UNIVERSITATEA BABEŞ-BOLYAI
BABEŞ-BOLYAI TUDOMÁNYEGYETEM
BABEŞ-BOLYAI UNIVERSITÄT
TRADITIO ET EXCELLENTIA

Dynamics in social systems: a computational physics approach

Ph.D. thesis summary

A DISSERTATION PRESENTED
BY
LEVENTE VARGA
TO
THE DEPARTMENT OF PHYSICS

SCIENTIFIC SUPERVISOR
PROFESSOR DR. ZOLTÁN NÉDA

BABEŞ-BOLYAI UNIVERSITY
CLUJ-NAPOCA, ROMANIA

2019

Thesis advisor: Professor Dr. Zoltán Néda Levente Varga

Dynamics in social systems: a computational physics approach

Abstract

Social systems are investigated using a wide range of computational and statistical physics tools and methods. We present studies of evolutionary game theory, human mobility, and citation networks to examine dynamics in complex social systems.

After a general introduction in Chapter 1, we study strategy distributions in case of multiagent spatial evolutionary games, during which the players are located on different types of networks and change their strategies using imitation or logit strategy update rules. Players can interact with a certain number of neighbors, which leads to cooperation, defection, or invasion within the investigated system.

In Chapter 2, we analyze the universal laws of commuting and human mobilities. We examine various road, air and commuting networks across Hungary, Italy, Europe, USA and worldwide. We investigate the averaged apparent speed as a function of travel distance by processing the data over these transportation networks. We also study the distribution of commuter fluxes depending on population density by using census datasets, and job opening information. In addition, we use different radiation type models to evaluate and explain the obtained results.

Finally, in Chapter 3 we argue on the necessity of field-based normalization for comparing scientific production in different research areas. Individual indicators of various scientific articles can be different within a wide range of disciplines, such as physics and mathematics, where the number of citations is very diverse. Studying benchmarks and publication networks, we present some clustering methods that help in separating scientific fields. We also report the presence of universal scaling rules in scientific publications and Facebook posts.

Keywords : universalities, scaling, social systems, complex systems, dynamics, computational physics, statistical physics, evolutionary games, human mobility, commuting patterns, networks, citation networks, social networks

Ph.D. thesis contents

Contents

Introduction

The social system – according to Parsons, who has given the concept of "system" in modern sociology – is an orderly arrangement with an interrelationship of parts, where every part has a fixed place and definite role to cooperation [1]. Examples of everyday social systems are families, communities, nations, industries, which depend on diverse shared characteristics such as location, cultural norms, religion, socioeconomic status, etc.

Social systems are known examples of complex systems with a lot of exciting research areas. Several electronically available datasets help researchers reveal universalities and test their models [2]. In the light of that, we process available datasets, simulate experimental data and develop models for identifying and studying the dynamical behavior in systems. The main domains of physics used in such studies are computational physics together with evolutionary game theory, statistical physics, and complex network theory analysis. We study phenomena such as cooperation, competition, commuting, traveling, clustering, normalizing, and fitting. Besides these, we also focus on solving problems of strategy invasion, human mobility, transportation, citation, and visualization.

In the present thesis, we offer new models, approaches, and data analysis methods in the field of social system dynamics. We discuss and present our scientific results in the three research areas listed above.

First area is the evolutionary game theory, which is an application of the mathematical theory of games by placing population in a biological context [3]. The players are imitating some evolutionary rules considering the Darwinian selection or varying their strategies following individual rationality of social systems. Players of games are placed typically on a lattice or network, and they interact with a certain number of neighbors.

The methods of statistical physics are directly applied to evolutionary games, e.g., two-dimensional Ising model in a magnetic field with up and down spins [4]. The spin states represent the players' strategies, and Glauber dynamics allows one-site spin (strategy) flips with a calculated probability [5]. Probabilities are calculated for the imitation or logit strategy update rules, and during the games, players try to maximize their income.

Another important ingredient of statistical physics in the evolutionary games is the Monte Carlo simulations [6]. Using this type of simulations, we can quickly reproduce on computer the played games.

In the present thesis, the social systems provided by evolutionary game theory are examined by measuring the distributions of different strategies. During our examinations, we can observe the shapes of the domains of cooperation or competition for players, who follow the same strategic goal, for example.

Second area is the human mobility, which is important in many social systems. If we think about the daily traffic or even economic and urban development, it is important to understand first the operation of the mobility world around us [7]. The everyday journeys of individuals are also significant, as presented in the study of González et al. [8]. Individuals often return to a few highly frequented locations, and their trajectories indicate good traceability.

Some popular starting points in the study of human mobility are census datasets with commuting routes, population density, road and air networks, or travel times and delays in nodes. These data can be analyzed with the methods of statistical physics and understood in the framework of various models: such as the gravity model [9], the radiation model and the radiation model with selection [10, 11].

In the present study, two further generalizations of the original radiation model are given: one is the travel cost optimized radiation model where one takes into account for the selection of commuting jobs also the involved travel costs [12]. The other one is the flow and jump model which is based on a simple master equation [13].

In the present society, in order to get a better traffic optimization, for choosing the suitable transportation mode, or for the selection of the appropriate job, it is important to understand mobility and commuting patterns. In our investigation, we use models from statistical physics to understand the human mobility in social systems.

Third area is the network representations of complex systems in particular social systems, which are widespread in social science, information theory, neuroscience, biology, etc. [14, 15]. Connections between elements by the nodes of networks and the edges between them is a helpful visualization for the core of any complex systems.

For understanding the complex world of scientific networks, one can use both citation of collaboration and social networks. Computational physics approaches together with classical network processing methods, such as visualization, clusterization or statistical distributions, help in the analysis of structural features of networks [15]. Identifying communities is relevant for an easier processing and for understanding better the topology of complex networks.

Computer programming combined with statistical physics methods offer a variety of frameworks, software tools, and algorithms for network analysis. By using these tools, we can identify communities, subnetworks, and structures for benchmarking and comparing real-world networks.

There are many ways to compare the communities created by the clustering of networks. A great example for comparison is the cross-field normalization of scientometric indicators [16–18]. The distribution of studied indicators from each scientific areas can be normalized using a power-law-like function [19]. After normalization, a comparison can be made on the same scale between calculated indicators from two or more different fields, such as physics and mathematics, where the given indicators can be offset by a proportionality constant. Similarly, the distribution of various examined areas, such as citation and Facebook data, or publications of a selected journal or author can be described by Tsallis-Pareto distribution, which is a proper probability density function with a power-law-like tail [20].

The thesis is organized in three extended chapters, according to the studied problematics: in Chapter 1 the Monte Carlo simulations, the theoretical calculations, and the obtained results of evolutionary games are discussed. In Chapter 2 we present the data analysis of human mobility and commuter fluxes. Here we discuss the power of some simple models in fitting the experimental data. In Chapter 3 the community detection, the normalization of individual scientific indicators, and the similarity of scientific publications with Facebook posts are discussed. Finally, at the end of the thesis, we give a summary of the obtained results.

"Evolutionary game theory is a way of thinking about evolution at the phenotypic level when the fitnesses of particular phenotypes depend on their frequencies in the population."

John Maynard Smith

1 Evolutionary games

In general, we study multiagent spatial evolutionary games. On these systems, the players are placed on a lattice or network, and they interact with a certain number of neighbors, as shown in Figure 1.1. During such a game at each time-step individuals participate in all potential games choosing one possible strategy. Their pair interaction with the selected neighbors is represented by the payoff matrix:

$$\mathbf{A} = \begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix} \tag{1.1}$$

In the above example (1.1), the $\mathbf{A}$ matrix shows what is the players' gain if they select one of the possible strategies. The evolutionary games can be deterministic or stochastic with synchronized or random sequential dynamical update rule, depending on imitation or myopic strategy selection to achieve the higher individual gain.

The imitation update method is used in analogy with biological processes, when the players can imitate their more successful neighbors. The myopic selection update rule is the so-called logit rule from economy and physics models [21, 22]. In this situation, each player can calculate their payoff, and they prefer to choose a higher income strategy depending on the expected payoff values.

To analyze the evolutionary games, researchers consider very often Monte Carlo (MC) type simulations on networks of site N. In most of the cases, the simulation starts from a random initial condition, but it may also happening, that the initial condition is a predefined pattern. During the simulations, after a suitable thermalization time t_{th} we start measuring the statistical data over a sampling time t_s. These values are adjusted to the system's parameters, and we usually use periodic boundary conditions.

During the MC simulations, we have used the dynamical cluster methods or generalized mean-field approximations [5]. For these methods, we calculate all the configuration probabilities in a way that we numerically solve a set of equation of motions, and as a result, we can give analytical predictions for the studied quantities.

Using the above methods, in the thesis, we study a few questions from evolutionary game theory and discuss the obtained results. First, we present the matching-pennies game on two types of networks. Then we consider an evolutionary rock-paper-scissors game with synchronized strategy updating. The relationship between the strategies of an extended spatial evolutionary prisoner's dilemma game is also discussed.

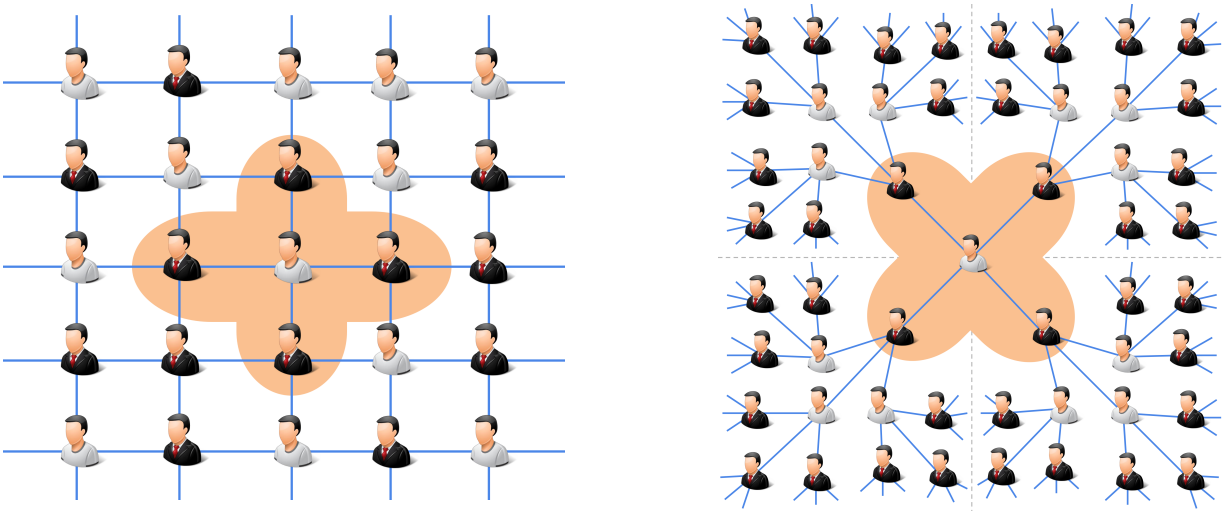

Figure 1.1: Evolutionary game theory players with interacting neighbors on a two-dimensional square lattice (left), and on a loop-free Bethe lattice (right).

Here the introduced win-stay-lose-shift strategy evaluates the outcome of the last round, and by changing their strategies players can improve the gain. Lastly, we have studied invasion processes on a square lattice in case of two-strategy evolutionary games.

And now a little bit in more details: **First**, we analyze an evolutionary game, known as the matching-pennies game, where two types of players, X and Y, are located on bipartite networks [23]. We study these evolutionary games on a square lattice, and on a bipartite loop-free Bethe lattice as seen on Figure 1.1. In both cases, each player of type X interacts with her four connected neighbors of type Y like the white and black squares on a chessboard. On left subfigure of Figure 1.1 is visible a random player distribution regarding the type of players, but on right subfigure of Figure 1.1 each player of the X type is connected to her four neighbors of Y type (the white player from the center is connected her four neighbors black players).

Using these networks and the two types of players, we examine the effects of matching-pennies interactions, and then we study the spatial distribution of the two choices called heads (H) and tails (T). However, how does this game work? In the traditional matching-pennies game, the two players first agree who will be the winner if the sides of the coins are the same or different. Then they conceal a coin in their palms with the side heads or tails upward and reveal their choices simultaneously. The winner receives the opponent's penny. Player of the X type wins with (H,H) or (T,T) strategy choices and player of the Y type wins with (H,T) or (T,H) strategy combinations.

Therefore, we study the strategy distribution by the effect of the matching-pennies game for two types of networks when varying the noise level in the Glauber type dynamical rule [24]. As you can see an example snapshot on Figure 1.2, we measure the presence of the strategies like the white (H) and black (T) boxes on the figure.

The studied systems are analyzed by performing MC simulations on square lattices of $N = L \times L$ sites with periodic boundary conditions and bipartite random regular graphs of site N. During the simulations, N is varied from $N = 2.5 \times 10^5$ up to 4×10^6. The simulation started from a random initial

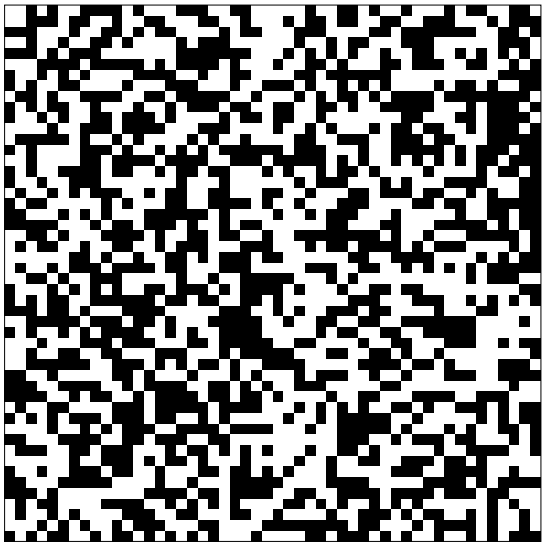

Figure 1.2: Strategy distribution in case of the matching-pennies game. The 50×50 snapshot presents a white (H) and black (T) boxes distribution.

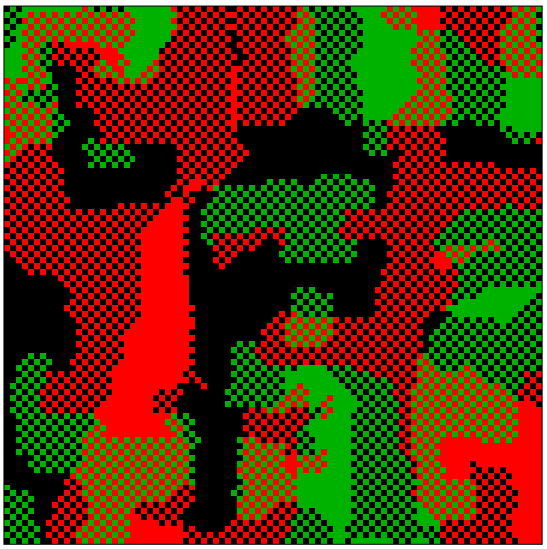

Figure 1.3: A typical snapshot about the spatial distribution of strategies in a block of 90×90 sites of a square lattice at noise $K = 0.4$. Black color stands for the R strategy, red for P, while green indicates the S strategy.

state, and after a thermalization time t_{th}, the statistical data are obtained by averaging over a sampling time t_s. This t_{th} value is changed typically between 2 000 and 10^6 Monte Carlo steps (MCSs) and the t_s value is modified from 50 000 to 3×10^6 MCSs where within the time unit (MCS), each site has in average a chance to modify its state.

The MC simulations have confirmed quantitatively the absence of correlations between the nearest neighbor sites. Nevertheless, we found weak negative correlations between the second and third order neighbors.

After the simulation results, we review the theoretical predictions. From these computed results, we determined the values of the correlation function again. Then we made the comparison to the results of the MC simulations.

The studies are repeated on a bipartite random regular graph for the same number of neighbors $k = 4$. This was necessary for the analysis of the topological features of the connectivity network. The most striking differences are the smaller magnitudes of the correlations on the bipartite random regular graphs and their absence between the third neighbor sites.

Second, we studied a spatial evolutionary rock-paper-scissors game on a square lattice where the player obtains their payoff from the games played with their four neighbors [25]. They use synchronized strategy updating according to the logit rule. For the traditional RPS game, players simultaneously choose one of three strategies named as rock (R), paper (P) and scissors (S). According to the rules of the game, a strategy is superior to another and inferior to the third one: rock beats scissors, scissors beat paper, and paper beats rock, i.e., the game represents the cyclic dominance of strategies.

Systems with RPS interaction show a global oscillation in the strategies frequencies. Generally, in a structured population when the players are placed on a network, the oscillation evolves towards a limit

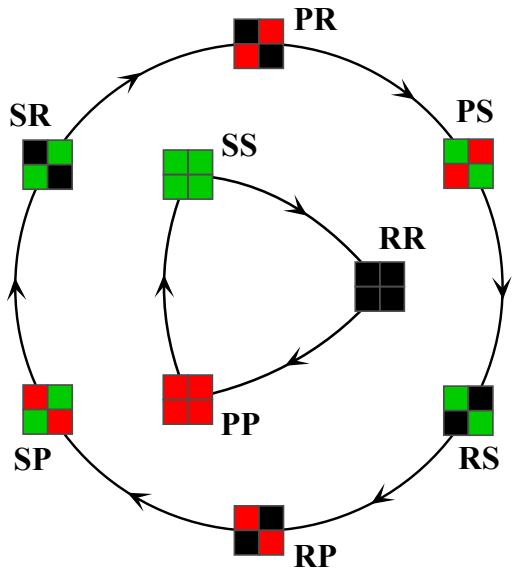

Figure 1.4: Flow diagram for the cyclic behavior of domains. Solid (black) lines show how the domain composition changes in every generation due to the logit strategy adoption rule. Color code is the same as in Figure 1.3. Domain types are denoted using the appropriate strategy names. Note the different names for the chessboard and anti-chessboard structures. Pure domains return to the same state in three generations, chessboard-like associations accomplish this in six generations. Interestingly, the nine available strategy associations form two disjoint cycles instead of a full cycle.

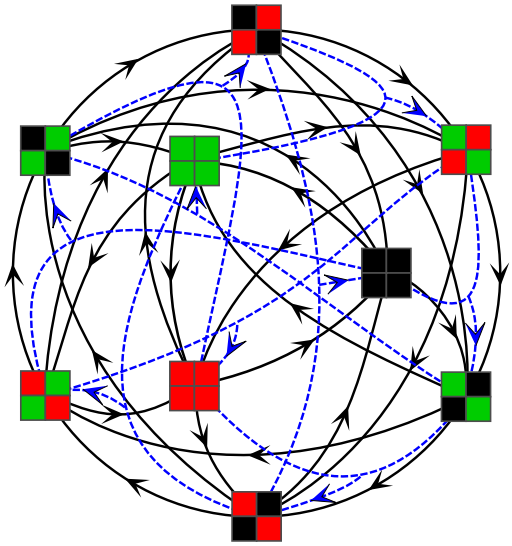

Figure 1.5: Flow diagram for the invasion and speciation processes between the different strategy associations (domains). Arrows on the solid (black) lines show the winning association when the two associations connected by the black line meet each other. Dashed (blue) lines display the cases when the encounter of two associations results in the appearance of a third species. The arrow points toward the newly established association. The meaning of the strategy colors is the same as in Figure 1.3.

cycle or its size increase until only one strategy remains alive. In the two-dimensional case, numerical simulations show the survival of all three strategies by self-organizing patterns.

Using this system's oscillations within these strategies domains, we studied a nine-species competitive relationship, where there are three pure R, P, S strategies and six other mixed strategy species with a chessboard-like structure, where the black and white squares of the chessboard are occupied by two different strategies, as can be seen on Figure 1.4. The players are located on a square lattice with periodic boundary conditions, and they collect their income from the played RPS games with nearest neighbors.

We studied the model by performing MC simulations on a square lattice with $N = 4 \times 10^4$ up to 10^6 sizes. Usually, the simulation started from a random strategy state, but in case of invasions, we applied predefined initial conditions too. We also used a transient or thermalization time $t_{th} = 1\,000 - 5\,000$ MCSs for the initial stabilization of the system. Over the $t_s = 10^4 - 10^6$ sampling time we measured the simulation data.

Starting from a random initial state, we observed a cluster-formation process that indicates a synchronous oscillation at low noise levels. During the domain-growing process, a self-organized pattern is created in the system. An example snapshot is shown in Figure 1.3, where all nine types of strategy can be distinguished.

Following the update rules, players choose the strategy providing the highest income. First, we consider the homogeneous domains, which are denoted by the same strategy pairs RR, PP and SS positioned on the two sublattices. Here we can observe a cyclic behavior, for instance, in $t = 0$ time all players use strategy R, then in the next $t = 1$ everyone will choose P and afterward in $t = 2$ all strategy is S. These cyclic changes are illustrated in the inner ring in Figure 1.4.

In the second case, these pairs can be different. For example of mixed strategies, the RP denotes a uniform dispersion of R and P strategies on the two sublattices. Within this domain, the P players do not wish to change their strategies, but the disaffected R players are enforced to modify their strategies to S. Consequently, the system evolves into the SP state. In any case, this mixed cyclic behavior is likewise shown in the outer ring in Figure 1.4.

Depending on the noise K value, we separate two cases. At low noise levels, the simulation results show that the point defects stay inside the defined homogeneous and mixed cyclic patterns. As the value K increases, at high noise levels, we can observe small islands of other states inside the outlined domains. This reasonably complex structure of strategy pairs and invasions is visible in Figure 1.5.

In the following let's look at the meaning of the intertwined invasion and speciation relations between the different domain types denoted by the black and blue dashed lines with arrows. The black lines with the arrow on them indicate the direction of these invasions. As an example, we can see that the RS strategy pair invades the RP pair. The blue dashed lines with arrows on them represent the collision of two strategy domains, which gives birth to a third one and invades the initial strategy pairs.

Summing up the above two cases, we can mention a third case, when the third emerged strategy pair does not invade any of the initial strategies. In this situation, the new strategy pair plays as a catalyst and goes through the system in an invader-defender role, but finally, the superior domain consumes him as well.

Third, we tried to map the evolution of cooperation with the assistance of a three strategy evolutionary prisoner's dilemma game [26]. Traditionally, during the prisoner's dilemma (PD) game, two participating players choose between cooperation (C) and defection (D) strategies. Similarly with this, in our model, players can adopt three strategies: always-cooperating (AllC), always-defecting (AllD) and the newly introduced win-stay-lose-shift (WSLS) strategy. WSLS cooperates in the first round and in the subsequent game she evaluates the last rounds outcome. Then he/she changes his/her strategy if the given payoff is smaller than a defined threshold.

In a one-shot game, the players select between cooperation (C) and defection (D) and earn different payoffs depending on their simultaneous decisions. The mutual cooperation's result is the reward R, and the defection's income is the punishment P. A defector exploiting a cooperator earns the temptation T while the victim of the exploitation receives the sucker's payoff S. The PD satisfy the $T > R > P > S$ ranking. According to the generality, we can fix the reward ($R = 1$) and the punishment ($P = 0$) values remaining with two payoff parameters.

We study our spatial evolutionary model on a square lattice with four nearest neighbors. Players play an iterated PD (IPD) game. In addition to the evolutionary games discussed above, we study both (imitation and logit) update mechanisms in this case.

On an $N = L \times L$ sized square lattice, players can use AllD, AllC, and WSLS strategies. However,

at the WSLS strategy we established a friendliness parameter w in between $R = 1$ and $P = 0$, which defines the probability of cooperation in the first round; accordingly, high w can be associated with a friendly behavior. In most cases, this parameter is set to $w = 0.5$, but we analyze different $w = 0.1$ and 0.9 values as well.

In our MC simulations, we examine the model on a square lattice characterized by a periodic boundary condition. The used system size is $N = 200 \times 200 = 40\,000$ limited by our calculation capacity. In one MCS, each player has the option to change her strategy once on average. For the phase diagrams, stationary strategy concentrations were obtained by averaging the strategy distribution. The simulation runs over a transient time of $t_{tr} = 50\,000$ MCSs and averaging happens over $t_s = 10\,000$ MCSs sampling time.

Using the $T > R > P > S$, $R = 1$ and $P = 0$ relations, we studied analytically the potential game character of the game in case of $w = 0.5$ by the logit rule update. We studied the interaction and equivalence between three AllD, AllC and WSLS strategies.

We simulated the theoretical results, and we observe the competition between the AllD and WSLS strategies. It is visible that the dominated AllC strategy is present in an extremely low frequency and practically does not disturb the competition of AllD and WSLS.

We studied also the $w \neq 0.5$ friendliness parameters for $w = 0.1$ and 0.9. In case of $w = 0.1$, the area of the unfriendly WSLS strategy has widened, moreover, in the case of $w = 0.9$, the territory of the friendly WSLS strategy has narrowed. This parameter selection is what we expect if one wants to survive better in addition to the AllD strategy.

Next, we discuss simulation results using the imitation update rule. The obtained results are different from the results of the logit rule. One of the important factors that we should mention in case of imitation is that the parameter w no longer pays a role. A significant difference is that the WSLS strategy is widening. These simulations prove the positive impact of the imitation update of cooperative WSLS attitude.

Last, we investigate the invasion processes by changing the strengths of the self-dependent and cross-dependent components [27]. Referring to the whole $T - S$ parameter space, here we restrict our investigation to the region of $T < 1$ and $S < 0$.

We study a two-strategy evolutionary game with imitation update rule. Players are located on a square lattice, and they interact with their nearest neighbors. We examine here the interface motion and invasion velocities for the different homogeneous phases.

During our simulations, we used a square lattice with $N = L \times L$ sites, where equivalent players choose one of the two (1 or 2) strategies. The system has periodic boundary conditions. Players play matrix games with their four nearest neighbors.

We applied imitation strategy update with a small change. Here we choose two neighboring players, and the selected player follow, i.e., adopt her neighbor's strategy with the a calculated probability.

All of our simulations are started from an artificial initial state in which one of the strategies forms a circular island in the sea of the other strategy. In the case when the strategy 1 (black) is the sea and the strategy 2 (white) is the circle island, we can see a domain shrink and its disappearance at the horizontal and vertical interfaces. In the opposite situation if the two strategies shifts the sea and island roles, there

is also a domain that reduces and shrinks, but now it is surprisingly along the ± 1 tilted interfaces with different and faster contraction.

We also studied the invasion behavior when the square lattice was replaced by other two-dimensional lattices. Practically, we observed the same attitude of players and strategies. Differences in the orientation of invasions were observed only when the players collected payoffs from a game with their nearest or next-nearest neighbors.

"We live in a world where great incompatibles co-exist: the human scale and the superhuman scale, stability and mobility, permanence and change, identity and anonymity, comprehensibility and universality."

Kenzo Tange

2

Human mobility

Similarly with other human mobility studies, here we are also dealing with understanding and revealing universal laws that govern human mobilities. More specifically, we propose to analyze different modes of transportation and the distribution of the mobility fluxes between the settlements.

In the first part of our study, we analyze several human transportation networks such as the road-network of Hungary and the interstate road-network of continental USA, the direct flights between the major airports of Europe and the air travel network of USA.

Examining the traveling time as a function of the travel distance, we study how the traveling speed increases with the travel distance due to the travel time lost in the main hubs, the structure of the travel networks and the speed limit of the roads and vehicles. We investigate this speed as a function of the travel distance, i.e., geodesic distance and the driving distance, observing a significant difference.

Another important problem in the field of human mobility is the daily commuting. In most cases, the raw data originate from census data, which contain a lot of useful informations, like the home place and workplace fields of the commuters.

Using these data and the spatial distribution data for population density, we investigate commuter fluxes at different distances. We are processing commuting and population databases from Hungary, Italy, and the USA to measure the distribution of mobility fluxes. Then, we compare existing, and novel models for fitting the averaged commuter flux data. Such existing models are the original Radiation Model (RM) [10], the Gravity Model (GM) [9] and the generalized Radiation Model with Selection (RMwS) [11], which have been successfully used for understanding the job selection for the individuals. Moreover, we offer two new generalizations of the radiation model, one with the travel cost optimization method and the other with application of a master equation allowing both jump and local flow processes.

First, we revealed a universal relation between the average speed of the mobility and the travel distance [28]. The travel is an integral part of our lives, and everybody learned that the traveling time does not scale linearly with the travel distance [7]. It is well known for example that sometimes it takes several hours to travel at distances of a few hundreds of kilometers, but not significantly more to travel at the opposite side of the Earth.

To illustrate this idea, Figure 2.1 shows the roughly estimated distance and velocity scales for different human traveling modes from walking up to cosmic journeys. The overall trend is a power-law with an average 0.5 exponent on the distance and velocity scales.

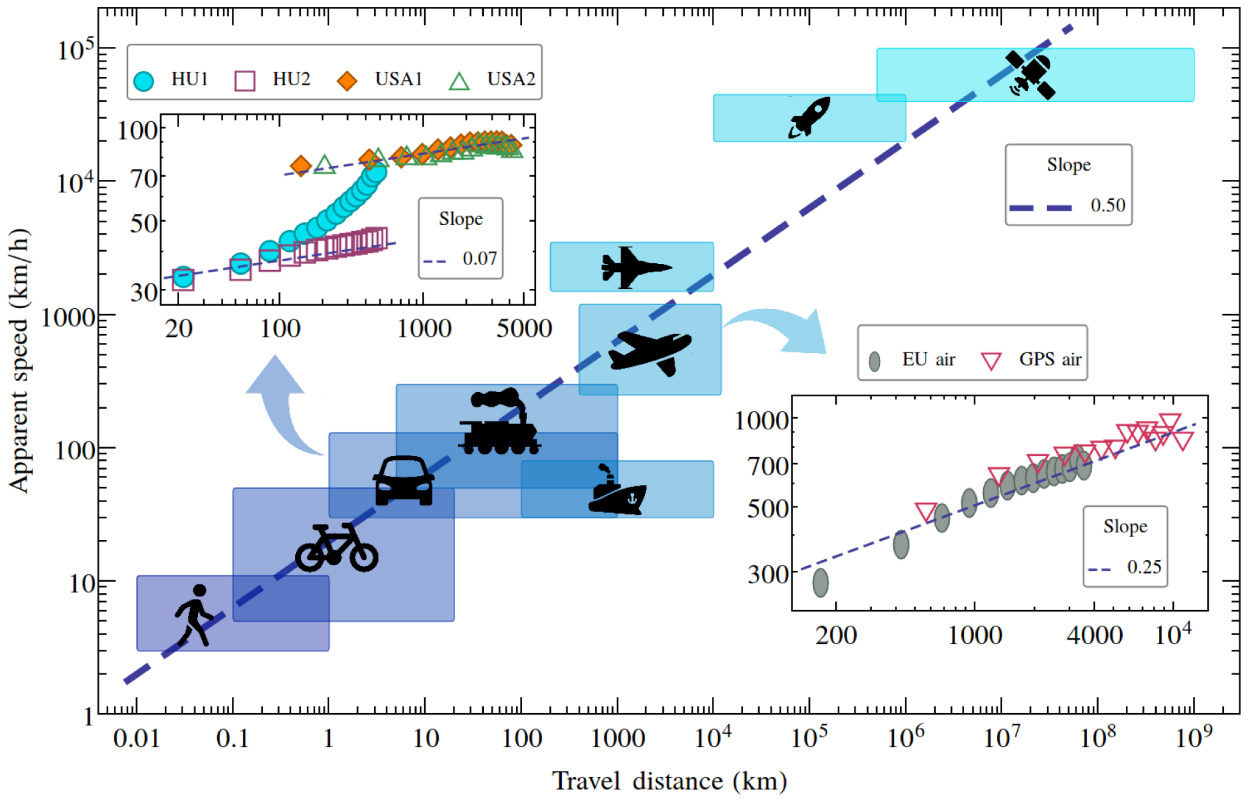

Figure 2.1: Velocity and distance scales of human travel. The apparent travel speed (estimated as the traveled distance on the geodesic line divided by the travel time) as a function of the travel distance. Boxes indicate intervals for different traveling modes. The two inset figures present some averaged results on the two most popular traveling modes: car and air travel. The dashed lines in these insets indicate power-law trends with the specified exponent. Dashed lines with different slopes are not fitting results; they indicate power-law trends with the specified exponents only for guiding the eyes. Please note the logarithmic axes.

Additionally, a similar power-law-like is founded inside the small boxes of Figure 2.1, which represent the various transportation modes separately. In our investigation, we focus on the two most popular traveling modes, the car, and the air travel. In both cases, we obtained the same sublinear travel time increase as a function of the travel distance like in the overall big picture. For example, if we travel from one city to another one that is far away for a large part of the travel we use highways where the average speed is high. However, when we drive out from a city, there are many stops, traffic jams reducing the average traveling speed drastically for short travels. For air travel, the smaller distances are served with smaller planes with lower traveling altitude and smaller cruising speed, and the average traveling speed is greatly reduced for short travels, because of the takeoff, landing and parking maneuvers.

As shown in the two inset frames of Figure 2.1, we collected and analyzed traveling data of country and highway roads, and direct flights between airports. Results for car travel were calculated on the country, highway and interstate road-networks of Hungary and the continental USA. The spatial distribution of the locations and the structure of the road-network is shown in Figure 2.2. In the case of Hungary, we measured travel between cities of 174 statistical subregions on the territory of the country [29]. Here we are distinguished results with and without allowed highway travel (maps HU1 and HU2, respectively).

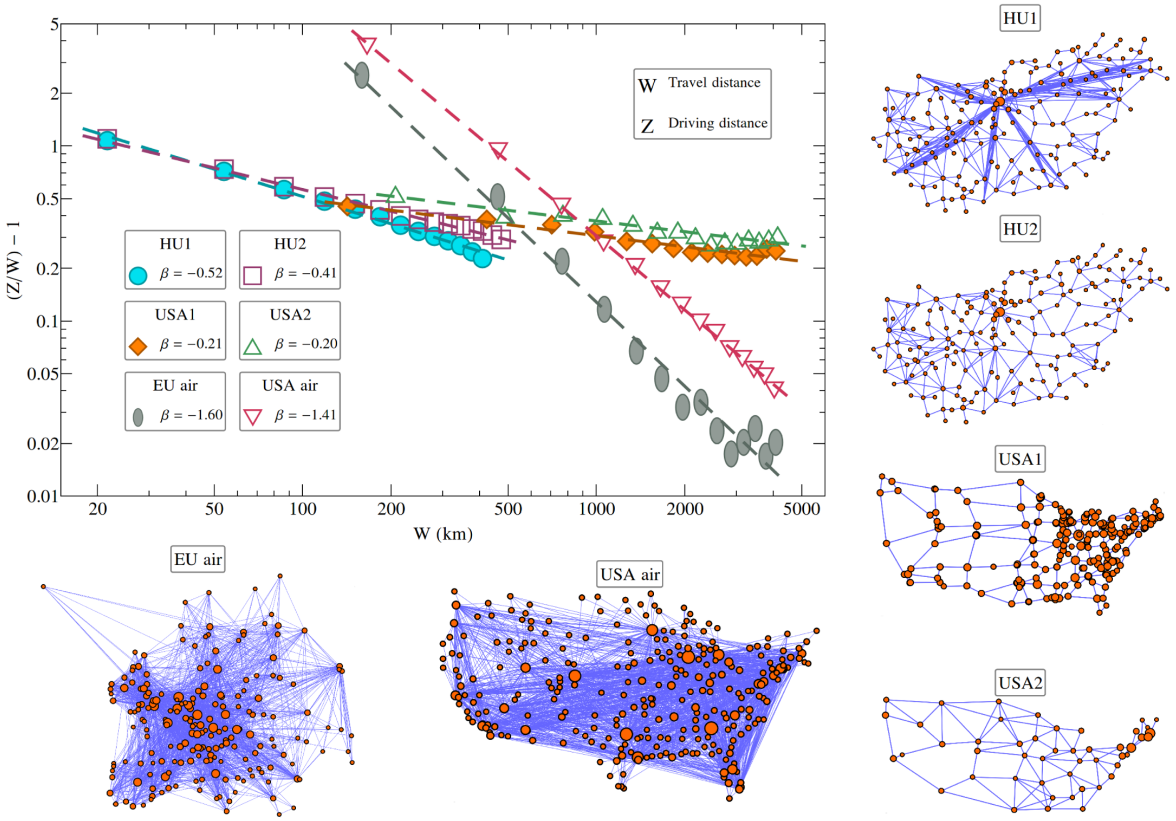

Figure 2.2: The topology of travel networks. Scaling between the driving distance (z) and geodesic distance (w) on different travel networks. The curves suggest the validity of equation (2.1) relation. The figure also illustrates the topological structure of some human transportation networks used in the present study. Please note the logarithmic axes.

For the USA we studied 241 locations in the neighbors of the major interstate roads junction points (map USA1), and 48 state capitals (map USA2).

For air travel, we considered only direct flights from airports of Europe, USA and worldwide. The spatial location of these airports is illustrated on maps of Figure 2.2. For Europe, we used timetable data and GPS coordinates of 203 major airports [30], as seen on the EU air map. In case of USA, we worked with 282 airports on the territory of USA (USA air map) but only with the topology data of the air travel network, since we did not have access to timetable data. Therefore, we were looking for other GPS tracking data with the useful flight time data recorded over 500 flights from all over the world [31]. This data is not shown on the spatial map, and we have to note that USA air data and map is used for distance calculation, but the GPS air data was used for the traveling speed calculation.

Consequently, the two inset panels in Figure 2.1 display the car and air travel data for the apparent traveling speed as a function of the travel distance. We obtained a power-law-like trend with the 0.07 scaling exponent for the roads and 0.25 for the flights, shown by dashed lines. However, there is the noticeable difference for the case of different road infrastructures, more precisely in case of HU1 data fit is a much stepper increasing trend due to use of both roads and highways (for the roadmap see Figure 2.2).

Due to the specific topology of the road-networks, there is usually no straight-line between cities, so we must distinguish two kinds of distance. The travel distance (w) means the distance on geodesic lines between the source and target points, and the driving distance, denoted as z refers for the minimal length of the path in the network. In view of the different distances, we can define different velocities. The apparent speed, denoted by v is calculated as the (travel distance)/(travel time), while the cruising or driving speed (u) is computed as the (driving distance)/(travel time).

Let us first consider the relationship between these two distances. During the topology analysis of transportation networks sketched in Figure 2.2, we find that the travel distance and the driving distance are on average related to each other in the form of a scaling relation as

$$\frac{z}{w} - 1 = C \cdot w^{-\beta} \tag{2.1}$$

where the β exponent of the power-law-like trend is expressing the relation of the driving distance to the geodesic line.

As visually illustrated in Figure 2.2, the obtained $\beta \approx 1.4 - 1.6$ (air travel) and $\beta \approx 0.2 - 0.5$ (road travel) exponents are suggesting that for air travel the value of z converges more quickly to w than in the case of the road travel, i.e., the flight paths are rather along geodesic lines than the roads.

The increasing trend for the driving speed as a function of driving distance is resulting from the combined effect of the two types of delay. First, both ends of a trajectory the source and target nodes generate delays. In our case, the nodes are cities, in which the complexity of the traffic causes smaller and larger delays depending on the size of the cities. Second, the delays are also on the trajectories as well.

These obtained results suggest that the driving or cruising speed is increasing as a function of the length of the link in case of the direct links in a network. Our results support the hypothesis according to which further we travel the faster we go. All the above data proves the increase of the averaged apparent speed as a function of distance both for different modes of traveling and also taken them together.

Second , we offer a new generalization of the radiation model considering also the effect of the travel costs [12], and we examine the applicability of this model on a complete commuter database in Hungary.

The original Radiation Model (RM) [10] is based on the simple assumption that jobseekers are optimizing their income by accepting the closest job offer that has a better salary than the one that is currently available in their current location. Assuming a $p_{\leq}(z)$ cumulative distribution function for the incomes in the studied society the probability $P_{>}(z|n)$ that an individual with income z refuses the closest n job offers is

$$P_{>}(z|n) = [p_{\leq}(z)]^n \tag{2.2}$$

By using the probability density function for incomes, $p(z) = \frac{\partial p_{\leq}(z)}{\partial z}$, we can now calculate the probability of not accepting the closest n jobs. Then, accepting the hypothesis that the number of job openings in a territory is proportional with the W population ($n = \mu W$), the radiation model probability is

$$P_{>}(W)_{RM} = \frac{1}{\mu W + 1} \text{ (one parameter fit)} \tag{2.3}$$

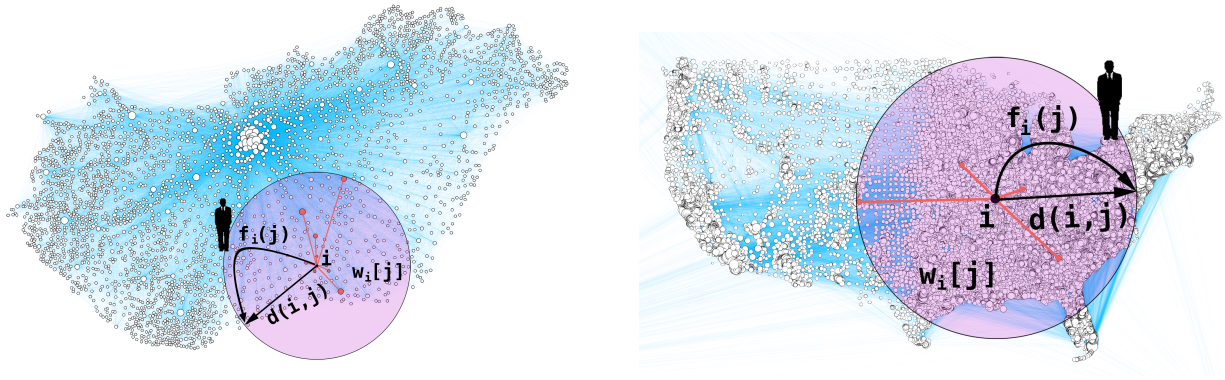

Figure 2.3: Settlements and data processing method in the commuting network of Hungary (left) and USA (right). Disks of different radius $d(i,j)$, starting from a given settlement and reaching the other j settlements, are constructed. The population $w_i[j]$ inside these disks and the commuter number, starting from settlement i and traveling to settlement j, $f_i(j)$, is recorded.

that an individual commutes to a location that is outside of a disk centered on its current location and containing a total population W.

Assuming that the jobseekers are selective in their choices and they are willing to accept better job offers only with a probability q, we get the Radiation Model with Selection (RMwS) [11]. Following the original radiation model calculations with the newly introduced selection criteria we get (for more details please consult the thesis):

$$P_{>}(W)_{RMwS} = \frac{1-(1-q)^{\mu W+1}}{(\mu W+1)\,q} \text{ (a two-parameter fit)} \tag{2.4}$$

For $q = 1$, we get back the original radiation model. The RMwS model describes better the distribution of the commuting fluxes in the USA with its two parameters [11]. This is however not a surprise since one would naturally expect that a two parameter model offers a better fit than a one-parameter model. Starting again from the original radiation model, we introduced another variation of it, which we named the Travel Cost Optimized Radiation Model (TCORM). In contrast with the original radiation model, where the job acceptance is independent of the distance between the residence and workplace, here we introduce a stronger condition for commuting: the individuals will choose the commute if there is a better income after subtracting the travel costs. In this manner, the travel costs depend also as a function of the traveled distance, and not only the transited job offers n. Using this assumption and following the reasoning of the radiation models hypothesis we get (for more details please consult the thesis):

$$P_{>}(W)_{TCORM} = \frac{1+\lambda\sqrt{W}}{\mu W+1} \text{ (a two-parameter fit)} \tag{2.5}$$

For testing the above presented models, we process a complete commuter and population database from Hungary. We analyzed the 2011 population census data [32], and we also used data with 1 km^2 resolution for population distribution in the census year 2011 [33].

During data processing, we select one by one the settlements i as source for commuting and construct

the disks with radius $d(i,j)$, reaching to the target settlement j. This is sketched in Figure 2.3. We count the total population $w_i[j]$ inside this disk and record the number of commuters $f_i(j)$ starting from settlement i and traveling to settlement j.

With the data $d(i,j)$, $f_i(j)$, and $w_i[j]$ for all the settlement pairs (i,j) we calculate the experimental $P_>(W)$ probabilities.

Using the data analyses algorithms based on the above calculations, we process to construct the $P_>(W)$ curve experimentally from the 2011 census data. The computed values are compared with the best fits acquired from the original RM model (2.3), the RMwS model (2.4) and our TCORM model (2.5). The boundary effects become important for large W values (the disks centered on the settlements become largely incomplete since they extend over the borders of Hungary). As a consequence, we considered the data only up to $W = 1\,000\,000$.

The obtained results indicates that the TCORM model performs better than the simple RM and RMwS models. The RM and RMwS models are only capable of describing a portion of the W population interval, in contrast, the TCORM model is offering a visually good fit for the whole interval.

Last, we introduce yet another model for commuting fluxes and extend our investigation using large-scale population densities and commuter fluxes from other geographical regions as well [13]. Based on the above presented analyses, we process complete commuter databases from USA, Italy, and Hungary.

Beside the RM model (2.3), the RMwS model (2.4) and the TCORM model (2.5), we study the behavior of the very popular gravity model (GM) [9], as well. We introduce and test also the model introduced here, the Flow and Jump Model (FJM).

The GM model assumes that the number of commuters $f_i(j)$ between cities i and j is written as:

$$f_i(j) = F(W_i)\frac{(W_j)^\alpha}{(r_{i,j})^\beta} \tag{2.6}$$

where W_i is the population of the settlement i and $r_{i,j}$ is the distance between settlements i and j. $F(x)$ is a monotonically increasing kernel function, and α and β are fitting exponents.

Using the $f_i(j)$ data we can also compute the $P^i_>(W)_{GM}$ probability, that an individual living in location i commutes to a location that is outside of a disk centered at its current location and containing a population W:

$$P^i_>(W)_{GM} = 1 - \frac{\sum_{j\neq i}^{(w_i[j]<W)} f_i(j)}{\sum_j f_i(j)} = 1 - \frac{\sum_{j\neq i}^{(w_i[j]<W)} \frac{(W_j)^\alpha}{(r_{i,j})^\beta}}{\sum_j \frac{(W_j)^\alpha}{(r_{i,j})^\beta}} \tag{2.7}$$

This probability is independent of the $F(x)$ kernel function. We denoted by $w_i[j]$ the total population inside the disk. Now, the averaged probability that commuters travel outside the disk with W population is

$$P_>(W)_{GM} = \langle P^i_>(W)_{GM}\rangle_i \tag{2.8}$$

We introduce now a novel one-parameter model as an alternative for the simple RM model. We name this model the Flow and Jump model (FJM). It is based on simple master equation for the $\rho(n,t) = -dP_>(n,t)/dn$ probability density. We use here the same notations as the one used previously for the

RM model. Based on the assumptions of the "growth and reset type models", which are introduced in review [34], we assume an inverse process: a backward probability flow supplemented by a jump process from the origin to any state with a given n value. The continuous master equation is the following (for more details please consult [34]):

$$\frac{d\rho(n,t)}{dt} = \frac{\partial(\eta(n)\rho(n,t))}{\partial n} + [\gamma(n)\rho(n,t)]\rho(0,t) \tag{2.9}$$

This master equation (2.9) specifies a process where there is a local net probability density flow from each state towards the $n = 0$ state and a jump probability from the origin $(n = 0)$ to an n state. For the state dependent $\eta(n)$ (flow rate) and $\gamma(n)$ (jump rate) rates we consider simple kernels that are realistic for the commuting process. The transitions $0 \to n$ governed by the $\gamma(n)\rho(n,t)$ rates describes the probability that workers choose a commuting job. $\gamma(n)$ should decrease with distance, and the proportionality with $\rho(n,t)$ suggests that the popular commuting places have many good jobs. We therefor choose the following forms for $\eta(n)$ and $\gamma(n)$:

$$\rho_s(n) = \frac{\eta(0)\rho_s(0)}{\eta(n)} e^{-\int_0^n \frac{\gamma(x)\rho_s(0)}{\eta(x)} dx} \tag{2.10}$$

After performing the calculations, based on the assumption $n(r) = \mu W(r)$ we get a slightly modified expectation for probability (for more details please consult the thesis)

$$P_>(W)_{FJM} = \frac{1}{(\mu W + 1)^{(a-1)}} \tag{2.11}$$

In the following we consider the FJM model with the universal parameter $a = 7/4$ which offers a much-improved fit for the real commuting data. In the case of the $a = 2$, we get back the original radiation model. The model is a two-parameter one, however, if we set the universality of a it becomes similarly with RM a one-parameter model.

For verifying the assumptions of the model apart of the smaller-size data available for Hungary, we are using a larger-scale dataset for the USA and one smaller-size dataset for Italy.

For USA we analyzed the estimated population census data between 2006 and 2010 [35] using $Q = 73\,803$ settlements (nodes) (white circles in Figure 2.3) and $4\,156\,426$ commuter routes (edges) (blue lines between white circles in Figure 2.3). For studying the spatial distribution of population, we used a database from years between 2006 and 2010. This database provided an estimated population of continental USA divided in $11\,078\,286$ cells of 1 km^2 area [36]. In order to speed up our calculations, we have spatially renormalized this data and obtained a less accurate resolution with 4 km^2 size cells. This is done by collapsing the data of four neighboring cells and averaging their latitudinal and longitudinal coordinates. As result we ended up with $1\,230\,920$ cells containing a total population $W = 308\,745\,231$.

Italy data contains $Q = 8\,093$ settlements, $556\,120$ commuter routes and it is from the Italian population census realized in 2011 [37]. The total population $W = 55\,605\,065$ is mapped in cells of 1 km^2 area [38].

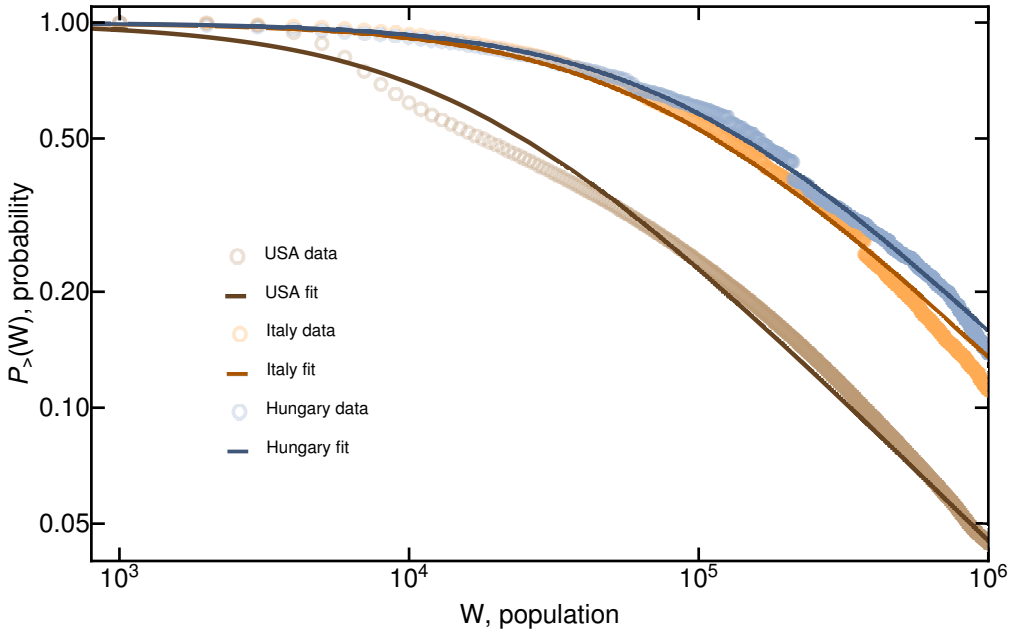

Figure 2.4: Visual comparison between the FJM model prediction and experimental data for all three countries (USA, Italy, Hungary). The faint lines composed of circles show the $P_>(W)$ experimental data and the simple dark colored lines are the best fits with the FJM model prediction (2.11). We fixed $a = 7/4$ and the best fit μ values are given in Table 2.1.

	USA	Italy	Hungary
μ	0.000062	0.000013	0.000011
$\mathbf{R^2}$	0.993	0.997	0.998

Table 2.1: Fitting parameters and goodness of the fits shown in Figure 2.4, considering the functional form given by equation (2.11) and fixing $a = 7/4$.

The experimental data processing is based on the steps sketched in the above calculations for Hungary. With the data $d(i, j)$, $f_i(j)$, and $w_i[j]$ for all the settlement pairs (i, j) (see details on Figure 2.3) we compute the $P_>(W)$ probabilities.

First, the $P_>(W)$ probabilities computed for USA are compared with the best fit results obtained from the original RM model (2.3), the RMwS model (2.4), the TCORM model (2.5), the GM model (2.8) and the novel FJM model (2.11). In the FJM model the $a = 7/4$ parameter is fixed for all studied datasets, so the only free parameter of this model is μ. For the GM model fitting was realized by considering a progressive mesh method for various $\alpha \in [-1.0, 2.5]$ and $\beta \in [-1.0, 2.5]$ values.

In order to minimalize the boundary effect, we examine the data only up to $W_{max} = 1\,000\,000$, and for the short commuting routes we set a lower threshold of $W_{min} = 1\,000$. Fitting is performed in the $[W_{min}, W_{max}]$ interval.

The obtained statistics are in favor of the FJM and GM models. The FJM model offers a good description of our studied experimental data. The fact that FJM model over performs the approximation given by the RM model is originating from the fixed parameter $a = 7/4$. Nevertheless, the FJM model overperforms also the RMwS and TCORM two-parameter models as well. The studied GM model also offers a good fit, but during the fitting for different countries we cannot fix α and β parameters.

The FJM model for $a = 7/4$ also works better for the commuting data processed for Hungary and Italy. The goodness of the fits and fit parameters are shown in Table 2.1. Additionally, Figure 2.4 shows the FJM fit for the experimentally determined $P_>(W)$ curves for all three investigated databases.

The best description from the GM and FJM models shows that if one uses the framework of the RM model it is important to take into account the fact that the selection of jobs is distance, cost and size dependent.

The fit results of the FJM model for USA, Italy, and Hungary are summarized in Table 2.1, where the best fit parameter μ characterizes both the availability of jobs per population and the attractiveness of these jobs to jobseekers.

"The more complex the network is, the more complex its pattern of interconnections, the more resilient it will be."

Fritjof Capra

3

Citation dynamics and networks

This chapter is devoted to the investigation of publication and social networks. First, in this chapter, we present our community detection methods based on graph Voronoi diagrams and stochastic graph Voronoi tessellations [39].

Based on the methodology, if a community detection algorithm is ready for the tests first we consider it on different benchmarks and real-world networks. We generate benchmark networks by the benchmark software framework provided by Lancichinetti et al. [40, 41]. This a first simple application since here we know exactly which of the nodes belongs to each cluster. In contrast, in the case of real-world networks we do not usually know the nodes affiliation to clusters, but it is also important that our algorithms work well on real-world networks.

As a specific problem where clustering is important, we studied the normalization of scientometric indicators in case of individual publications. We proposed here a local cluster detection algorithm to identify the scientometric community of an article. After we detected the local cluster, we calculate many article indicators on it. Then a normalization method was applied for these values.

We studied link evolution dynamics on the scientific publications and the Facebook friendship network, by suggesting the existence of two simple laws: preferential linking and exponential growth of the number of nodes. We find that the distribution of shares for the Facebook posts and the distribution of scientific citations both fit well with the Tsallis-Pareto probability distribution function with the exponent $g = 1.4$ [20].

First, we present a geometric solution to graph community detection based on graph Voronoi diagrams [39, 42]. This method usually is used to partition metric spaces into regions (Voronoi cells) around given seed points, as illustrated in Figure 3.1A. Each point of the space pertains to her closest seed. We apply this solution to graphs, where all edges have a positive length and the distance between two nodes is equal to the shortest path between them, see Figure 3.1B.

In order to introduce the clustering method, first, we present the Voronoi partition method in 2D Euclidean space. We consider a set of points in a 2D plane, distributed to form local groups. We calculate the local density of points inside plaquettes, and we select the Voronoi cell seeds (generator points) inside their neighborhood with radius r, as seen on Figure 3.1A. Then, we assign each non-seed point of the plane to the Voronoi cell belonging to the seed closest to the point. As a result, the points are partitioned into groups by the Voronoi cells.

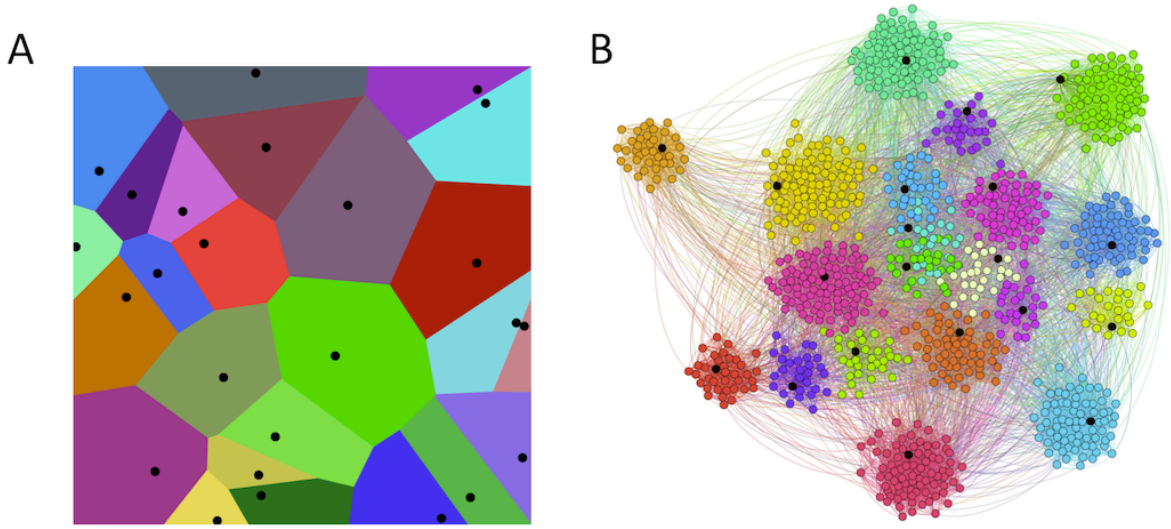

Figure 3.1: Voronoi diagrams. (A) Illustration of Voronoi partitioning in 2D Euclidean space. (B) Graph Voronoi diagram as represented by the graph drawing application Gephi using the ForceAtlas2 layout algorithm [43]. Generator nodes are shown in black.

To apply the previous theory, for networks, we need to transform this graph into metric space. First, we count a distance value between any two nodes, which is the length of the shortest path between them. The length of any path is equal to the sum of the length of edges along the path, where we defined edge length as the inverse of edge clustering coefficient (ECC) presented in [44]. The ECC of an edge between node i and node j is computed as

$$C_{i,j} = \frac{z_{i,j} + 1}{min[(k_i - 1), (k_j - 1)]} \tag{3.1}$$

where k_i, k_j are the degrees of the nodes, $z_{i,j}$ is the number of triangles the edge belongs to and $min[...]$ is the number of potential triangles it could belong to, as it is the smaller value of the degrees of the two adjacent nodes, minus one (the examined edge). If the $1/C_{i,j}$ value is large, then it likely means that the edge connects nodes in different clusters.

After we have defined the distance, the next step is to determine generator points. During our calculations, we used a Voronoi seed selection method based on the relative local density of nodes [45]. It operates on a subgraph consisting of the first neighbors of node i. For this subgraph we determine

$$\rho_i = \frac{m}{m + k} \tag{3.2}$$

where m is the number of edges inside the neighborhood, and k is the number of edges going out of the subgraph.

If we determined the generator points as a function of distance $1/C_{i,j}$ and a radius r, then we have a clustered graph. It is apparent, that varying the r parameter will influence the number of communities within the graph. We studied the influence of the value of r and obtained that relatively small r provides a good partitioning in both benchmarks and real-world networks. The best strategy however for

the Voronoi partitioning is by increasing the value of r and observing the quality function of the given clustering, for example with a cluster modularity [46].

We tested our methods on benchmark and real-world networks. We used the benchmark networks generator algorithms with a large variety of different parameter settings [40, 41]. We have generated 100 graphs with different parameters. These graphs were partitioned with increasing r values and monitored the quality functions between our partitioning and the predefined communities on the graph. For comparison, we used the modularity function [46] and the mutual information between our clustering.

When our method reached the maximum value of modularity, the mutual information was close to 1 at the optimal value of r, indicating that our method identified the original communities successfully on benchmarks. However, it is noticeable that in some cases the community structure is not very clear, due to the fact that there are a lot of small clusters.

Contrary to the benchmarks in the case of real-world networks clustering is more complicated. In such cases, we do not know the optimal community structures, and the goodness of our method can be judged only by comparing with other already accepted methods. We tested our algorithm on several real-world networks with different structure and origin, and we compared it to five other widely used algorithms. As a result, we obtained that our method works well for relatively small r values, and only, when we have large mixing, does not find the right clusters.

Second, we consider a stochastic version of the above presented community detection method [47]. The essence of the new method is the random selection of the generator points and the Voronoi cohesion matrix. Each node has the possibility to be a generator point. Then for a certain selection of the generator nodes, we calculate the Voronoi cohesion matrix, which is the probability of co-location of a pair of nodes, i.e., it tells the probability of intra-community and inter-community pairs. In this form, the values are larger for intra-community pairs than inter-community pairs. Following these calculations, we repeat it and modify the network topology until we get a clusterization through some stochastic steps.

Knowing that each node can be randomly selected as a generator point, we detail know the steps of the algorithm:

- we randomly choose g generator nodes, and we perform a graph-Voronoi tessellation by the distance along the shortest path. Then, for each pair of nodes, we determine their co-location.
- we repeat this tessellation R_e times and calculate the Voronoi cohesion matrix, which is defined as the average of the co-location matrices. These cohesion matrices are plotted so that nodes are ordered by the ground truth information.
- we identified that the intra-module and inter-module nodes are somewhat separated. Starting from that, we apply an iterative modification to the network's topology. A small percentage of the edges with low cohesion is moved in between unconnected nodes with high association cohesion. As a result of the modification, the community structure is preserved, but the separation between them are even clearly defined.

We tested our algorithm on benchmark networks. We generated a multilevel benchmark network with 28 and 10 communities on the first and second levels. We evaluated the cohesion matrix and the topological relocation at every cycle. As a result, we got 9 communities in both cases (with and without topological relocation), but the cohesion matrix indicates better clustering by the topological ordering.

Before we test real-world networks, let's see how the Voronoi cohesion matrix is assembled. One element of the matrix is a probability for co-location of a pair of nodes, i.e., probability for sharing the same Voronoi cell. The matrix contains diagonal and off-diagonal probabilities. After the ground truth information ordering method, the nodes, belonging to same cluster, are located on the matrix diagonal with higher probabilities.

If it is necessary when the given cohesion matrix does not show clearly the communities, we are using a contrast boosting method during which we topological relocate the weak edges, and after that, we separate the communities with a threshold value. During the contrast boosting method communities can be determined as follows: all nodes get a separate community label, and in a loop, over all nodes, the community label of the current node is assigned to the nodes that have not changed their label yet and whose cohesion with the current node exceeds a threshold.

Clustering of real-world networks remains still difficult. During our community detection, the plain stochastic graph-Voronoi method generates the cohesion map, where the gap between the inter and intra pairs is not significant.

In the light of these, the plain stochastic graph-Voronoi method works well on benchmarks. In the case of real-world networks, the contrast boosting technique is also needed. In some extreme cases, the combined method is also required.

Third, we study the normalization of scientometric indicators of individual publications [19]. Increasingly used bibliometric indicators, such as the impact factor [48], eigenfactor [49] or h-index [50] cannot assure a direct comparison of different disciplines with each other. Also, within the same journal, the published articles have widely different citation numbers.

The representation of the publications and citations through the publication network is widely used. Every publication and citation in this network is indicated by nodes and edges between them, respectively. Using this identification, the individual scientometric indicators become calculable possible on such networks. More precisely, the article evaluation is realized with the citation number, i.e., the input degree of the node or the local PageRank [51] measures closely related to the detected scientific domains.

Based on the idea of the parallelized local community detection method [52], we introduce a specific local cluster detection (LCD) algorithm. This is a shell spreading method outward from a starting publication, and using this, the algorithm detects communities without partitioning the entire network. Then, we tested the LCD algorithm on both benchmark and real-world networks too.

After community detection is done, we calculate the relevant scientometric indicators on the detected local community. The first indicator is the citation of one article which reflects its impact on the scientific community. This indicator is equal with the input degree of nodes, i.e., the number of references on the whole network. The second is the simplified (local) version of the PageRank [51].

The first important statistical quantity to normalize is the citation statistics of the studied articles. In our case, the probability distribution of an article indicator represents the occurrence probability of articles by the article indicator values. Therefore, these probability distributions will describe the citation behavior in different scientific domains.

We mention that the input degree n_i distribution function of clusters on the same benchmark network

is a straight line on the log-log plot. Starting from the power-law distribution of the networks, the tail of this distribution function is described with the

$$p(n_i) = An_i^{\alpha} \tag{3.3}$$

power-law, where A is a proportionality constant. For simplicity, we rescale the n_i values, so that each distribution is fitted with the $A = 1$ normalization constant. Accordingly, each input degree has to be multiplied with a scaling factor $\xi = A^{1/\alpha}$, as follows:

$$p(n_i) = An_i^{\alpha} = (A^{1/\alpha}n_i)^{\alpha} = (\xi n_i)^{\alpha} \tag{3.4}$$

This rescaling method is tested on different real-world citation networks, for example on the Web of Science citation network with 771 914 articles and 7 779 703 citations. We started the LCD algorithm, then we measured the input degree and the local PageRank distributions. The input degree distributions shows that in case of WoS network it is necessary to scale the citation numbers to achieve a normalized distribution with the $A = 1$ proportionality constant.

The similarly studied local PageRank distribution function is also fitted with power-law. The data is fitted with the $p(n_i) = 10^{-7}n_i^{-1.4}$ function. The scaling and the obtained scaling factors differ significantly due to the different nature of the local PageRank article indicator.

Last, we briefly present our results concerning the statistics of citations and Facebook shares [53]. It is well known that citations evaluate the articles, authors or institutions. Following the pattern of scientific citations, the Facebook shares also rate the posts or users. Citations or shares of the publications or posts quantify and characterize the quality of them. In our work, we focus on these distributions and look for universalities among them.

Starting from earlier studies revealing some universalities in the citation distributions [54], we also examined the distribution of citations and shares. Previous studies have shown that citations received by several academic institutions and journals fall on a common curve if they are renormalized relative to the mean. In other words, they calculated the probability density $f(x)$ for one paper with x citations and plotted the $\langle x\rangle f(x)$ value as a function of $x/\langle x\rangle$. $\langle x\rangle$ is the mean value of x. The results plotted in Figure 3.2 shows that the different sets rescale into the same curve.

We observe also that the plotted distributions exhibit a clear power-law trend for high citation numbers in the $x/\langle x\rangle > 10$ domain. We have shown that the entire distribution can be successfully fitted with the Tsallis-Pareto (TP) type distribution [20]

$$f(x) = \frac{g}{(g-1)\langle x\rangle}\left(1 + \frac{x}{(g-1)\langle x\rangle}\right)^{-1-g} \tag{3.5}$$

which is a probability density function (PDF) with a power-law-like tail. This is not completely a scale-free distribution, the scale-free properties fulfill only for $g > 1$ and large enough $x/\langle x\rangle$ values. It is therefore more appropriate to call these heavy-tail distributions [55].

Assuming the above distribution function, we processed many datasets as shown in Figure 3.2. First,

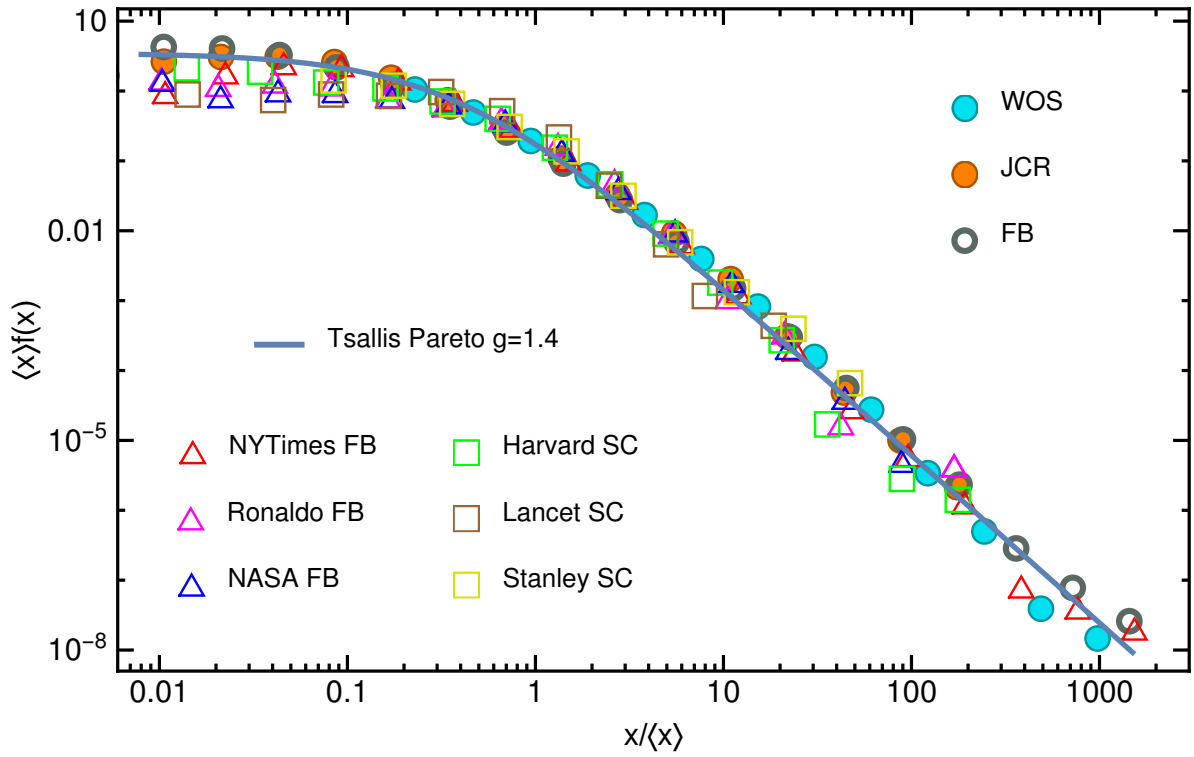

Figure 3.2: Rescaled distribution of the citation (share) numbers. $f(x)$ is the probability density (PDF) for one paper (post) to have x citations/shares. We present the $\langle x \rangle f(x)$ value as a function of $x/\langle x \rangle$ ($\langle x \rangle$ the mean value, or first moment of the PDF). For high citation number a clear power-law trend is visible. Different symbols are for different datasets as illustrated in the legend. For high $x/\langle x \rangle$ a clear power-law trend is visible. The entire curve can be well-fitted with a TP distribution (3.5) with $g \approx 1.4$ and $\langle x \rangle = 1$.

we processed more than 600 000 ISI Web of Science (WoS) publications [56], ten years long citations statistics for all ISI indexed (approx. 12 000) journals from the Journal Citation Report (JCR) [57] and more than 150 000 posts from 16 different Facebook users (pages) [58]. All three sets of data suggests that the one-parameter TP type PDF is appropriate for data fitting with $g \approx 1.4$ parameter, as shown by the continuous line and data points in Figure 3.2.

Besides the large datasets, we collected data from 16 Facebook pages. Here we show three different types, one from news the New York Times (NYTimes FB) page, one from sports celebrity the Cristiano Ronaldo (Ronaldo FB) page and one from science the NASA (NASA FB) page. In the case of each small network, it is visible that our selected TP type probability density function fits very well also these smaller datasets.

In addition to these, we studied the citations for articles published in 1990 with authors from Harvard University (Harvard SC), for papers published in The Lancet Elsevier journal in 1990 (Lancet SC) and for a single author from physics, Prof. H. E. Stanley from the Boston University, USA, who has 965 ISI publications and 62 996 ISI citations (Stanley SC). These data also fits well with the TP type PDF.

For all datasets, we constructed the experimental probability distribution function by a logarithmic binning method with 2^n size bins. Aside from small fluctuations in the examined data in case of small datasets, all collected data followed the same trend with $g \approx 1.4$.

To model the obtained results, we assume that the illustrated growth process results from the exponential growth of publications (posts) number as a function of time and citations following a rich gets richer multiplicative growth. For the case of Facebook, the exponential growth is highlighted by the presentation of Mark Zuckerberg according to which the information sharing activity on Facebook is also growing exponentially [59]. The Matthew effect: "For to all those who have, more will be given" in many social systems including citation in science has also been discussed in many previous works [60, 61].

4

Conclusion

In the present thesis, we studied dynamics in social systems using computational physics approaches. The hard task we had to face was to logically synthesize and present in a pedagogical manner the studies we have performed in the related fields: multiagent evolutionary games, human mobilities, social networks, and scientometrics. Besides modeling, a challenge that we faced was data mining, data processing and developing numerical and analytical tools for comparing model and real-world data.

We have studied several types of evolutionary games using both imitation and logit strategy update rules on different type of networks. We investigated the strategy distributions of different type of players through Monte Carlo simulations and analytical calculations.

We discussed correlations in the distribution of strategies, we applied synchronized strategy update rules, and we demonstrated the invasion and speciation processes. We illustrated the positive impact of the imitation and logit update rules for the maintenance of cooperation. Then, we examined the invasions in horizontal, vertical and tilted directions.

As a second topic, we have analyzed the universal laws that govern human mobilities, and we have computed the probability of commuting through a certain number of population. We studied different human mobilities on road-networks, and air travel networks. We investigated commuting patterns between the settlements of three geographical regions, and elaborated two original model. To compare in a critical manner our theoretical models with reality, we used human transportation networks, census datasets, jobs and distance information.

As a first result, we confirmed our hypothesis according to which there is a universal rule: “further we travel the faster we go”. Based on the travel distance and traveling time dependence, we proved that the averaged apparent speed is increasing as a function of the distance, following a power-law-like trend.

We than studied the distribution of commuter fluxes. We used radiation type models, the gravity model, and the flow and jump model for fitting the experimentally observed data. The Flow and Jump model proposed by us offered the best fits for all three experimental data we have used.

Finally, we have investigated community appearance and clusterization effects on networks. We proposed a field-based normalization of scientific articles, and we presented similar scaling rules for scientific publications and Facebook posts. The research was performed on publication and social networks, and we tested our algorithms on benchmarks and real-world networks too.

For publication networks we proposed two communities detection methods. These methods can be used to efficiently group nodes in different clusters. We also developed a local cluster detection method, which is working without processing the entire large network. This proved to be important for determining and normalizing the individual indicators. In order to achieve that, we normalized the probability distributions of scientific indicators according to the power-law type distributions attributes of data to eliminate the differences between scientific areas.

Besides that, we studied citation/share behavior in the publication and social networks, and we observed that they show the same popularity pattern. Following a master equation approach, we were able to model the characteristic statistics in both systems, and we found that the Tsallis-Pareto type probability distribution function describes well their normalized distribution.

In conclusion, we introduced several novel approaches to characterize the dynamics in social systems, revealing novel and interesting aspects of these systems of very different natures.

5

Publications related to the thesis

5.1 Scientific papers

- Szabó, Gy., **Varga**, **L**., and Borsos, I. Evolutionary matching-pennies game on bipartite regular networks. *Phys. Rev. E*, 89, 042820 (2014).
- Deritei, D., Lázár, Zs. I., Papp, I., Járai-Szabó, F., Sumi, R., **Varga**, **L**., Ravasz Regan, E., and Ercsey-Ravasz, M. Community detection by graph Voronoi diagrams. *New J. Phys.*, 16, 1–17 (2014).
- **Varga**, **L**., Vukov, J., and Szabó, Gy. Self-organizing patterns in an evolutionary rock-paper-scissors game for stochastic synchronized strategy updates. *Phys. Rev. E*, 90, 042920 (2014).
- Vukov, J., **Varga**, **L**., Allen, B., Nowak, M. A., and Szabó, Gy. Payoff components and their effects in a spatial three-strategy evolutionary social dilemma. *Phys. Rev. E*, 92, 012813 (2015).
- **Varga**, **L**., Kovács, A., Tóth, G., Papp, I., and Néda, Z. Further We Travel the Faster We Go. *PLoS ONE*, 11, e0148913 (2016).
- Szabó, Gy., **Varga**, **L**., and Szabó, M. Anisotropic invasion and its consequences in two-strategy evolutionary games on a square lattice. *Phys. Rev. E*, 94, 052314 (2016).
- **Varga**, **L**., Tóth, G., and Néda, Z. An improved radiation model and its applicability for understanding commuting patterns in Hungary. *Regional Statistics*, 6, 27–38 (2016).
- Lázár, Zs. I., Papp, I., **Varga**, **L**., Járai-Szabó, F., Deritei, D., and Ercsey-Ravasz, M. Stochastic graph Voronoi tessellation reveals community structure. *Phys. Rev. E*, 95, 022306 (2017).
- Néda, Z., **Varga**, **L**., and Biró, T. S. Science and Facebook: The same popularity law! *PLoS ONE*, 12, e0179656 (2017).
- **Varga**, **L**., Tóth, G., and Néda, Z. Commuting patterns: The Flow and Jump model and supporting data. *EPJ Data Science*, 7, 37 (2018).

5.2 Conference proceedings papers

- **Varga**, **L**., Deritei, D., Ercsey-Ravasz, M., Florian, R., Lázár, Zs. I., Papp, I., and Járai-Szabó, F. Normalizing scientometric indicators of individual publications using local cluster detection methods on citation networks. *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, 12(9), 1189–1198 (2018).

5.3 Conference contributed talks

- "Commuting patterns: a modified radiation model and supporting data", MaCS 2018, 12th Joint Conference on Mathematics and Computer Science, Săcuieu, Romania, 14-17 June 2018.
- "Normalizing scientometric indicators of individual publications using local cluster detection methods on citation networks", ICCSIB 2018 : 20th International Conference on Cybermetrics, Scientometrics, Informetrics and Bibliometrics, Barcelona, Spain, 29-30 October 2018.

5.4 Conference posters

- "Local Cluster Detection Method for Normalizing Scientometric Indicators" and "Community detection using stochastic graph Voronoi tessellation", NetSci 2015, Zaragoza, Spain, 1-5 June 2015.
- "Testing human mobility models with commuters data" and "Velocity versus distance scaling in human travel", XXXVI Dynamic Days Europe, Corfu, Greece, 6-10 June 2016.
- "Genetic-like algorithm applied on citation networks for evaluating scientific publications", Conference of Complex Systems 2016, Beurs Van Berlage, Amsterdam, The Netherlands, 19-22 September 2016.
- "An improved radiation model and its applicability for understanding commuting patterns", meco43: 43rd Conference of the Middle European Cooperation in Statistical Physics, Kraków, Poland, 1-4 May 2018.

Selected References

[1] T. Parsons. *The Social System*. Routledge, London, 1991.

[2] M. Barthélemy. Spatial networks. *Phys. Rep.*, 449:1–101, 2011.

[3] J. Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, UK, 1982.

[4] C. Domb. Ising model. In C. Domb and M. S. Green, editors, *Phase Transitions and Critical Phenomena, Vol. 3*, pages 357–484. Academic Press, London, 1974.

[5] Gy. Szabó and G. Fáth. Evolutionary games on graphs. *Phys. Rep.*, 446:97–216, 2007.

[6] K. Binder and D. W. Heermann. *Monte Carlo Simulation in Statistical Physics. An Introduction*. Springer-Verlag, Berlin, 1992.

[7] R. Gallotti and M. Barthélemy. Anatomy and efficiency of urban multimodal mobility. *Sci. Rep.*, 4:6911, 2014.

[8] M. C. González, C. A. Hidalgo, and Barabási A. L. Understanding individual human mobility patterns. *Nature*, 453:779–782, 2008.

[9] F. Lukermann and P. W. Porter. Gravity and potential models in economic geography. *Contributions to probability and statistics*, 50:493–504, 1960.

[10] F. Simini, M. C. González, A. Maritan, and A. L. Barabási. A universal model for mobility and migration patterns. *Nature*, 484:96–100, 2012.

[11] F. Simini, A. Maritan, and Z. Néda. Human mobility in a continuum approach. *PLoS ONE*, 8(3):e60069, 2013.

[12] L. Varga, G. Tóth, and Z. Néda. An improved radiation model and its applicability for understanding commuting patterns in hungary. *Regional Statistics*, 6:27–38, 2016.

[13] L. Varga, G. Tóth, and Z. Néda. Commuting patterns: The flow and jump model and supporting data. *EPJ Data Science*, 7:37, 2018.

[14] A. L. Barabási. *Network Science*. Cambridge University Press, United Kingdom, 2012.

[15] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, New York, 2010.

[16] R. N. Kostoff. Citation analysis cross-field normalization: a new paradigm. *Scientometrics*, 39(3): 225–230, 1997.

[17] F. Radicchi, S. Fortunato, and C. Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proc. Natl Acad. Sci.*, 105(45):17268–17272, 2008.

[18] A. Schubert and T. Braun. Cross-field normalization of scientometric indicators. *Scientometrics*, 36(3):311–324, 1966.

[19] L. Varga, D. Deritei, M. Ercsey-Ravasz, R. V. Florian, Zs. I. Lázár, I. Papp, and F. Járai-Szabó. Normalizing scientometric indicators of individual publications using local cluster detection methods on citation networks. *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, 12(9):1189–1198, 2018.

[20] S. Thurner, F. Kyriakopoulos, and C. Tsallis. Unified model for network dynamics exhibiting nonextensive statistics. *Phys. Rev. E.*, 76:036111, 2007.

[21] L. E. Blume. The statistical mechanics of strategic interactions. *Games Econ. Behav.*, 5:387–424, 1993.

[22] L. E. Blume. The statistical-mechanics of best-response strategy revision. *Games Econ. Behav.*, 11: 111–145, 1995.

[23] Gy. Szabó, L. Varga, and I. Borsos. Evolutionary matching-pennies game on bipartite regular networks. *Phys. Rev. E*, 89:042820, 2014.

[24] R. J. Glauber. Time-dependent statistics of the Ising model. *J. Math. Phys*, 4:294–307, 1963.

[25] L. Varga, J. Vukov, and Gy. Szabó. Self-organizing patterns in an evolutionary rock-paper-scissors game for stochastic synchronized strategy updates. *Phys. Rev. E*, 90:042920, 2014.

[26] J. Vukov, L. Varga, B. Allen, M. A. Nowak, and Gy. Szabó. Payoff components and their effects in a spatial three-strategy evolutionary social dilemma. *Phys. Rev. E*, 92:012813, 2015.

[27] Gy. Szabó, L. Varga, and M. Szabó. Anisotropic invasion and its consequences in two-strategy evolutionary games on a square lattice. *Phys. Rev. E*, 94:052314, 2016.

[28] L. Varga, A. Kovács, G. Tóth, I. Papp, and Z. Néda. Further we travel the faster we go. *PLoS ONE*, 11:e0148913, 2016.

[29] Hungarian central statistical office, regional atlas - Hungary's territorial breakdown by statistical micro-regions for regional development. https://www.ksh.hu/regional_atlas_microregions, 2015.

[30] T. Dusek. The time frame of the air travel in Europe. *Private communication*, 2009.

[31] GPSlib, GPS tracks hosting service. http://www.gpslib.net, 2015.

[32] Census tract flow, commuting data, Hungary. http://www.ksh.hu, 2011.

[33] Population distribution, Hungary. http://ec.europa.eu/eurostat/cache/GISCO/geodatafiles/GEOSTAT-grid-POP-1K-2011-V2-0-1.zip, 2011.

[34] T. S. Biró and Z. Néda. Unidirectional random growth with resetting. *Physica A*, 499:355–361, 2018.

[35] CTPP Census tract flows, commuting data, American Community Survey. https://www.fhwa.dot.gov/planning/census_issues/ctpp/data_products/2006-2010_tract_flows/, 2006-2010.

[36] Population distribution, American Community Survey. https://www.census.gov/geo/maps-data/data/tiger-data.html, 2006-2010.

[37] Census tract flow, commuting data, Italy. http://www.istat.it/storage/cartografia/matrici_pendolarismo/matrici_pendolarismo_2011.zip, 2011.

[38] Population distribution, Italy. http://ec.europa.eu/eurostat/cache/GISCO/geodatafiles/GEOSTAT-grid-POP-1K-2011-V2-0-1.zip, 2011.

[39] F. Aurenhammer. Voronoi diagrams – a survey of a fundamental geometric data structure. *ACM Comput. Surv.*, 23:345–405, 1991.

[40] A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Phys. Rev. E*, 80:016118, 2009.

[41] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmarks graphs for testing community detection algorithms. *Phys. Rev. E*, 78:046110, 2008.

[42] D. Deritei, Zs. I. Lázár, I. Papp, F. Járai-Szabó, R. Sumi, L. Varga, E. Ravasz Regan, and M. Ercsey-Ravasz. Community detection by graph voronoi diagrams. *New J. Phys.*, 16:1–17, 2014.

[43] M. Bastian, S. Heymann, and M. Jacomy. Gephi: an open source software for exploring and manipulating networks. *Proceedings of ICWSM: AAAI Int. Conf. on Weblogs and Social Media*, pages 361–362, 2009.

[44] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proc. Natl Acad. Sci.*, 101:2658–2663, 2004.

[45] S. Fortunato. Community detection in graphs. *Phys. Rep.*, 486:75–174, 2010.

[46] M. E. J. Newman. Modularity and community structure in networks. *Proc. Natl Acad. Sci.*, 103: 8577–8582, 2006.

[47] Zs. I. Lázár, I. Papp, L. Varga, F. Járai-Szabó, D. Deritei, and M. Ercsey-Ravasz. Stochastic graph voronoi tessellation reveals community structure. *Phys. Rev. E*, 95:022306, 2017.

[48] E. Garfield. The impact factor and using it correctly. *Der Unfallchirurg*, 101(6):413–414, 1998.

[49] C. T. Bergstrom, J. D. West, and M. A. Wiseman. The eigenfactor metrics. *Journal of Neuroscience*, 28(45):11433–11434, 2008.

[50] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proc. Natl Acad. Sci.*, 102(46):16569–16572, 2005.

[51] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. *Technical Report, Stanford InfoLab*, 1999-66, 1999.

[52] J. P. Bagrow and E. M. Bollt. Local method for detecting communities. *Phys. Rev. E*, 72(4):046108, 2005.

[53] Z. Néda, L. Varga, and T. S. Biró. Science and facebook: The same popularity law! *PLoS ONE*, 12:e0179656, 2017.

[54] A. Chatterjee, A. Ghosh, and B. K. Chakrabarti. Universality of citation distributions for academic institutions and journals. *PloS ONE*, 11:0146763, 2016.

[55] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51:661–703, 2009.

[56] Web of Science. http://www.webofknowledge.com, 2015.

[57] InCites, Journal citation report, Thomson Reuter. https://jcr.incites.thomsonreuters.com/JCRJournalHomeAction.action, 2016.

[58] Facebook. https://www.facebook.com, 2016.

[59] M. Zuckerberg. Online sharing is growing at an exponential rate. https://www.youtube.com/watch?v=HNy9uxcRedU, 2012.

[60] H. Jeong, Z. Néda, and A. L. Barabási. Measuring preferential attachment in evolving networks. *Europhys. Lett.*, 61:567–572, 2003.

[61] J. Wang. Unpacking the Matthew effect in citations. *J. of Informetics*, 8:329–339, 2014.