.6 corresponds to the analytical version of this model. Even though the cyclic behavior of the model remains the same, the analytical model has the property of trust

Figure 4.8: Altruist agents dominate other types

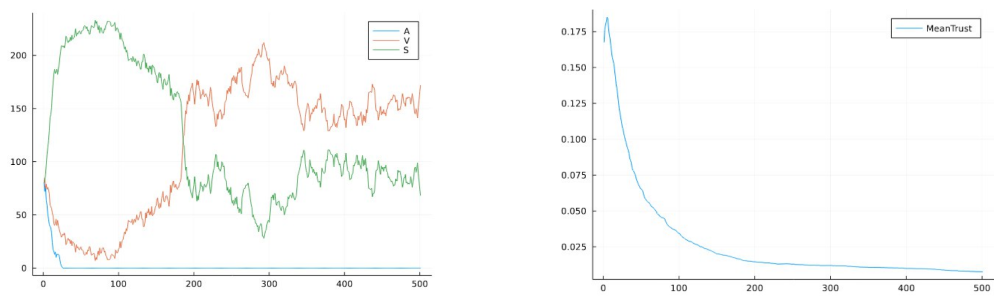


(a) Number of agents over time (b) Mean trust over time

Results for $\sigma^a = 0.33, \sigma^s = 0.33, \sigma^v = 0.34, \bar{b}_t = 0.1, \theta^a = 1, \theta^v = 0.8, \alpha = 0.02, \beta = 0.05, \pi_{cd} = 0, \pi_{dd} = 1.1, \pi_{cc} = 2, \pi_{dc} = 3$. A refers to altruist agents, V refers to vulnerability-responsive agents while S refers to selfish agents.

being a part of the cycle, while in the agent-based model, trust goes to very low levels. However, the main part of the cycle has the same result in both models. The difference between trust results could be due to how the agent-based model requires each agent to calculate their payoffs, while population-based dynamics use the representative agent when calculating how trust is updated.

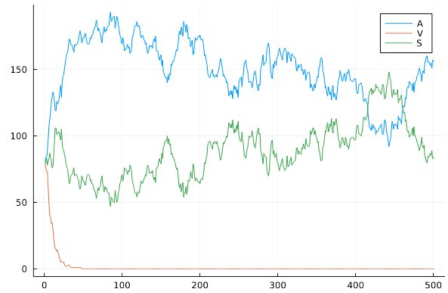
Figure 4.9: Selfish and vulnerability-responsive agents co-exist



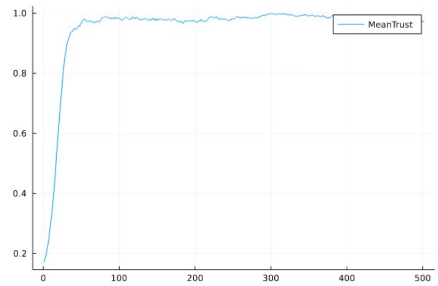
(a) Number of agents over time (b) Mean trust over time

Results for $\sigma^a = 0.33, \sigma^s = 0.33, \sigma^v = 0.34, \bar{b}_t = 0.1, \theta^a = 1, \theta^v = 0.8, \alpha = 0.02, \beta = 0.05, \pi_{cd} = -5, \pi_{dd} = 1, \pi_{cc} = 2, \pi_{dc} = 5.3$.

Figure 4.10: Altruism and selfishness exist together



(a) Number of agents over time



(b) Mean trust over time

Results for $\sigma^a = 0.33, \sigma^s = 0.33, \sigma^v = 0.34, \bar{b}_t = 0.1, \theta^a = 1, \theta^v = 0.8, \alpha = 0.02, \beta = 0.05, \pi_{cd} = -2, \pi_{dd} = -1, \pi_{cc} = 1, \pi_{dc} = 2.$

Figure 4.10 shows a situation where altruist and selfish agents co-exist in a cyclic situation. This co-existence creates a very different result compared to our previous model in 4.9 even though they look the same. Here, trust goes to 1 compared to the previous model, where it goes to 0. The difference is caused by altruist agents surviving here, while vulnerability-responsive agents are surviving in the previous model. We show that when altruist agents survive, trust has a chance to approach one even though selfish agents still exist. These results indicate the same results as our analytical model from Figure 4.7, where altruist and selfish agents co-exist, and trust goes to one.

5 CONCLUSION

In this paper, we have created an analytical framework to study the correlation between trust and trustworthiness. This correlation may seem trivial; however, we have shown that this correlation is deeper than it seems. First, the correlation tends to be cyclic rather than a positive one. This is because a high trust environment creates opportunity for a malicious agent, who can decrease the trust level of the population again. Therefore, a cyclic relationship appears. Second, our primary model is analytical; therefore, it is less prune to error. Since an analytical model is a generalized model where one can study various interactions between parameters, our model can be used to find how various trust levels affect a population's reciprocal preferences.

Both theoretical and empirical studies frequently employ the concepts of trust and trustworthiness. However, this paper is the first to develop a coevolutionary model between them. We reveal a deep and complex connection using both analytical and numerical studies, which would be very hard to study otherwise. Even though the correlation seems evident initially, the issues and cyclic behavior due to high trust creating new opportunities for malicious agents are important to this literature. This cycle is due to malicious agents benefiting more when the trust level in the society is high. Our analytical solutions point to different equilibria where multiple reciprocal behaviors can either eliminate each other or co-exist. These equilibria are then tested in numerical simulations. Both our phase diagrams of the analytical model and agent-based models result in parallel.

Altruism is a very strong motive in human interactions. We have tested another common motive, vulnerability-responsiveness against altruism in this paper. We have shown that a pure altruistic motive is still very strong compared to both selfish motives and vulnerability-responsive motives. Nevertheless, a vulnerability-responsive agent or a selfish agent can still find her place in the population if the trust is low enough.

Our analytical results point to a situation where altruist and vulnerability-responsive agents, altruist and selfish agents or vulnerability-responsive and selfish agents can co-exist; however, we do not observe all types existing in the equilibrium. Moreover, domination of a type is observed. We believe these results are significant due to their possible implications in agent-based models. An agent-based model is often criticized for not having a solid base. Regardless, when analytical solutions provide simple and generalized constraints, agent-based models allow us to test different behaviors under various initial conditions. This paper follows a straightforward framework suggested by O'Connor (2017). In this framework, an agent-based model is used for robustness after an analytical solution is presented. Since our analytical results are consistent with our numerical simulations, we can suggest that our model is robust to error. Moreover, different initialisation of different parameters yield very interesting results, while the main topic remains the same.

Both the effects and components of trust are crucial aspects of society. We show that even when some agents betray each other, as in the case in Section 3, trust can increase over time and if there are enough altruists in the population the full trust can be a stable. Furthermore, a very simple set of rules are enough to replicate real-world dynamics of social trust. Here, we created an agent-based model to see such dynamics in action. Agent-based models can be used to develop policies when they are also linked by real-world data (Dawid & Gatti, 2018). This paper intends not to suggest policies but to show how agent-based modelling can be used to both predict and analyze social components of bilateral behaviours.

To conclude, we can say that there could be many implementations of this framework for policymakers. Since Algan and Cahuc (2010), social trust has been a relatively popular topic in economics due to its relatively unpredicted effect on macroeconomic indicators. However, it is not easy to predict social capital using traditional methods. For example, even the causality between inequality and trust has been a serious dispute among economists (Bergh & Bjørnskov, 2014). Under these circumstances, we find it safe to say that agent-based modelling could be one of the strongest candidates for such a challenging task. Finally, another further implementation of this model could be done by using social network theory, as it uses interpersonal dynamics and studies them thoroughly.

REFERENCES

Alesina, A., & Giuliano, P., (2015), Culture and institutions, *Journal of Economic Literature*, *53*(4), 898–944.

Alesina, A., & La Ferrara, E., (2002), Who trusts others?, *Journal of Public Economics*, *85*(2), 207–234.

Algan, Y., & Cahuc, P., (2010), Inherited trust and growth, *American Economic Review*, *100*(5), 2060–2092.

Axelrod, R., & Hamilton, W. D., (1981), The evolution of cooperation, *Science*, *211*(4489), 1390–1396.

Becker, G. S., (1976), Altruism, egoism, and genetic fitness: Economics and sociobiology, *Journal of Economic Literature*, *14*(3), 817–826.

Berg, J., Dickhaut, J., & Mccabe, K., (1995), Berg, Dickhaut, Mccabe - 1995 - Trust, Reciprocity and Social History.pdf.

Bergh, A., & Bjørnskov, C., (2014), Trust, welfare states and income equality: Sorting out the causality, *European Journal of Political Economy*, *35*, 183–199.

Bhandari, H., & Yasunobu, K., (2009), What is social capital? A comprehensive review of the concept, *Asian Journal of Social Science*, *37*(3).

Bosse, T., Jaffry, S. W., Siddiqui, G. F., & Treur, J., (2012), Comparative analysis of agent-based and population-based modelling in epidemics and economics, *Multiagent and Grid Systems*, *8*(3), 223–255.

Bourdieu, P., (1986), The forms of capital, *Cultural theory: An anthology*, *1*, 81–93.

Brown, J. S., Cunningham, J. J., & Gatenby, R. A., (2016), Aggregation effects and population-based dynamics as a source of therapy resistance in cancer, *IEEE Transactions on Biomedical Engineering*, *64*(3), 512–518.

Carrasco, J. A., Harrison, R., & Villena, M., (2018), Interdependent preferences and endogenous reciprocity, *Journal of Behavioral and Experimental Economics*, *76*, 68–75.

Charness, G., & Rabin, M., (2002), Understanding social preferences with simple tests, *Quarterly Journal of Economics*, *117*(3), 817–869.

Coleman, J. S., (1988), Social capital in the creation of human capital, *American journal of sociology*, *94*, S95–S120.

Cox, J. C., Kerschbamer, R., & Neururer, D., (2016), What is trustworthiness and what drives it?, *Games and Economic Behavior*, *98*, 197–218.

Danielson, P., (2002), Competition among cooperators: Altruism and reciprocity, *Proceedings of the National Academy of Sciences*, *99*(suppl_3), 7237–7242.

Dawid, H., & Gatti, D. D., (2018), Agent-based macroeconomics, *Handbook of computational economics*, *4*, 63–156.

Di Cagno, D., & Sciubba, E., (2010), Trust, trustworthiness and social networks: Playing a trust game when networks are formed in the lab, *Journal of Economic Behavior & Organization*, *75*(2), 156–167.

Dinesen, P. T., (2012), Does generalized (dis) trust travel? examining the impact of cultural heritage and destination-country environment on trust of immigrants, *Political Psychology*, *33*(4), 495–511.

Dinesen, P. T., Schaeffer, M., & Sønderskov, K. M., (2020), Ethnic diversity and social trust: A narrative and meta-analytical review, *Annual Review of Political Science*, *23*(1), 441–465.

Dohmen, T., Falk, A., Huffman, D., & Sunde, U., (2008), Representative trust and reciprocity: Prevalence and determinants, *Economic Inquiry*, *46*(1), 84–90.

Dohmen, T., Falk, A., Huffman, D., & Sunde, U., (2012), The intergenerational transmission of risk and trust attitudes, *The Review of Economic Studies*, *79*(2), 645–677.

Douds, K., & Wu, J., (2018), Trust in the Bayou City: Do Racial Segregation and Discrimination Matter for Generalized Trust?, *Sociology of Race and Ethnicity*, *4*(4), 567–584.

Dutt, A. K., (2006), Aggregate demand, aggregate supply and economic growth, *International review of applied economics*, *20*(3), 319–336.

Ehsan, A., Klaas, H. S., Bastianen, A., & Spini, D., (n.d.), Social capital and health: A systematic review of systematic reviews, *SSM - Population Health*, *8*(April).

Fehr, E., & Fischbacher, U., (2003), The nature of human altruism, *Nature*, *425*(6960), 785–791.

Fehr, E., & Schmidt, K. M., (2006), The economics of fairness, reciprocity and altruism–experimental evidence and new theories, *Handbook of the Economics of Giving, Altruism and Reciprocity*, *1*, 615–691.

Fukuyama, F., (2001), Social capital, civil society and development, *Third world quarterly*, *22*(1), 7–20.

Gamage, S. K. N., Prasanna, R., Jayasundara, J., Ekanayake, E., Rajapakshe, P., GAKNJ, A., Kumudumali, S., & Nedelea, A. M., (2020), Social capital and sme: A systematic literature review and research directions, *Ecoforum Journal*, *9*(3).

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L., (2000), Measuring trust, *Quarterly Journal of Economics*, *115*(3), 811–846.

Guiso, L., Sapienza, P., & Zingales, L., (2006), Does culture affect economic outcomes?, *Journal of Economic perspectives*, *20*(2), 23–48.

Guzmán, R., Harrison, R., Abarca, N., & Villena, M. G., (2020), A game-theoretic model of reciprocity and trust that incorporates personality traits, *Journal of Behavioral and Experimental Economics*, *84*, 101497.

Hagen, J. M., & Choe, S., (1998), Trust in japanese interfirm relations: Institutional sanctions matter, *Academy of management Review*, *23*(3), 589–600.

Hardin, R., (1999), From bodo ethics to distributive justice, *Ethical Theory and Moral Practice*, *2*(4), 399–413.

Helbing, D., (2012), Agent-based modeling. In *Social self-organization* (pp. 25–70), Springer.

Hennes, D., Morrill, D., Omidshafiei, S., Munos, R., Perolat, J., Lanctot, M., Gruslys, A., Lespiau, J.-B., Parmas, P., Duenez-Guzman, E., et al., (2019), Neural replicator dynamics, *arXiv preprint arXiv:1906.00190*.

Hirsch, M., & Smale, S., (1974), Differential equations, dynamical systems, and linear algebra (pure and applied mathematics, vol. 60).

Isoni, A., & Sugden, R., (2019), Reciprocity and the paradox of trust in psychological game theory, *Journal of Economic Behavior & Organization*, *167*, 219–227.

Jaffry, S. W., & Treur, J., (2013), Agent-based and population-based modeling of trust dynamics, *Transactions on Computational Collective Intelligence IX*, 124–151.

Knack, S., & Keefer, P., (1997), Does social capital have an economic payoff? a cross-country investigation, *The Quarterly journal of economics*, *112*(4), 1251–1288.

Krause, J., Croft, D. P., & James, R., (2007), Social network theory in the behavioural sciences: Potential applications, *Behavioral Ecology and Sociobiology*, *62*(1), 15–27.

Krishna, A., & Shrader, E., (1999), Social capital assessment tool, *Conference on social capital and poverty reduction*, *2224*.

Kwon, S.-W., & Arenius, P., (2010), Nations of entrepreneurs: A social capital perspective, *Journal of Business Venturing*, *25*(3), 315–330.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. W., (1997), Legal determinants of external finance, *The journal of finance*, *52*(3), 1131–1150.

Leonardi, R., Nanetti, R. Y., & Putnam, R. D., (2001), *Making democracy work: Civic traditions in modern italy*, Princeton University Press Princeton, NJ, USA.

Levine, D. K., (1998), Modeling altruism and spitefulness in experiments, *Review of Economic Dynamics*, *1*(3), 593–622.

Lippert, S., & Spagnolo, G., (2011), Networks of Relations and Word-of-Mouth Communication, *Games and Economic Behavior*, *72*(1), 202–217.

Lütz, A. F., Amaral, M. A., & Wardil, L., (2021), Moderate immigration may promote a peak of cooperation among natives, *Physical Review E*, *104*(1), 1–9.

Mishra, S., (2020), Social networks, social capital, social support and academic success in higher education: A systematic review with a special focus on 'underrepresented' students, *Educational Research Review*, *29*, 100307.

O'Connor, C., (2017), The cultural red king effect, *The Journal of Mathematical Sociology*, *41*(3), 155–171.

Paldam, M., (2000), Social capital: One or many? definition and measurement, *Journal of Economic Surveys*, *14*(5), 629–653.

Phelps, S., Marcinkiewicz, M., Parsons, S., & McBurney, P., (2005), Using population-based search and evolutionary game theory to acquire better-response strategies for the double-auction market, *Proceedings of IJCAI-05 workshop on trading agent design and analysis (TADA-05)*.

Poder, T. G., (2011), What is Really Social Capital? A Critical Review, *American Sociologist*, *42*(4), 341–367.

Putnam, R., (2001), Social capital: Measurement and consequences, *Canadian Journal of Policy Research*, *2*(1), 41–51.

Putnam, R. D., et al., (2000), *Bowling alone: The collapse and revival of american community*, Simon; schuster.

Rabin, M., (1993), Incorporating fairness into game theory and economics, *The American Economic Review*, 1281–1302.

Rathbun, B. C., (2011), *Trust in international cooperation: International security institutions, domestic politics and american multilateralism* (Vol. 121), Cambridge University Press.

Ritzberger, K., & Weibull, J. W., (1995), Evolutionary selection in normal-form games, *Econometrica: Journal of the Econometric Society*, 1371–1399.

Robison, L. J., Schmid, A. A., & Siles, M. E., (2002), Is social capital really capital?, *Review of social economy*, *60*(1), 1–21.

Schelling, T. C., (1971), Dynamic models of segregation, *Journal of mathematical sociology*, *1*(2), 143–186.

Sethi, R., & Somanathan, E., (2001), Preference evolution and reciprocity, *Journal of Economic Theory*, *97*(2), 273–297.

Sethi, R., & Somanathan, E., (2003), Understanding reciprocity, *Journal of Economic Behavior & Organization*, *50*(1), 1–27.

Simon, H. A., (1993), Altruism and economics, *The American Economic Review*, *83*(2), 156–161.

Slemrod, J., & Katuščák, P., (2005), Do trust and trustworthiness pay off?, *Journal of Human Resources*, *40*(3), 621–646.

Sobel, J., (2005), Interdependent preferences and reciprocity, *Journal of Economic Literature*, *43*(2), 392–436.

Sønderskov, K. M., (2011), Does generalized social trust lead to associational membership? unravelling a bowl of well-tossed spaghetti, *European Sociological Review*, *27*(4), 419–434.

Takeuchi, N., & Hogeweg, P., (2012), Evolutionary dynamics of rna-like replicator systems: A bioinformatic approach to the origin of life, *Physics of Life Reviews*, *9*(3), 219–263.

Uslaner, E. M., (2002), The moral foundations of trust, *Available at SSRN 824504*.

Zak, P. J., & Knack, S., (2001), Trust and growth, *The economic journal*, *111*(470), 295–321.

A Equilibrium of the sequential prisoners' dilemma

Here, we will present the equilibrium of the sequential prisoners' dilemma when agents are rational and choose their strategies as follows: first movers act based on their belief of cooperative behaviour and second movers act given their intrinsic preferences representing their trustworthiness. This situation can be described a Bayesian game where players can be of different types. Their utilities can be calculated based on the expected cooperative action by the first movers $\mathbb{E}b_t$.

The utilities of playing different strategies for an **altruist agent** i against an opponent j are:

$$u_{i,t}(CU, s_j) = (1 + \theta^a)\mathbb{E}b_t\pi_{cc} + (1 - \mathbb{E}b_t)(\pi_{cd} + \theta^a\pi_{dc}) \tag{18}$$

$$u_{i,t}(CC, s_j) = (1 + \theta^a)\mathbb{E}b_t\pi_{cc} + (1 - \mathbb{E}b_t)(1 + \theta^a)\pi_{dd} \tag{19}$$

$$u_{i,t}(DU, s_j) = \mathbb{E}b_t(\pi_{dc} + \theta^a\pi_{cd}) + (1 - \mathbb{E}b_t)(1 + \theta^a)\pi_{dd}. \tag{20}$$

- Against a cooperating opponent, CC or CU provides the same utility; against a defecting opponent, if the agent plays CC, they will both defect and the agent will have $(1 + \theta^a)\pi_{dd}$ utility and if the agent plays CU, she will cooperate and have $(1 + \theta^a\frac{\pi_{dc}}{\pi_{cd}})\pi_{cd}$. The utility of playing CU is higher than the utility of CC if and only if $\theta^a > \frac{\pi_{dd}-\pi_{cd}}{\pi_{dc}-\pi_{dd}}$.

- Against a defecting opponent, CC or DU provides the same utility; against a cooperating opponent, if the agent plays CC, they will both cooperate and the agent will have $(1 + \theta^a)\pi_{cc}$ utility and if the agent plays DU, she will defect and have $(1 + \theta^a\frac{\pi_{cd}}{\pi_{dc}})\pi_{dc}$. The utility of playing CC is higher than the utility of DU if and only if $\theta^a > \frac{\pi_{dc}-\pi_{cc}}{\pi_{cc}-\pi_{cd}}$.

The utilities of playing different strategies for a **selfish agent** i against an opponent j

are:

$$u_{i,t}(CU, s_j) = \mathbb{E}b_t\pi_{cc} + (1 - \mathbb{E}b_t)\pi_{cd} \tag{21}$$

$$u_{i,t}(CC, s_j) = \mathbb{E}b_t\pi_{cc} + (1 - \mathbb{E}b_t)\pi_{dd} \tag{22}$$

$$u_{i,t}(DU, s_j) = \mathbb{E}b_t\pi_{dc} + (1 - \mathbb{E}b_t)\pi_{dd} \tag{23}$$

From the conditions on the payoffs of the game ($\pi_{dc} > \pi_{cc} > \pi_{dd} > \pi_{cd}$), it is trivial that DU brings the highest utility for a selfish agent. The utilities of playing different strategies for a **vulnerability-responsive agent** i against an opponent j are:

$$u_{i,t}(CU, s_j) = \mathbb{E}b_t\pi_{cc}(1 + \theta^v - \mathbb{E}b_t\theta^v) + (1 - \mathbb{E}b_t)(1 - \theta^v\mathbb{E}b_t)\pi_{cd} \tag{24}$$

$$u_{i,t}(CC, s_j) = \mathbb{E}b_t\pi_{cc}(1 + \theta^v - \mathbb{E}b_t\theta^v) + (1 - \mathbb{E}b_t)(\pi_{dd} - \theta^v\mathbb{E}b_t\pi_{cd}) \tag{25}$$

$$u_{i,t}(DU, s_j) = \mathbb{E}b_t(\pi_{dc} + \theta^v\pi_{cd} - \mathbb{E}b_t\theta^v\pi_{cc}) + (1 - \mathbb{E}b_t)(\pi_{dd} - \mathbb{E}b_t\theta^v) \tag{26}$$

The utility of CC is higher than CU since $\pi_{dd} > \pi_{cd}$. The utility of CC is higher than DU if $\theta^v > \dfrac{\pi_{dc} - \pi_{cc}}{\pi_{cc} - \pi_{cd}}$.

As far as first movers are concerned, there are two alternatives; either they will cooperate blindly or they will defect blindly. Suppose that the above conditions are satisfied for second movers, then the following utilities will be provided for each alternative. The expected utility of a first mover cooperating agent i is

$$\mathbb{E}u_{i,t} = (1 - p^s_{i,t})\pi_{cc} + p^s_{i,t}\pi_{cd} \tag{27}$$

and the expected utility of a first mover defecting agent i is

$$\mathbb{E}u_{i,t} = (1 - p^a_{i,t})\pi_{dd} + p^a_{i,t}\pi_{dc} \tag{28}$$

If $p^v_{i,t} > p^a_{i,t}\frac{-\pi_{cc}+\pi_{dc}}{\pi_{cc}-\pi_{dd}} + p^s_{i,t}\frac{-\pi_{cd}+\pi_{dd}}{\pi_{cc}-\pi_{dd}}$, then the first agent will choose to cooperate and otherwise the agent will defect.

B Steady States the System

The steady state is reached when $\dot{\bar{b}} = 0, \dot{\sigma^a} = 0, \dot{\sigma^s} = 0$. The system of equations have 5 steady states $(\bar{b}^*, \sigma^{a*}, \sigma^{s*}, \sigma^{v*})$.

The first trust dynamics is at rest if

$$\dot{\bar{b}} = \alpha\bar{b}_t\left(\bar{b}_t(1 - \sigma^s_t)(\pi_{cc} - \pi_{cd}) + (1 - \bar{b}_t)\sigma^a_t(\pi_{dc} - \pi_{dd})\right) = 0. \tag{29}$$

There are two solutions to this equation. The first is the trivial

$$\bar{b_t} = 0 \Rightarrow b^* = 0 \tag{30}$$

and the second is given by

$$\begin{aligned}
&\bar{b_t}(1-\sigma_t^s)(\pi_{cc}-\pi_{cd}) + (1-\bar{b_t})\sigma_t^a(\pi_{dc}-\pi_{dd}) = 0 \\
&\Rightarrow b^* = \frac{-\sigma^{a*}(\pi_{dc}-\pi_{dd})}{(\pi_{cc}-\pi_{dd}) - \sigma^{s*}(\pi_{cc}-\pi_{cd}) - \sigma^{a*}(\pi_{dc}-\pi_{dd})}
\end{aligned} \tag{31}$$

The share of altruist agents does not change if

$$\begin{aligned}
\dot{\sigma^s} = \beta\sigma^a \Big(&(1-\sigma^a)\left(\pi_{cd} + \bar{b_t}(1-\theta^v+\bar{b_t}\theta^v)(\pi_{cc}-\pi_{cd}) + (1-\bar{b_t})(\pi_{dc}-\pi_{dd})\right) \\
&-\bar{b_t}\sigma^s\left(\pi_{dc}-\pi_{cc}-(1-\bar{b_t})\theta^v(\pi_{cc}-\pi_{cd})\right)\Big) = 0
\end{aligned} \tag{32}$$

There are two solutions to this equation. The first is the trivial

$$\sigma^a = 0 \Rightarrow \sigma^{a*} = 0 \tag{33}$$

and the second is given by

$$\begin{aligned}
&(1-\sigma^a)\left(\pi_{cd} + \bar{b_t}(1-\theta^v+\bar{b_t}\theta^v)(\pi_{cc}-\pi_{cd}) + (1-\bar{b_t})(\pi_{dc}-\pi_{dd})\right) \\
&-\bar{b_t}\sigma^s\left(\pi_{dc}-\pi_{cc}-(1-\bar{b_t})\theta^v(\pi_{cc}-\pi_{cd})\right) = 0 \\
&\Rightarrow \sigma^{a*} = 1 - \frac{-\bar{b}^*\sigma^{s*}\left(\pi_{dc}-\pi_{cc}-(1-\bar{b}^*)\theta^v(\pi_{cc}-\pi_{cd})\right)}{\pi_{cd} + \bar{b}^*(1-\theta^v+\bar{b}^*\theta^v)(\pi_{cc}-\pi_{cd}) + (1-\bar{b}^*)(\pi_{dc}-\pi_{dd})}
\end{aligned} \tag{34}$$

The share of selfish agents does not change if

$$\begin{aligned}\dot{\sigma^s} =& \beta\sigma^s \left(\bar{b_t}(1-\sigma^s)\left(\pi_{dc}-\pi_{cc}-(1-\bar{b_t})\theta^v(\pi_{cc}-\pi_{cd})\right)\right.\\ &\left.-\sigma^a\left(\pi_{cd}+\bar{b_t}(1-\theta^v+\bar{b_t}\theta^v)(\pi_{cc}-\pi_{cd})+(1-\bar{b_t})(\pi_{dc}-\pi_{dd})\right)\right)=0 \end{aligned} \quad (35)$$

There are two solutions to this equation. The first is the trivial

$$\sigma^s = 0 \Rightarrow \sigma^{s*} = 0 \quad (36)$$

and the second is given by

$$\begin{aligned}&\bar{b_t}(1-\sigma^s)\left(\pi_{dc}-\pi_{cc}-(1-\bar{b_t})\theta^v(\pi_{cc}-\pi_{cd})\right)\\ &-\sigma^a\left(\pi_{cd}+\bar{b_t}(1-\theta^v+\bar{b_t}\theta^v)(\pi_{cc}-\pi_{cd})+(1-\bar{b_t})(\pi_{dc}-\pi_{dd})\right)=0\\ &\Rightarrow \sigma^{s*} = 1-\frac{\sigma^a\left(\pi_{cd}+b^*(1-\theta^v+b^*\theta^v)(\pi_{cc}-\pi_{cd})+(1-b^*)(\pi_{dc}-\pi_{dd})\right)}{b^*\left(\pi_{dc}-\pi_{cc}-(1-b^*)\theta^v(\pi_{dc}-\pi_{dd})\right)}\end{aligned} \quad (37)$$

This system of equations (equations 30, 31, 33, 34, 36, 37) admits six solutions.

- $(0,1,0,0)$: This solution corresponds to one agent cooperating and the other defecting outcome where only altruistic behaviours survive with no trust in the society.
- $\left(\frac{\pi_{dc}-\pi_{dd}}{\pi_{dc}-\pi_{cc}},1,0,0\right)$: This solution corresponds to all agents behaving altruistically with a certain level of trust.
- $(0,0,\sigma^{s*},1-\sigma^{s*})$: This solution corresponds to the uncooperative outcome where only selfish and vulnerability-responsive behaviours are viable with no trust in the society.
- $\left(1-\frac{\pi_{dc}-\pi_{cc}}{\theta^v(\pi_{cc}-\pi_{cd})},0,\frac{\pi_{cc}-\pi_{dd}}{\pi_{cc}-\pi_{cd}},\frac{\pi_{dd}-\pi_{cd}}{\pi_{cc}-\pi_{cd}}\right)$: This solution corresponds to only selfish and vulnerability-responsive behaviours with a certain trust in the society.
- $(\frac{\pi_{dd}-\pi_{dc}-\pi_{cd}}{2(\pi_{cc}-\pi_{dc})+\pi_{dd}-\pi_{cd}},1-\sigma^{s**},\sigma^{s**},0)$ where $\sigma^{s**}=\frac{\pi_{cd}(\pi_{cc}-\pi_{dd})+(\pi_{cc}+\pi_{dd}-\pi_{dc})(\pi_{dd}-\pi_{dc})}{\pi_{cd}(\pi_{cc}-\pi_{cd})+(\pi_{dd}-\pi_{dc})(\pi_{cc}+\pi_{cd}-\pi_{dc})}$: This rest point depicts a sit-

uation where altruist and selfish agents co-exist. This co-existence causes vulnerability-responsive agents to be removed from the system.

- $(M, \sigma^{a*}, 0, 1-\sigma^{a*})$ where $\sigma^{a*} = \frac{M(\pi_{cc}-\pi_{dd})(2(\pi_{dc}-\pi_{cc})-(\pi_{dd}-\pi cd)-(\pi_{dc}+\pi_{cd}-\pi_{dd}))}{(1+M)(\pi_{dc}\pi_{dd})(\theta^v(\pi_{cd}\pi_{cc})+2\pi_{dc}-\pi_{cc}-\pi_{dd}+\pi_{cd})-M\pi_{cc}}$[10]: Finally, in our last steady state altruist and vulnerability-responsive agents co-exist.

C The Local Stability Analysis

We denote the Jacobian matrix by $J = [J_{ij}]$. According to our dynamical system, the matrix dimension is 3×3:

$$
\begin{aligned}
J_{11} = \frac{\partial \dot{\sigma}^a}{\partial \sigma^a_t} =& \beta\left((1-2\sigma^a_t)(u^a_t - u^v_t) - \sigma^s_t(u^s_t - u^v_t)\right) \\
J_{12} = \frac{\partial \dot{\sigma}^a}{\partial \sigma^s_t} =& \beta\sigma^a_t(u^v_t - u^s_t) \\
J_{13} = \frac{\partial \dot{\sigma}^a}{\partial b_t} =& \beta\sigma^a_t\left((1-\sigma^a)\left((1+2\bar{b}_t\theta^v)(\pi_{cc}-\pi_{cd}) + \theta^v\pi_{cd} + \pi_{dd} - \pi_{dc}\right)\right. \\
& \left. -\sigma^s_t\left(2\bar{b}_t\theta^v(\pi_{cc}-\pi_{cd}) - \pi_{cc} + \theta^v\pi_{cd} + \pi_{dc}\right)\right) \\
J_{21} = \frac{\partial \dot{\sigma}^s}{\partial \sigma^a_t} =& \beta\sigma^s_t(u^v_t - u^a_t) \\
J_{22} = \frac{\partial \dot{\sigma}^s}{\partial \sigma^s_t} =& \beta\left((1-2\sigma^s_t)(u^s_t - u^v_t) - \sigma^a_t(u^a_t - u^v_t)\right) \\
J_{23} = \frac{\partial \dot{\sigma}^s}{\partial \bar{b}_t} =& \beta\sigma^s_t\left((1-\sigma^s)\left(\theta^v(2\bar{b}_t - 1)(\pi_{cc}-\pi_{cd}) + \pi_{dc} - \pi_{cc}\right)\right. \\
& \left. -\sigma^a\left((1+2\bar{b}_t\theta^v - \theta^v)(\pi_{cc}-\pi_{cd}) + \pi_{dd} - \pi_{dc}\right)\right) \\
J_{31} = \frac{\partial \dot{\bar{b}}}{\partial \sigma^a_t} =& \alpha\bar{b}_t(1-\bar{b}_t)(\pi_{dc}-\pi_{dd}) \\
J_{32} = \frac{\partial \dot{\bar{b}}}{\partial \sigma^s_t} =& \alpha\bar{b}^2_t(\pi_{cd}-\pi_{cc}) \\
J_{33} = \frac{\partial \dot{\bar{b}}}{\partial b_t} =& \alpha\left(2(1-\sigma^s)(\pi_{cc}-\pi_{cd})\bar{b}_t + \sigma^a(1-2\bar{b}_t)(\pi_{dc}-\pi_{dd})\right)
\end{aligned}
$$

where $u^a_t - u^v_t = \bar{b}_t(1-\theta^v+\bar{b}_t\theta^v)\pi_{cc} + (1-\bar{b}_t)\left((1+\bar{b}_t\theta^v)\pi_{cd} + \pi_{dc} - \pi_{dd}\right)$ and $u^s_t - u^v_t = \bar{b}_t\left(\pi_{dc} - (1+\theta^v - \bar{b}_t\theta^v)\pi_{cc} + (1-\bar{b}_t)\theta^v\pi_{cd}\right)$.

[10] M is given by the solution of the quadratic equation $(\theta_v(\pi_{cc}-\pi_{cd})M^2 + (\pi_{cc}+\pi_{cd}-\theta^v(\pi_{cc}) - \pi_{dc}-\pi_{cd}+\pi_{dd})M + \pi_{dc}+\pi_{cd}-\pi_{dd}) = 0$

Jacobian Matrix for $(\bar{b}^*, \sigma^{a*}, \sigma^{s*}, \sigma^{v*}) = (0, 1, 0, 0)$ is given by

$$J_1 = \begin{bmatrix} \beta(\pi_{dd} - \pi_{cd} - \pi_{dc}) & 0 & 0 \\ 0 & \beta(\pi_{dd} - \pi_{cd} - \pi_{dc}) & 0 \\ 0 & 0 & \alpha(\pi_{dc} - \pi_{dd}) \end{bmatrix}$$

with the associated eigenvalues

$$\lambda_1 = \alpha(\pi_{dc} - \pi_{dd}) > 0$$
$$\lambda_2 = \lambda_3 = -\beta(\pi_{cd} + \pi_{dc} - \pi_{dd}).$$

As one of the eigenvalues is non-negative, this steady state is unstable.

Jacobian matrix for the equilibrium point $(1, 1, 0, 0)$ is given by

$$J_2 = \begin{bmatrix} -\beta\pi_{cc} & \beta(\pi_{cc} - \pi_{dc}) & 0 \\ 0 & \beta(\pi_{dc} - 2\pi_{cc}) & 0 \\ 0 & \alpha(\pi_{cd} - \pi_{cc}) & \alpha\left(2\pi_{cc} - \pi_{dd} - \pi_{dc}\right) \end{bmatrix}$$

The associated eigenvalues are:

$$\begin{aligned} \lambda_1 =& -\beta\pi_{cc} < 0 \\ \lambda_2 =& \beta(\pi_{dc} - 2\pi_{cc}) \\ \lambda_3 =& \alpha(2\pi_{cc} - \pi_{dc} - \pi_{dd}) \end{aligned}$$

λ_1 is always negative. λ_2 and λ_3 are negative if $\pi_{dc} < 2\pi_{cc}$ and $2\pi_{cc} < \pi_{dc} + \pi_{dd}$. Then the system will be stable if $\frac{\pi_{dc}}{2} < \pi_{cc} < \frac{\pi_{dc}+\pi_{dd}}{2}$.

Jacobian matrix for the equilibrium point $(\bar{b}^*, \sigma^{a*}, \sigma^{s*}, \sigma^{v*}) = (0, 0, \sigma^{s*}, 1 - \sigma^{s*})$:

$$\mathcal{J}_3 = \begin{bmatrix} \beta\sigma^{s*}(\pi_{dd} + \pi_{cd} - \pi_{dc}) \\ 0 \\ 0 \end{bmatrix.$$
$$\left.\begin{matrix} \beta\sigma^{s*}(\pi_{dd} - \pi_{cd} - \pi_{dc}) & 0 \\ 0 & 0 \\ \beta\sigma^{s*}(1 - \sigma^{s*})\left(\pi_{dc} - \pi_{cc} + \theta^v(\pi_{cc} - \pi_{cd})\right) & 0 \end{matrix}\right] \tag{38}$$

with the associated eigenvalues

$$\lambda_1 = \lambda_2 = 0$$

$$\lambda_3 = \beta(\pi_{cd} + \pi_{dc} - \pi_{dd}).$$

As two of the eigenvalues are non-negative, this steady state is unstable.

Jacobian matrix for the equilibrium point
$(\bar{b}^*, \sigma^{a*}, \sigma^{s*}, \sigma^{v*}) = \left(1 - \frac{\pi_{dc}-\pi_{cc}}{\theta^v(\pi_{cc}-\pi_{cd})}, 0, \frac{\pi_{cc}-\pi_{dd}}{\pi_{cc}-\pi_{cd}}, \frac{\pi_{dd}-\pi_{cd}}{\pi_{cc}-\pi_{cd}}\right)$:

$$\mathcal{J}_4 = \begin{bmatrix} \frac{2\beta\left((\theta^v+1)\pi_{cc}^2+\left((-\frac{\theta^v}{2}-2)\pi_{dc}+\frac{\pi_{dd}-\pi_{cd}}{2}-\pi_{cd}\theta^v\right)\pi_{cc}+\frac{\pi_{dc}}{2}((1+\theta^v)\pi_{cd}+2\pi_{dc}-\pi_{dd})\right)}{\theta^v(\pi_{cc}-\pi_{cd})} & \frac{-2\beta(\pi_{cc}-\pi_{dd})\left((\theta^v+1)\pi_{cc}^2+\left((-\frac{\theta^v}{2}-2)\pi_{dc}+\frac{\pi_{dd}-\pi_{cd}}{2}-\pi_{cd}\theta^v\right)\pi_{cc}+\frac{\pi_{dc}}{2}(\pi_{cd}\theta^v+\pi_{cd}+2\pi_{dc}-\pi_{dd})\right)}{(\pi_{cc}-\pi_{cd})^2\theta^v} & \frac{-\alpha(\theta^v(\pi_{cc}-\pi_{cd})+\pi_{cc}-\pi_{dc})(\pi_{cc}-\pi_{dc})(\pi_{dc}-\pi_{dd})}{{\theta^v}^2(\pi_{cc}-\pi_{cd})^2} \\ 0 & 0 & \frac{-\alpha(\theta^v(\pi_{cc}-\pi_{cd})+\pi_{cc}-\pi_{dc})^2}{{\theta^v}^2(\pi_{cc}-\pi_{cd})} \\ 0 & \frac{-\beta(\pi_{cc}-\pi_{dd})(\theta^v(\pi_{cc}-\pi_{cd})+\pi_{cc}-\pi_{dc})(\pi_{cd}-\pi_{dd})}{(\pi_{cc}-\pi_{cd})^2} & 0 \end{bmatrix} \quad (39)$$

$$\lambda_1 = \frac{2\beta\left((\theta^v+1)\pi_{cc}^2+\left((-\frac{\theta^v}{2}-2)\pi_{dc}+\frac{\pi_{dd}-\pi_{cd}}{2}-\pi_{cd}\theta^v\right)\pi_{cc}+\frac{\pi_{dc}((1+\theta^v)\pi_{cd}+2\pi_{dc}-\pi_{dd})}{2}\right)}{\theta^v(\pi_{cc}-\pi_{cd})}$$

$$\lambda_2 = \frac{\sqrt{\alpha\beta}\,(\theta^v(\pi_{cc}-\pi_{cd})+\pi_{cc}-\pi_{dc})\sqrt{\pi_{cd}-\pi_{dd}}\sqrt{(\theta^v(\pi_{cc}-\pi_{cd})+\pi_{cc}-\pi_{dc})(\pi_{cc}-\pi_{dd})}}{\theta^v(\pi_{cc}-\pi_{cd})^{3/2}},$$

$$\lambda_3 = -\frac{\sqrt{\alpha\beta}(\theta^v(\pi_{cc}-\pi_{cd})+\pi_{cc}-\pi_{dc})\sqrt{\pi_{cd}-\pi_{dd}}\sqrt{(\theta^v(\pi_{cc}-\pi_{cd})+\pi_{cc}-\pi_{dc})(\pi_{cc}-\pi_{dd})}}{\theta^v(\pi_{cc}-\pi_{cd})^{3/2}}$$

λ_2 and λ_3 are opposite signs of the same expression. Either one of them is positive and the other one is negative, or both of them are equal to 0. Since stability requires both eigenvalues to be negative, this equilibrium point cannot be stable.

$$\Bigg[\frac{\left(\frac{\pi_{dc}^2}{2}+\frac{\left(\left(-\frac{\theta^v}{2}-2\right)\pi_{cc}+\left(\frac{\theta^v}{2}+\frac{1}{2}\right)\pi_{cd}-\frac{\pi_{dd}}{2}\right)\pi_{dc}}{2}+\frac{\pi_{cc}\left((\theta^v+1)\pi_{cc}+\left(-\theta^v-\frac{1}{2}\right)\pi_{cd}+\frac{\pi_{dd}}{2}\right)}{2}\right)\beta(\pi_{cd}-\pi_{dd})(\pi_{cd}+\pi_{dc}-\pi_{dd})^2}{\left(\pi_{dc}^2+(-\pi_{cc}-\pi_{cd}-\pi_{dd})\pi_{dc}-\pi_{cd}^2+(\pi_{cc}+\pi_{dd})\pi_{cd}+\pi_{dd}\pi_{cc}\right)\left(\pi_{cc}-\frac{\pi_{cd}}{2}-\pi_{dc}+\frac{\pi_{dd}}{2}\right)^2}$$

$$\frac{\left(\frac{\pi_{dc}^2}{2}+\frac{\left(\left(-\frac{\theta^v}{2}-2\right)\pi_{cc}+\left(\frac{\theta^v}{2}+\frac{1}{2}\right)\pi_{cd}-\frac{\pi_{dd}}{2}\right)\pi_{dc}}{2}+\frac{\pi_{cc}\left((\theta^v+1)\pi_{cc}+\left(-\theta^v-\frac{1}{2}\right)\pi_{cd}+\frac{\pi_{dd}}{2}\right)}{2}\right)\beta(\pi_{cd}-\pi_{dd})(\pi_{cd}+\pi_{dc}-\pi_{dd})^2}{\left(\pi_{dc}^2+(-\pi_{cc}-\pi_{cd}-\pi_{dd})\pi_{dc}-\pi_{cd}^2+(\pi_{cc}+\pi_{dd})\pi_{cd}+\pi_{dd}\pi_{cc}\right)\left(\pi_{cc}-\frac{\pi_{cd}}{2}-\pi_{dc}+\frac{\pi_{dd}}{2}\right)^2}$$

$$-\frac{(2\pi_{cc}-\pi_{cd}-2\pi_{dc}+\pi_{dd})\beta\left(\pi_{dd}^2+(\pi_{cc}-\pi_{cd}-2\pi_{dc})\pi_{dd}+\pi_{cd}\pi_{cc}-\pi_{dc}(\pi_{cc}-\pi_{dc})\right)(\pi_{cd}-\pi_{dd})(\pi_{cd}+\pi_{dc}-\pi_{dd})}{\left((\pi_{cc}+\pi_{cd}-\pi_{dc})\pi_{dd}-\pi_{cd}^2+(\pi_{cc}-\pi_{dc})\pi_{cd}-\pi_{dc}(\pi_{cc}-\pi_{dc})\right)^2}$$

$$-\frac{(2\pi_{cc}-\pi_{cd}-2\pi_{dc}+\pi_{dd})\beta\left(\pi_{dd}^2+(\pi_{cc}-\pi_{cd}-2\pi_{dc})\pi_{dd}+\pi_{cd}\pi_{cc}-\pi_{dc}(\pi_{cc}-\pi_{dc})\right)(\pi_{cd}-\pi_{dd})(\pi_{cd}+\pi_{dc}-\pi_{dd})}{\left((\pi_{cc}+\pi_{cd}-\pi_{dc})\pi_{dd}-\pi_{cd}^2+(\pi_{cc}-\pi_{dc})\pi_{cd}-\pi_{dc}(\pi_{cc}-\pi_{dc})\right)^2}$$

$$\mathcal{J}_5 = \frac{-\frac{\beta}{2}\left(\pi_{dc}^2+\left(\left(-\frac{\theta^v}{2}-2\right)\pi_{cc}+\left(\frac{\theta^v+1}{2}\right)\pi_{cd}-\pi_{dd}\right)\pi_{dc}+\pi_{cc}\left((\theta^v+1)\pi_{cc}+\left(-\theta^v-\frac{1}{2}\right)\pi_{cd}+\frac{\pi_{dd}}{2}\right)\right)\left(\pi_{dc}^2+(-\pi_{cc}-2\pi_{dd})\pi_{dc}+(\pi_{cd}+\pi_{dd})\pi_{cc}-\pi_{dd}(\pi_{cd}-\pi_{dd})\right)(\pi_{cd}+\pi_{dc}-\pi_{dd}}{\left(\pi_{dc}^2+(-\pi_{cc}-\pi_{cd}-\pi_{dd})\pi_{dc}+(\pi_{cd}+\pi_{dd})\pi_{cc}-\pi_{cd}(\pi_{cd}-\pi_{dd})\right)\left(\pi_{cc}-\frac{\pi_{cd}}{2}-\pi_{dc}+\frac{\pi_{dd}}{2}\right)^2}$$

$$\frac{(2\pi_{cc}-\pi_{cd}-2\pi_{dc}+\pi_{dd})\beta\left(\pi_{dd}^2+(\pi_{cc}-\pi_{cd}-2\pi_{dc})\pi_{dd}+\pi_{cd}\pi_{cc}-\pi_{dc}(\pi_{cc}-\pi_{dc})\right)(\pi_{cd}-\pi_{dd})(\pi_{cd}+\pi_{dc}-\pi_{dd})}{\left((\pi_{cc}+\pi_{cd}-\pi_{dc})\pi_{dd}-\pi_{cd}^2+(\pi_{cc}-\pi_{dc})\pi_{cd}-\pi_{dc}(\pi_{cc}-\pi_{dc})\right)^2}$$

$$-\frac{\alpha(\pi_{cd}+\pi_{dc}-\pi_{dd})(2\pi_{cc}-\pi_{dc})(\pi_{dc}-\pi_{dd})}{(2\pi_{cc}-\pi_{cd}-2\pi_{dc}+\pi_{dd})^2}$$

$$-\frac{\alpha(\pi_{cd}+\pi_{dc}-\pi_{dd})^2(\pi_{cc}-\pi_{cd})}{(2\pi_{cc}-\pi_{cd}-2\pi_{dc}+\pi_{dd})^2}$$

$$\frac{(\pi_{dc}-\pi_{dd})(\pi_{cd}-\pi_{dd})(\pi_{cd}+\pi_{dc}-\pi_{dd})\alpha}{-\pi_{cd}^2+(\pi_{cc}-\pi_{dc}+\pi_{dd})\pi_{cd}-(\pi_{dc}-\pi_{dd})(\pi_{cc}-\pi_{dc})}\Bigg] \tag{40}$$

$$\lambda_1 = -2\beta\left(\left(\theta^v+1\right)\pi_{cc}^2+\left(\left(-\frac{\theta^v}{2}-2\right)\pi_{dc}+\left(-\theta^v-\frac{1}{2}\right)\pi_{cd}+\frac{\pi_{dd}}{2}\right)\pi_{cc} + \frac{\pi_{dc}\left(2\pi_{dc}+\pi_{cd}\left(\theta^v+1\right)-\pi_{dd}\right)}{2}\right)\frac{\left(\pi_{cd}+\pi_{dc}-\pi_{dd}\right)}{\left(2(\pi_{cc}-\pi_{dc})-\pi_{cd}+\pi_{dd}\right)^2}$$

$$\lambda_2 = \frac{\sqrt{\alpha}\sqrt{\pi_{cd}-\pi_{dd}}\sqrt{2\pi_{cc}-\pi_{cd}-2\pi_{dc}+\pi_{dd}}(\pi_{cd}+\pi_{dc}-\pi_{dd})}{4(-\pi_{cd}^2+(\pi_{cc}-\pi_{dc}+\pi_{dd})\pi_{cd}-(\pi_{dc}-\pi_{dd})(\pi_{cc}-\pi_{dc})(\pi_{cc}-\pi_{dc}+\frac{\pi_{dd}-\pi_{cd}}{2})} \Big(\Big(4\beta(\pi_{cc}-\pi_{dd})\pi_{cd}^3+\Big(-4\beta\pi_{cc}^2+4\beta\pi_{cc}\pi_{dd}+4((\pi_{dc}-2\pi_{dd})\beta - \frac{\alpha(\pi_{dc}-\pi_{dd})}{4}(\pi_{dc}-\pi_{dd}))\pi_{cd}^2+8(\pi_{dc}-\pi_{dd})\left(\beta\pi_{cc}^2+(\frac{-3\pi_{dc}+\pi_{dd}}{2})\beta + \frac{\alpha(\pi_{dc}-\pi_{dd})}{4}\right)\pi_{cc}+\frac{\beta}{2}(\pi_{dc}^2-\pi_{dc}\pi_{dd}+\pi_{dd}^2)-\frac{\alpha(\pi_{dc}-\pi_{dd})^2}{4}\Big)\pi_{cd} - 4\left(\beta\pi_{cc}^2+\left((-2\pi_{dc}+\pi_{dd})\beta+\frac{\pi_{dd}\alpha}{2}\right)\pi_{cc}+\pi_{dc}(\pi_{dc}-\pi_{dd})\beta - \frac{(\pi_{dc}-\frac{\pi_{dd}}{2})\alpha\pi_{dd}}{2}\right)(\pi_{dc}-\pi_{dd})^2\Big)^{1/2} + (\pi_{dc}-\pi_{dd})\left(2(\pi_{cc}-\pi_{dc})+\pi_{dd}-\pi_{cd}\right)\alpha(\pi_{cd}-\pi_{dd})\Big)$$

$$\lambda_3 = \frac{-\sqrt{\alpha}\sqrt{\pi_{cd}-\pi_{dd}}\sqrt{2\pi_{cc}-\pi_{cd}-2\pi_{dc}+\pi_{dd}}(\pi_{cd}+\pi_{dc}-\pi_{dd})}{4(-\pi_{cd}^2+(\pi_{cc}-\pi_{dc}+\pi_{dd})\pi_{cd}-(\pi_{dc}-\pi_{dd})(\pi_{cc}-\pi_{dc})(\pi_{cc}-\pi_{dc}+\frac{\pi_{dd}-\pi_{cd}}{2})} \Big(\Big(4\beta(\pi_{cc}-\pi_{dd})\pi_{cd}^3+\Big(-4\beta\pi_{cc}^2+4\beta\pi_{cc}\pi_{dd}+4((\pi_{dc}-2\pi_{dd})\beta - \frac{\alpha(\pi_{dc}-\pi_{dd})}{4}(\pi_{dc}-\pi_{dd}))\pi_{cd}^2+8(\pi_{dc}-\pi_{dd})\left(\beta\pi_{cc}^2+(\frac{-3\pi_{dc}+\pi_{dd}}{2})\beta + \frac{\alpha(\pi_{dc}-\pi_{dd})}{4}\right)\pi_{cc}+\frac{\beta}{2}(\pi_{dc}^2-\pi_{dc}\pi_{dd}+\pi_{dd}^2)-\frac{\alpha(\pi_{dc}-\pi_{dd})^2}{4}\Big)\pi_{cd} - 4\left(\beta\pi_{cc}^2+\left((-2\pi_{dc}+\pi_{dd})\beta+\frac{\pi_{dd}\alpha}{2}\right)\pi_{cc}+\pi_{dc}(\pi_{dc}-\pi_{dd})\beta - \frac{(\pi_{dc}-\frac{\pi_{dd}}{2})\alpha\pi_{dd}}{2}\right)(\pi_{dc}-\pi_{dd})^2\Big)^{1/2} - (\pi_{dc}-\pi_{dd})\left(2(\pi_{cc}-\pi_{dc})+\pi_{dd}-\pi_{cd}\right)\alpha(\pi_{cd}-\pi_{dd})\Big)$$

Taking $\pi_{cd} = 0$ as the lowest level of utility to simplify and illustrate we find that $\lambda_1 < 0$ if $\theta^v > \frac{(\pi_{cc}-\pi_{dd})(2\pi_{dc}-2\pi_{cc}-\pi_{dd})}{\pi_{cc}(2\pi_{cc}-\pi_{dc})}$. λ_2 and λ_3 are negative if $\frac{\beta}{\alpha} < \frac{\pi_{dd}(-2\pi_{dc}+2\pi_{cc}+\pi_{dd})}{4(\pi_{dc}-\pi_{cc})(\pi_{cc}+\pi_{dd}-\pi_{dc})}$.

Jacobian matrix for the equilibrium point $(\bar{b}^*, \sigma^{a*}, \sigma^{s*}, \sigma^{v*}) = (M, \sigma^{a*}, 0, 1-\sigma^{a*})$ where $\sigma^{a*} = \frac{M(\pi_{cc}-\pi_{dd})(2(\pi_{dc}-\pi_{cc})-(\pi_{dd}-\pi cd)-(\pi_{dc}+\pi_{cd}-\pi_{dd}))}{(1+M)(\pi_{dc}\pi_{dd})(\theta^v(\pi_{cd}\pi_{cc})+2\pi_{dc}-\pi_{cc}-\pi_{dd}+\pi_{cd})-M\pi_{cc}}$:

$$\left[\begin{matrix}-2\beta\left(\theta^{v}\left(\pi_{cc}-\pi_{cd}\right)M^{2}+\left(\left(-\pi_{cc}+\pi_{cd}\right)\theta^{v}+\pi_{cc}-\pi_{cd}-\pi_{dc}+\pi_{dd}\right)M+\pi_{cd}+\pi_{dc}-\pi_{dd}\right)\\ \left(LM-\frac{1}{2}\right)\\ \\ -L\beta M^{2}\left(\left(\pi_{cc}-\pi_{cd}\right)\left(-1+M\right)\theta^{v}-\pi_{cc}+\pi_{dc}\right)\\ \\ -2\beta L\left(\left(M-\frac{1}{2}\right)\left(\pi_{cc}-\pi_{cd}\right)\theta^{v}+\frac{\left(\pi_{cc}-\pi_{cd}-\pi_{dc}+\pi_{dd}\right)}{2}\right)M\left(LM-1\right)\end{matrix}\right.$$

$$0$$

$$\mathcal{J}_6 = \qquad \beta\left(\left(\theta^{v}\left(\pi_{cc}-\pi_{cd}\right)M^{2}+\left(\left(-\pi_{cc}+\pi_{cd}\right)\theta^{v}+\pi_{cc}-\pi_{cd}-\pi_{dc}+\pi_{dd}\right)M+\right.\right.$$
$$\left.\pi_{cd}+\pi_{dc}-\pi_{dd}\right)L-\theta^{v}\left(\pi_{cc}-\pi_{cd}\right)M+\theta^{v}\left(\pi_{cc}-\pi_{cd}\right)+\pi_{cc}-\pi_{dc}\Big)$$

$$0$$

$$\left.\begin{matrix}-\alpha M\left(-1+M\right)\left(\pi_{dc}-\pi_{dd}\right)\\ \\ -\alpha M^{2}\left(\pi_{cc}-\pi_{cd}\right)\\ \\ -2\alpha M\left(\left(M-\frac{1}{2}\right)\left(\pi_{dc}-\pi_{dd}\right)L-\pi_{cc}+\pi_{dd}\right)\end{matrix}\right] \tag{41}$$

where $L = \frac{M(\pi_{cc}(\pi_{dc}-2\pi_{cc})-\pi_{dd}(\pi_{dc}-2\pi_{cc}))}{1+M(\pi_{dc}(\pi_{cc}\theta^{v}\pi_{cd})\pi_{dd}(\pi_{cc}\theta^{v}\pi_{cd})+2\pi_{dc}-\pi_{cc}-\pi_{dd}+\pi_{cd})-M\pi_{cc}}$ and M is given by the solution of the quadratic equation $(\theta_v(\pi_{cc}-\pi_{cd})M^2+(\pi_{cc}+\pi_{cd}-\theta^v(\pi_{cc})-\pi_{dc}-\pi_{cd}+\pi_{dd})M+\pi_{dc}+\pi_{cd}-\pi_{dd})=0$.

$$
\begin{aligned}
\lambda_1 =& -\beta M((\theta^v(\pi_{cc}-\pi_{cd})M^2+((-\pi_{cc}+\pi_{cd})\theta^v+\pi_{cc}-\pi_{cd}-\pi_{dc}+\pi_{dd})M \\
& +\pi_{cd}+\pi_{dc}-\pi_{dd})L-\theta^v(\pi_{cc}-\pi_{cd})M+\theta^v(\pi_{cc}-\pi_{cd})+\pi_{cc}-\pi_{dc}) \\
\lambda_2 =& \frac{1}{2}\left(4L^2\beta^2\theta^{v2}(\pi_{cc}-\pi_{cd})^2M^6-8L\theta^v(((-\theta^v+1)\pi_{cd}+\theta^v\pi_{cc}-\pi_{cc}+\pi_{dc}-\pi_{dd})L+\right. \\
& \frac{\theta^v(\pi_{cc}-\pi_{cd})}{2})(\pi_{cc}-\pi_{cd})\beta^2M^5+((((4\theta^{v^2}-16\theta^v+4)\pi_{cd}^2+(-8\pi_{cc}\theta^{v^2} \\
& +(24\pi_{cc}-16\pi_{dc}+16\pi_{dd})\theta^v-8\pi_{cc}+8\pi_{dc}-8\pi_{dd})\pi_{cd}+4\pi_{cc}^2\theta^{v^2}-8\pi_{cc}(\pi_{cc}-2\pi_{dc} \\
& +2\pi_{dd})\theta^v+4(\pi_{cc}-\pi_{dc}+\pi_{dd})^2)L^2+8\theta^v(\pi_{cc}-\pi_{cd})((-\theta^v+1)\pi_{cd}+\theta^v\pi_{cc}-\pi_{cc} \\
& +\pi_{dc}-\pi_{dd})L+\theta^{v^2}(\pi_{cc}-\pi_{cd})^2)\beta^2-4L\alpha((\pi_{dc}-\pi_{dd})(\pi_{cc}-\pi_{cd}-\pi_{dc}+\pi_{dd})L \\
& -2(\pi_{cc}-\frac{\pi_{dc}}{2}-\frac{\pi_{dd}}{2})\theta^v(\pi_{cc}-\pi_{cd}))\beta+4L^2\alpha^2(\pi_{dc}-\pi_{dd})^2)M^4+((-8((-\theta^v+1)\pi_{cd} \\
& +\theta^v\pi_{cc}-\pi_{cc}+\pi_{dc}-\pi_{dd})(\pi_{cd}+\pi_{dc}-\pi_{dd})L^2+((-4\theta^{v^2}+16\theta^v-4)\pi_{cd}^2+(8\pi_{cc}\theta^{v^2} \\
& +(-24\pi_{cc}+16\pi_{dc}-16\pi_{dd})\theta^v+8\pi_{cc}-8\pi_{dc}+8\pi_{dd})\pi_{cd}-4\pi_{cc}^2\theta^{v^2}+8\pi_{cc}(\pi_{cc}-2\pi_{dc} \\
& +2\pi_{dd})\theta^v-4(\pi_{cc}-\pi_{dc}+\pi_{dd})^2)L-2\theta^v(\pi_{cc}-\pi_{cd})((-\theta^v+1)\pi_{cd}+\theta^v\pi_{cc}-\pi_{cc} \\
& +\pi_{dc}-\pi_{dd}))\beta^2-8((\pi_{dc}-\pi_{dd})(\pi_{cd}+\pi_{dc}-\pi_{dd})L^2+(((-\pi_{cc}+\frac{3\pi_{dc}}{4}+\frac{\pi_{dd}}{4})\theta^v \\
& +\pi_{cc}-\pi_{dd})\pi_{cd}+\pi_{cc}(\pi_{cc}-\frac{3\pi_{dc}}{4}-\frac{\pi_{dd}}{4})\theta^v-(\pi_{cc}-\pi_{dd})(\pi_{cc}-\pi_{dc}+\pi_{dd}))L \\
& +\frac{\theta^v(\pi_{cc}-\pi_{dd})(\pi_{cc}-\pi_{cd})}{2})\alpha\beta-4L((\pi_{dc}-\pi_{dd})L+2\pi_{cc}-2\pi_{dd})\alpha^2(\pi_{dc}-\pi_{dd}))M^3 \\
& +((4(\pi_{cd}+\pi_{dc}-\pi_{dd})^2L^2+8((-\theta^v+1)\pi_{cd}+\theta^v\pi_{cc}-\pi_{cc}+\pi_{dc}-\pi_{dd})(\pi_{cd}+\pi_{dc}-\pi_{dd})L \\
& +(\theta^{v^2}-4\theta^v+1)\pi_{cd}^2+(-2\pi_{cc}\theta^{v^2}+(6\pi_{cc}-4\pi_{dc}+4\pi_{dd})\theta^v-2\pi_{cc}+2\pi_{dc}-2\pi_{dd})\pi_{cd} \\
& +\pi_{cc}^2\theta^{v^2}-2\pi_{cc}(\pi_{cc}-2\pi_{dc}+2\pi_{dd})\theta^v+(\pi_{cc}-\pi_{dc}+\pi_{dd})^2)\beta^2+4\alpha((\pi_{dc}-\pi_{dd}) \\
& *(\pi_{cd}+\pi_{dc}-\pi_{dd})L^2+(((\frac{\pi_{dc}}{2}-\frac{\pi_{dd}}{2})\theta^v+\frac{\pi_{dc}}{2}+2\pi_{cc}-\frac{5\pi_{dd}}{2})\pi_{cd}-\frac{(\pi_{dc}-\pi_{dd})}{2} \\
& *(\theta^v\pi_{cc}-5\pi_{cc}-\pi_{dc}+5\pi_{dd}))L+(\pi_{cc}-\pi_{dd})((-\theta^v+1)\pi_{cd}+\theta^v\pi_{cc}-\pi_{cc}+\pi_{dc}-\pi_{dd}))\beta \\
& +((\pi_{dc}-\pi_{dd})L+2\pi_{cc}-2\pi_{dd})^2\alpha^2)M^2-4(((\pi_{cd}+\pi_{dc}-\pi_{dd})L+(-\frac{\theta^v}{2}+\frac{1}{2})\pi_{cd} \\
& +\frac{\theta^v\pi_{cc}}{2}-\frac{\pi_{cc}}{2}+\frac{\pi_{dc}}{2}-\frac{\pi_{dd}}{2})\beta+\frac{((\pi_{dc}-\pi_{dd})L+2\pi_{cc}-2\pi_{dd})\alpha}{2})\beta(\pi_{cd}+\pi_{dc}-\pi_{dd})M \\
& \left.+\beta^2(\pi_{cd}+\pi_{dc}-\pi_{dd})^{1/2}\right)+\frac{(-2\pi_{cc}+2\pi_{cd})\theta^vL\beta M^3}{2} \\
& +\frac{((((2\pi_{cc}-2\pi_{cd})\theta^v-2\pi_{cc}+2\pi_{cd}+2\pi_{dc}-2\pi_{dd})L+\theta^v(\pi_{cc}-\pi_{cd}))\beta}{2} \\
& +\frac{(-2\alpha\pi_{dc}+2\pi_{dd}\alpha)L)M^2+(((-2\pi_{cd}-2\pi_{dc}+2\pi_{dd})L+(-\pi_{cc}+\pi_{cd})\theta^v+\pi_{cc}-\pi_{cd}}{2} \\
& -\frac{\pi_{dc}+\pi_{dd})\beta+(\alpha\pi_{dc}-\pi_{dd}\alpha)L+2\pi_{cc}\alpha-2\pi_{dd}\alpha)M+(\pi_{cd}+\pi_{dc}-\pi_{dd})\beta}{2} \\
\lambda_3 =& -\lambda_2
\end{aligned}
$$

λ_2 and λ_3 are opposite signs of the same expression similar to eigenvalues of the forth equilibrium point. Either one of them is positive and the other one is negative, or both of them are equal to 0. Since stability requires both eigenvalues to be negative, this equilibrium point cannot be stable.

D THE SOURCE CODE

D.1 Phase Diagrams (Python)

```
import scipy.optimize
import matplotlib.pyplot as plt
import matplotlib.tri as tri
import numpy as np
import math

print("module loaded")

class simplex_dynamics:
    '''draws dynamics of given function and
    corresponding fixed points into triangle'''
    #corners of triangle and calculation of points
    r0=np.array([0,0])
    r1=np.array([1,0])
    r2=np.array([1/2.,np.sqrt(3)/2.])
    corners =np.array([r0,r1,r2])
    triangle = tri.Triangulation(corners[:, 0], corners[:, 1])
    refiner = tri.UniformTriRefiner(triangle)
    trimesh = refiner.refine_triangulation(subdiv=4)
    trimesh_fine = refiner.refine_triangulation(subdiv=4)

    def __init__(self,fun):
        self.f=fun
        self.calculate_stationary_points()
```

```
        self.calc_direction_and_strength()

    #barycentric coordinates
    def xy2ba(self,x,y):
        corner_x=self.corners.T[0]
        corner_y=self.corners.T[1]
        x_1=corner_x[0]
        x_2=corner_x[1]
        x_3=corner_x[2]
        y_1=corner_y[0]
        y_2=corner_y[1]
        y_3=corner_y[2]
        l1=((y_2-y_3)*(x-x_3)+(x_3-x_2)*(y-y_3))/((y_2-y_3)*(x_1-x_3)+(x_3-x_2)*(y_1-y_3))
        l2=((y_3-y_1)*(x-x_3)+(x_1-x_3)*(y-y_3))/((y_2-y_3)*(x_1-x_3)+(x_3-x_2)*(y_1-y_3))
        l3=1-l1-l2
        return np.array([l1,l2,l3])

    def ba2xy(self,x):
        ### x: array of 3-dim ba coordinates
        ### corners: coordinates of corners of ba coordinate system
        x=np.array(x)
        # print(self.corners.shape)
        # print(self.corners.T)
        # print(x)
        return self.corners.T.dot(x.T).T


    def calculate_stationary_points(self):
        fp_raw=[]
        border=1 #don't check points close to simplex border
```

```
        delta=1e-12
        for x,y in zip(self.trimesh.x[border:-border], self.trimesh.y[border:-border]):
            start=self.xy2ba(x,y)
            fp_try=np.array([])

            sol=scipy.optimize.root(self.f,start,args=(0,),method="hybr")#,xtol=1.49012e-10,maxfev=1000
            if sol.success:
                fp_try=sol.x
                #check if FP is in simplex
                if not math.isclose(np.sum(fp_try), 1.,abs_tol=2.e-3):
                    continue
                if not np.all((fp_try>-delta) & (fp_try<1+delta)):#only if fp in simplex
                    continue
            else:
                continue
            #only add new fixed points to list
            if not np.array([np.allclose(fp_try,x,atol=1e-7) for x in fp_raw]).any():
                fp_raw.append(fp_try.tolist())
        #add fixed points in correct coordinates to fixpoints list
        fp_raw=np.array(fp_raw)
        if fp_raw.shape[0]>0:
            self.fixpoints= self.corners.T.dot(np.array(fp_raw#).T).T
        else:
            self.fixpoints=np.array([])

    def calc_direction_and_strength(self):
```

```
        direction= [self.f(self.xy2ba(x,y),0) for x,y in zip(self.trimesh.x, self.trimesh.y)]
        self.direction_norm=np.array([self.ba2xy(v)/np.linalg.norm(v) if np.linalg.norm(v)>0 else np.array([0,0]) for v in direction])
        self.direction_norm=self.direction_norm
        #print(direction_ba_norm)
        self.pvals =[np.linalg.norm(v) for v in direction]
        self.direction=np.array([self.ba2xy(v) for v in direction])

    def plot_simplex(self,ax,cmap='viridis',typelabels=["A","B","C"],**kwargs):

        ax.triplot(self.triangle,linewidth=0.8,color="black")
        ax.tricontourf(self.trimesh, self.pvals, alpha=0.8, cmap=cmap,**kwargs)

        #arrow plot options:
        n=-2
        print(self.direction_norm.T[0])
        color = np.sqrt(((self.direction_norm.T[0]-n)/2)*2 + ((self.direction_norm.T[1]-n)/2)*2)
        Q = ax.quiver(self.trimesh.x, self.trimesh.y, self.direction_norm.T[0],self.direction_norm.T[1], angles='xy',pivot='mid')#pivot='tail')#


        ax.axis('equal')
        ax.axis('off')
        margin=0.01
        ax.set_ylim(ymin=-margin,ymax=self.r2[1]+margin)
        ax.set_xlim(xmin=-margin,xmax=1.+margin)
```

```

        if self.fixpoints.shape[0]>0:
            ax.scatter(self.fixpoints[:,0],self.fixpoints[:,
             ↪  1],c="black",s=70,linewidth=1, marker="x")
        #fig.colorbar(timescatter,label="time")
        ax.annotate(typelabels[0],(0,0),xytext=(-0.0,
         ↪  -0.02),horizontalalignment='center',va='top')
        ax.annotate(typelabels[1],(1,0),xytext=(1.0,
         ↪  -0.02),horizontalalignment='center',va='top')
        ax.annotate(typelabels[2],self.corners[2],
         ↪  xytext=self.corners[2]+np.array([0.0,0.02]),
         ↪  horizontalalignment='center',va='bottom')

#initialize simplex_dynamics object with function
dynamics=simplex_dynamics(f)

#plot the simplex dynamics
fig,ax=plt.subplots()
dynamics.plot_simplex(ax, typelabels=["A", "V", "S"])
plt.savefig('1_simplex_b=1.jpg', dpi=300, frameon=True)
plt.show()
```

D.2 Numerical Simulation of Population Dynamics (Python)

```
import matplotlib.pyplot as plt
import numpy as np
from numpy.linalg import eig
import math
from sympy import *

xA = [0.33]
xS = [0.33]
xV = [0.34]
```

```
xT = [0.5]
dt = 0.01
alpha = 0.05
beta = 0.02
pcc= 3
pdc= 5.5
pcd= -10
pdd= 1
theta_a = 1
theta_v = 1

for t in range(60000):
    uT = (1- xS[t])*pcc + xS[t]*pcd
    uD = xA[t]*pdc + (1 - xA[t])*pdd
    uM_trust = xT[t] * uT + (1 - xT[t]) * uD
    Trust = xT[t]  +  alpha * xT[t] * (uM_trust - pdd) * dt


    uA = xT[t]*pcc +(1-xT[t])*pcd + theta_a*(xT[t]*pcc +
    ↪  (1-xT[t])*pdc)
    uS = xT[t]*pdc +(1-xT[t])*pdd
    uV = xT[t]*(1 + theta_v - xT[t] * theta_v) * pcc + (1 -
    ↪  xT[t]) * (pdd - xT[t]*theta_v * pcd)
    uM_rec = xA[t] * uA + xS[t] * uS + xV[t] * uV
    fA = beta * xA[t] * (uA - uM_rec)
    fS = beta * xS[t] * (uS - uM_rec)
    fV = beta * xV[t] * (uV - uM_rec)

    if Trust < 0:
      Trust = 0
    elif Trust > 1:
      Trust = 1
    else:
```

```
        Trust = Trust

    xT.append(Trust)
    xA.append(xA[t] + fA * dt)
    xS.append(xS[t] + fS * dt)
    xV.append(xV[t] + fV * dt)

plt.plot(xA, 'g', label = 'altruist')
plt.plot(xS, 'b', label = 'selfish')
plt.plot(xV, 'r', label = 'vulnerable res')
plt.plot(xT, 'cyan', label = 'trust')
plt.title('Population Dynamics')
plt.legend(loc='best')
plt.grid()
```

D.3 Numerical simulation of agent-based model (Julia)

```

using Clp
using DataFrames
using Distributions
using LinearAlgebra
using GameTheory
using Ranges
using Random
using Agents
using DrWatson: @dict
using InteractiveDynamics
using Plots
using StatsPlots
using CSV

a = 4
```

```
b = 4
random_trust = Beta(a, b)
average_trust = a / (a+b)
Personas = [1, 2, 3]

mutable struct Agent <: AbstractAgent
    id::Int                  # The identifier number of the agent
    pos::NTuple{2,Float64} # The x, y location of the agent on a 2D grid
    persona::Int             # Type of agent: 1=Altrust, 2=Vulnerable-Responsive, 3=Selfish
    trust::Float64           # Between 0 and 1. 0 means no trust while 1 means perfect trust
    energy::Float64          # Cumulation of payoffs
    last_ft_action::Int
    last_st_action::Int
    last_payoff::Float64
    turn_order::Int
    summary::Vector{Float64}
end
function model_initiation(;
numagents = 240,
pcc = 5,
pdc = 6,
pcd = -5,
pdd = 0,
mutation_rate = 100000,
extent = (2, 2),
spacing = 1,
seed = 3,
average_trust = 0.5,
trust_parameter = 0.1,
param_a =1.0,
```

```
param_v =1)

rng = Random.MersenneTwister(seed)
properties = @dict(
    numagents,
    pcc,
    pdc,
    pcd,
    pdd,
    mutation_rate,
    average_trust,
    trust_parameter,
    param_a,
    param_v)

space2d = ContinuousSpace(extent, spacing)
model = ABM(Agent, space2d; properties, rng, scheduler,
↪   Schedulers.Randomly())

for n in 1:numagents
pos = random_position(model)
persona = Personas[mod(n-1, 3)+1]
trust = rand(random_trust, 1)[1]
last_ft_action = 0
last_st_action = 0
last_payoff = 0.0
turn_order = 0
energy = 0.0
agent = Agent(n,pos,persona,
↪   trust,energy,last_ft_action,last_st_action,
↪   last_payoff,turn_order,[persona, trust]
)
add_agent!(agent, model)
```

```
end
return model
end

function model_step!(model)

for (a1, a2) in list_pairs(model)
    game!(a1, a2, model)
end

for a in allagents(model)
mu_rate = model.mutation_rate
self_payoff = a.last_payoff
random_opponent = random_agent(model)
opp_payoff = random_opponent.last_payoff
imit_prob = 1 / (1 + exp((self_payoff - opp_payoff)/mu_rate))
if a.turn_order == 1
a.trust = a.trust + model.trust_parameter * a.trust *
 ↪  (a.last_payoff - model.pdd)
if a.trust > 1
a.trust =1
elseif a.trust<0
a.trust =0
end
end

if a.turn_order == 2
if imit_prob > 1 - rand()
a.persona = random_opponent.persona
end
end
end
end
```

```
function list_pairs(model)
    pair_list = []
    shuffled_list = shuffle([a for a in allagents(model)])
    for a in 1:2:model.numagents
        push!(pair_list, (shuffled_list[a], shuffled_list[a+1]))
    end
    return pair_list
end
function game!(agent1::Agent, agent2::Agent, model)

pdd = model.pdd
pcd = model.pcd
pdc = model.pdc
pcc = model.pcc

    agent1.turn_order = 1
    agent2.turn_order = 2

    if agent1.trust > 1 - rand()
        agent1.last_ft_action = 1 # Trust
    else
        agent1.last_ft_action = 2 # Distrust
    end


    if agent2.persona == 1
        agent2.last_st_action = 1 # CU
    elseif agent2.persona == 2
        agent2.last_st_action = 2 # CC
    else
        agent2.last_st_action = 3 # DU
    end

```

```

    if agent1.last_ft_action == 1 && agent2.last_st_action == 1
        agent1.last_payoff = pcc
        agent2.last_payoff = pcc + model.param_a * (pcc)
    elseif agent1.last_ft_action == 1 && agent2.last_st_action ==
    ↪ 2
        agent1.last_payoff = pcc
        agent2.last_payoff = pcc + model.param_v * (pcc -
        ↪ model.trust_parameter * pcc)
    elseif agent1.last_ft_action == 1 && agent2.last_st_action ==
    ↪ 3
        agent1.last_payoff = pcd
        agent2.last_payoff = pdc
    elseif agent1.last_ft_action == 2 && agent2.last_st_action ==
    ↪ 1
        agent1.last_payoff = pdc
        agent2.last_payoff = pcd + model.param_a * (pdc)
    elseif agent1.last_ft_action == 2 && agent2.last_st_action ==
    ↪ 2
        agent1.last_payoff = pdd
        agent2.last_payoff = pdd
    else # elseif agent1.last_ft_action == 2 &&
    ↪ agent2.last_st_action == 3
        agent1.last_payoff = pdd
        agent2.last_payoff = pdd
    end
end

isaltruist(a) = a.persona == 1
isvresponsive(a) = a.persona == 2
isselfish(a) = a.persona == 3
isfirstmover(a) = a.turn_order == 1
issecondmover(a) = a.turn_order == 2
```

```
isfmcooperative(a) = a.last_ft_action == 1 ? true : false
issmcooperative(a) = a.last_st_action == 1 ? true : false
adata = [
(:last_payoff, mean),
(isaltruist, count),
(isvresponsive, count), #
(isselfish, count), #
(isfmcooperative, count, isfirstmover),
(issmcooperative, count, issecondmover),
(:trust, mean)
# (aggro, mymean, clb),
# (fdr, count)
]

mdata=[
:param_a,
:trust_parameter
]
models = [model_initiation(trust_parameter=c, param_v=v,
↪  mutation_rate=m, seed=t) for c in [0.1;0.5] for v in
↪  [0.8;0.05;1.0] for m in [0.01;0.1;1] for t in 1:10]

@time begin

adf, mdf = ensemblerun!(models, dummystep, model_step!, 50;
↪  adata, mdata)
end
merged = innerjoin(
  adf,
  mdf,
  on = [:step, :ensemble])
describe(merged)
merged2 = merged[merged.ensemble .== 100, :]
```

```
describe(merged2)
rename!(merged2,:count_isaltruist => :Altruist)
rename!(merged2,:count_isvresponsive => :VulnerableResponsive)
rename!(merged2,:count_isselfish => :Selfish)

@df merged2 plot([:Altruist :VulnerableResponsive :Selfish])
# rename!(merged2,:mean_trust => :MeanTrust)

@df merged2 plot([:MeanTrust], label="MeanTrust")
```

CURRICULUM VITAE

Özlem YAŞAR

E-mail: ozlemyasar@protonmail.com

Date of Birth: 28/11/1994

Place of Birth: Merkez/Çorum

Address: Altınova Mah., Kazım Ayvaz Sk., No:12, Daire:3, Süleymanpaşa/Tekirdağ.

Education

Master's Degree Galatasaray University - Economics (2018-2022)

Bachelor's Degree: Galatasaray University - Mathematics (2012-2017)