Recommendation System for Social Network Analysis Based on the Observed Behaviors

Luciene Cristina Alves Rinaldi
[1]*Centro Universitário FEI*
[2]*Universidade de São Paulo (USP)*
Brasil

Augusto Erni Klaus Einsfeldt
[1]*Centro Universitário FEI*
Brasil

Abstract

This paper presents a system based on detected information in social networks to support experiments in primate behavior psychology. The study aims to the learning processes through observation (e.g., usage of tools for breaking encapsulated fruits) and experiments that may emerge in an autonomous agents group (represented by capuchin monkeys). The experiment focuses on simulation data analysis for the prediction of social behavior associated with some cognitive skills. The dissemination of information occurs through the group social structure, which is represented by a graph object that has been built based on theorems and algorithms from the graph theory. The objective is to suggest items that may be of interest to the user. The finding of these contents uses techniques that analyze and combine data to reach the best possible degree of recommendation. These benefits coupled with the large volume of data focused on the right audience can boost business' profits or surveys. These benefits are the drivers for the high demand of such systems in different domains like films (e.g., NetFlix), social networks (e.g., Facebook), sales (e.g., IBM Watson), and others.

1. Introduction

The present computational model simulates the behavior of the observed primates, which can be useful to study situations and to test hypotheses that are difficult to observe. It is widely accepted, however, that computational models can reproduce only some characteristics of the environment, using rules consistent with observed behaviors. Even detailed computational models will not be able to represent all the specific properties of the *Real Environment* (RE). However, it is remarkable how well-refined and adequately proposed computational models can be of great help in researching this field or other areas, despite their inevitable limitations.

The work developed in this work was inspired by the study on social learning in robust capuchin monkeys in the *Tiete Ecological Park* (PET), which has been carried out by researchers from the LEC of IPUSP [3] [4]. Field behavior observation is a working method of ethology that studies animal behavior based on naturalistic observations or field experiments with an evolutionary bias that can give rise to valuable data from which it is possible to make deductions.

The first scenario considered as a RE is the island of PET where 38 robust capuchin monkeys live. Most were born there although some came from apprehensions by the Federal Police (from arrested smugglers or illegal commerce) or have come from donations [1].

IPUSP's LEC researchers go to the park an average of three times a week for a period of approximately two hours and record in a digital camera what is happening in the environments and include their notes to the data collected. The method used is the focal scan method [2], which considers a particular chosen monkey (focal animal), during a time interval, referring to the monkeys that are close enough to him (1m or 10m radius) and determine the level of proximity between them.

Based on the observed proximity data, two analytical tools were constructed to represent the social network of this community. The social network graph of the proximity matrix of the first developed tool is called Horacio and uses the average proximity level of the encounters and the second tool called Cuzco uses the *Minimum Generating Forest* (FGM).

The other observation data contains the monitoring of the behavior of the activities in which the monkeys were performing and in which location they happened. These are the data exchanged between the monkeys by the social network according to some criteria (rules) of the environment. Such behaviors are the daily activities that monkeys perform like eating, foraging, sleeping and others. Some behaviors are innate (e.g., sleep and to nurse) and others derive from learning from the observation of other monkeys, that is, they have some relation through a level of proximity that can be observed through the social network model.

These studies motivated the development of the Horacio platform that represents the second scenario treated as *Virtual Environment* (VE). The idea of the Horacio platform was to propose a model that finds the influence of *other activities* (e.g. social play, observation of coconuts breaking, social tolerance, among others), called *peripheral activities*, that helps in the transmission of a *specific knowledge* (*coconut break*). There *peripheral activities* are those that are

not directly related to the transmission of knowledge of the *specific activity*. It is important to note that in the presented model the agent used presents a *unique knowledge* (called *specific knowledge*) that is used to represent the *proficiency* (monkeys that learned to break coconuts) of the agent in the accomplishment of the *specific activity*. This knowledge is used to compare the *actual result* with the *simulated result* (*proficiency*).

In [3][4] it was verified the hypothesis that the activity of social play could influence the learning of coconut breakage due to the observation that monkeys intercalated the episodes of coconut break with social play. This led to the motivation to create a computational system that could help in the *transmission of knowledge about the* coconut breakage *specific activity*, using as reference the observed data.

The developed simulator is feed with field-generated raw data in a pseudo-randomized fashion with gaussian distribution, modulated, however, by the observed data from the RE, which serve to define rules consistent with RE, including statistical information about proximity and behavior profile of the agents represented in the VE (simulated).

The Horacio platform is composed by the simulation that works, allowing generate the information of the virtual scenario compatible with those that were observed in the real scenario. Thus, it allows the identification of essential characteristics of this group of primates and can be compared and tested to find the best expected result.

This work is initially related to the analysis of a social network model used to transfer knowledge of behaviors that is consistent with the data observed in the field of this group of primates.

Using the data observed in the field the simulated proximity relation was obtained such as that performed with the RE data, which is the same used by the social network model to transfer knowledge from one agent to the other only when they are close to each other (e.g. at the same site as the coconut break site).

Following the relationship with proximity is verified the knowledge between the agents (proficiency). Next, the model considers the activities that the two agents are performing at that moment. The transmission of knowledge is given exclusively by the activities being carried out and the proximity between the two agents, where a specific knowledge transmission coefficient is verified. This coefficient is found from the *Genetic Algorithm* (GA). The knowledge of the receiving agent is increased according to the knowledge transmission model that uses an equation with the parameters referring to the coefficient of knowledge found by the GA and is related to the activities the agents were executing at that moment [7] [8].

The idea of collaboration with LEC of IPUSP researchers was to use ethological data collected in the field by being consolidated in long theoretical studies on natural phenomena using methodologies and providing safe requirements for synthetic experiments. Most simulations that do not use real data can abuse the assumptions that are usually characterized by the enormous degree of freedom in their development [5]. Thus, starting from solid bases, it was intended to show that the simulations help as a laboratory of virtual experiments where real-world situations sometimes do not allow tests like these virtual laboratories can offer.

2. Development of the simulation environment and model

2.1. Knowledge transmission model

According to [6] cited in [3] [4], the environment in which monkeys live, social position, affiliative ties, among other factors, can influence the transmission of behavioral information among monkeys in a group. In addition, interaction among subjects, observation of their behaviors, physical closeness and the time they remain together may increase the chances of transmitting behavioral information.

In [3] is used the term 'information transmission' (area jargon). In this work, the term 'knowledge transmission' is used. It is important to note how it is used here. It was assumed that when two monkeys are close enough to each other, there may be transfer or transmission of knowledge (in the sense of competence or acquisition of pragmatic knowledge) from one monkey to the other according the activities that each one is doing. It should be noted that here were not used the same characteristics previously mentioned in [5]

- A knowledge (competence) is transmitted from one subject to another only when they are close (related) and varies according to the activities that each one is performing;
- Knowledge is transmitted in a directional way, that is, from the subject with greater knowledge (sender) to the subject with lesser knowledge (receiver), thus, the knowledge of the two involved is never diminished;
- The knowledge of the subject receiving never surpasses the knowledge of the emitting subject.

Once established that two subjects are sufficiently close an equation is used for the transmission of knowledge that depends on the activities that the two are performing. The following is the eq. (1) that is used for the transmission of knowledge. For this, it is

necessary to define some basic sets that are directly related to the model.

$S = \{s_{1,..,}s_{N_S}\}$ is the set of the N_S subjects used in the model;

$A = \{a_{1,..,}a_{N_A}\}$ is the set of the N_A activities that can be performed by the subjects;

$L = \{l_{1,..,}l_{N_L}\}$ is the set of the N_L places where the subjects can perform the activities.

Being A the set of activities and the subjects $S_i, S_j \in S$, the variation of knowledge of S_i is given by:

$$\triangle c_i = \begin{cases} (c_j - c_i)\, M_{a_i,a_j} & , \; if \; c_j > c_i \\ 0 & , otherwise \end{cases} \quad (1)$$

Where,

$\triangle c_i \in [0,1]$: is the variation in the increase of knowledge of subject S_i;

$c_i \in [0,1]$: it is the knowledge of the subject S_i;

$c_j \in [0,1]$: it is the knowledge of the subject S_j;

$a_i \in A$: represents the activity being performed by the subject S_i;

$a_j \in A$: represents the activity being performed by the subject S_j;

$M \in [0,1]$: it is the matrix of the coefficients of transmission of knowledge composed of the total number of N_A activities.

Note that M is the matrix that stores the parameters of the transmission coefficients of knowledge used in the transmission of knowledge in eq. (1), and in the used model the transmission of knowledge occurs exclusively considering the activities performed by the subjects.

The agent e_1 is performing the activity $a_1 = L = locomotion$ and the e_2 is performing the activity $a_2 = BS = social\ play$ when they are close to each other (within a radius of 1m or 10m) at the site $l = SQ = coconut\ break\ site$ (Fig. 1). It is important to note that the *transmission coefficients of knowledge* of M are found by the GA.

It is verified which of the two subjects, e_1 or e_2 has the highest value of knowledge, represented by c_1 or c_2, to establish who will be the transmitting or receiving subject. In this example the transmitter is the e_1. Then it is verified the activity that the two subjects e_1 and e_2 are doing. The activity a_1 which is the most important will be that of the transmitting subject e_1. For example, if $M_{ij} > M_{ik}$, means that when the transmitting subject e_1 (column) is performing the activity a_1, the coefficient of knowledge found in the transmission matrix M (which is *0.88*) is transmitted more effectively (directed) than that of the subject e_2 (row) that is performing the activity a_2. Once the coefficient of transmission is found, applies the eq. (1) and the knowledge c_2 of the subject e_2 is modified, that is increased to *0.88*.

The wanted matrix M has the knowledge transmission coefficients used in the previously mentioned knowledge transmission model, particularly that which represents the fraction of the knowledge that the transmitting subject, when performing the activity a_j, passes to the receiving subject when they perform the activity a_i. The equation $C_R = C_R + M_{ij}\,(C_T - C_R)$ summarizes this idea, where C_R is the knowledge of the receiver, the M_{ij} is the transmission matrix element (represented in Fig. 1 as an example by *0.88*) and C_T is the knowledge of the transmitter. The coefficients of the matrix are found by the GA.

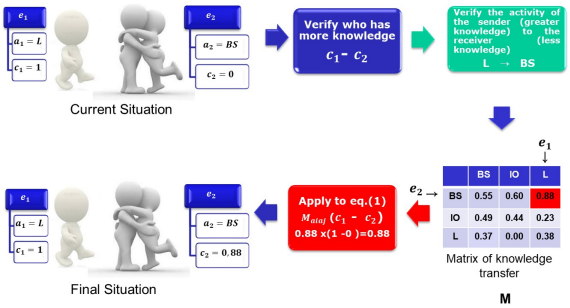

Figure 1. Example of knowledge transmission model

2.2. Observed Data and Notation Used

The proximity matrix $P^O \in [0,1]$ stores the information on the frequencies of the meetings between the N_S subjects, where P_{ij}^O indicates the probability of the subject S_i approaching the subject S_j. Note that P^O is a symmetric matrix. In this context, the proximity matrix is a graph representing the relationships of encounters between the subjects.

For each subject were observed which activities were being performed and in which place they happened. In the system, these data were used to find the behavioral profile of each subject S_i denoted by B^O, given by eq. (2):

$$B_i^O = \{U_i^O, V_i^O\},\ for\ i = 1,..,N_S \quad (2)$$

Where,

$U_i^O \in [0,1]^{N_A}$: is the profile of activities, and U_{ij}^O indicates the proportional time that the subject S_i spent doing the activity a_j;

$V_i^O \in [0,1]^{N_L}$: is the profile of sites, and V_i^O indicates the proportional time that the subject S_i spent on the site l_i.

Note that, because they represent proportions, it results in $\sum_{j=1}^{N_A} U_{ij} = 1$ and $\sum_{j=1}^{N_L} V_{ij} = 1$. We also defined the set of all profiles of the subjects as shown in eq. (3) given:

$$B^O = \{B_i^O\},\ for\ i = 1,..,N_S \quad (3)$$

In addition to the observed data regarding the proximity between the subjects and their profiles, the system also uses observations related to the proficiencies of each subject in performing the activity that directly expresses the related knowledge (e.g., breaking coconuts), referred here as specific

activity. These activities are observed at the beginning and at the end of the observation period and are called *initial proficiency* CI^O, and *final proficiency*, CF^O, respectively. The *initial observed proficiencies*, $CI^O \in [0,1]^{Ns}$, are calculated according to eq. (4):

$$CI_j^O = \frac{QI_j^O}{TI_j^O} \; for \; j = 1,..,N_S \quad (4)$$

Where,

QI_j^O: is the number of observations, at the beginning of the observation period, in which the subject S_j was successful in carrying out the specific activity;

TI_j^O: is the total number of observations, at the beginning of the observation period, in which the subject S_j performs the specific activity.

The final proficiencies observed, $CF^O \in [0,1]^{Ns}$ are calculated according to eq. (5):

$$CF_j^O = \frac{QF_j^O}{TF_j^O} \; for \; j = 1,..,N_S \quad (5)$$

Where,

QF_j^O: is the number of observations, at the end of the observation period, in which the subject S_j was successful in carrying out the specific activity;

TF_j^O: is the total number of observations, at the end of the observation period, in which the subject S_j performs the specific activity.

The observation periods for obtaining CI^O and CF^O should be of short duration (for example, a week or a month) compared to the period used to observe subjects' profiles (one or two years), because it is needed an initial measure of how much the subject knew about the specific activity being analyzed. Likewise, regarding the final period. Also, the period associated with CI^O should be located at the beginning of the observation or a little earlier. Similarly, CF^O should be located at the end or shortly after obtaining the profiles.

Note that the system is dealing exclusively with a single knowledge and that it is associated with the specific activity. Thus, while CI^O is used to represent the knowledge of each subject at the beginning of the period considered by the system, CF^O represents the knowledge of each subject in the end of the observation period. Therefore, the difference $CF^O - CI^O$ represents how much each subject learned during the period of interest.

2.3. The simulated environment

The simulated environment (as a platform or the proposed model) is composed of agents called e_1, e_2, e_3 and e_4, which represent the observed subjects, that are, the robust capuchin monkeys. The observed proximity data P^O, are the activities a_1, a_2, a_3 and a_4, which are carried out on site l_1 (coconut breaking site). The agents have the knowledge that is represented by c_1, c_2, c_3 and c_4. The simulated environment uses the actual data from the social model to govern the behavior of the agents in the simulator so that they are consistent with the observed real environment. Note that e_1 is close to e_2 and due to the proximity from the social network, or a relationship between the two obtained from the profile of the real data, they can meet up to 10 times during the simulation where there may be a transmission of knowledge of e_1 to e_2.

To provide greater coherence in the simulation (trying to avoid the occurrence of unforeseen situations), it was proposed the implementation of coherence rules that are verified at each iteration of the simulator to ensure that very unlikely events can be avoided. In this way, it is obtained a mechanism that allows the simulator to manage a behavior that is closer to what is expected. The current implementation of this concept is partial, applying to situations in which the activity requires a partner to be properly executed (Fig. 2).

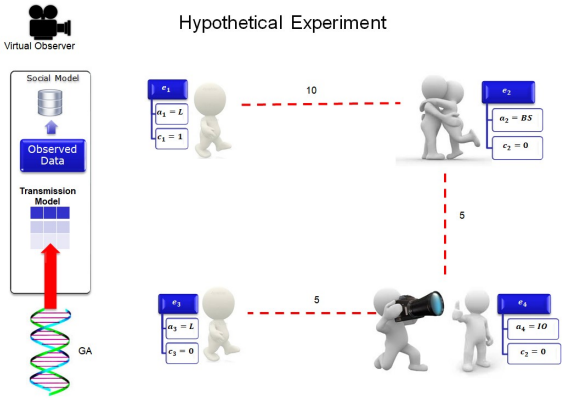

Figure 2. Simulated environment (developed platform)

Next, the social model, the agents used and how the simulation is performed will be detailed. Also, will be explained which and how the simulated data are extracted from the simulated environment that are annotated by a type of virtual observer to validate the simulator with the real environment.

2.4. The social model

Observed data on observed initial proficiency CI^O are used to define the amount of specific knowledge that each agent has at the beginning of the simulation. After performing the simulation, the performance of the social model used by the simulated environment is evaluated by comparing the observed final proficiencies CF^O and simulated final proficiencies CF^S which are read from the agents at the end of the environment run. The social model W stores the necessary data for the creation and execution of the virtual environment given by eq. (6):

$$W = \{CI^O, P^O, B^O, CF^O\} \quad (6)$$

Where,

CI^O: is the vector with the observed initial proficiencies;

P^O: is the matrix with observed proximities;

B^O: is the set of all the profiles obtained from the observations;

CF^O: is the vector with the observed final proficiencies;

CI^O: is the vector with the observed initial proficiencies.

2.5. The agent

In the simulated environment, each subject S_i is represented by an agent e_i. The state of each agent at the simulated time t is given by eq. (7):

$$e_i(t) = \{B_i^O, c_i(t), a_i(t), l_i(t)\}, \text{ for } i = 1, .., N_S \quad (7)$$

Where,

B_i^O : is the observed profile of the subject S_i;

$c_i(t) \in [0,1]$: is the knowledge of the agent in the instant t;

$a_i(t) \in A$: is the activity performed by the agent at the instant t;

$l_i(t) \in L$: is the place where the agent is in the instant t.

Note that the agent profile is obtained from the observed data and remains constant throughout the simulation. In addition to storing the current state, the agent is responsible for suggesting the activity he wants to perform and the location in which he wants to be at the time instant following the current one. Both wishes are chosen by following the statistics of the observed proximities noted in the agent's profile. It was built in this way to maintain consistency between the simulation and the real environment.

Given the subject S_i and its activity profile being given by B_i^O, the likelihood $Prob$ in S_i to choose the activity $a_j \in A$ for the moment $t+1$ is given by eq. (8):

$$Prob\ a_j = U_{ij}^O, \text{ for } j = 1, .., N_A \quad (8)$$

In the same way that the choice of an agent's activity suggestion is made based on the activity profile, the choice of the location suggestion is based on the location profile. Thus, given the subject S_i and being its locality profile given by V_i^O, the likelihood $Prob$ in S_i to choose the location $l_j \in L$ for the moment $t+1$ is given by eq. (9):

$$Prob\ l_j = V_{ij}^O, \text{ for } j = 1, .., N_L \quad (9)$$

Both the choice of activity desired by the agent for the next simulated iteration $t+1$, and the locality, are performed in a weighted way according to the occurrence of the activities and localities observed and that they do not depend on the agent's current activity or location. Finally, the simulator is the one who applies the knowledge transmission model when the agents meet.

2.6. The simulation

The simulation of the environment follows the observed data contained in the social model defined in eq. (6). For such, N_S are generated. It is important to emphasize that the agents are initiated with the observed initial proficiencies (obtained from the real environment), defined in eq. (4). This algorithm has three important points:

- **The choice of agent:** For each iteration of time in the simulator, the sequence in which the agents are selected to act is different, thus avoiding that the sequence of actions creates addictive dynamics in accordance with the position the agents have in the vector that stores it.
- **The choice of state suggestion by the agent**: After an agent has been chosen, the environment demands to present a suggestion of state for the next iteration. This state is composed of an activity and a place. However, some activities can only be performed with the participation of a second agent. For such activities the environment is responsible for choosing which agent will play the role of partner. It should be emphasized here that the effectiveness of the action in the context of this simulation does not require that it be effectively implemented (as said before, they are abstractions of the real environment).
- **The choice of the partner:** For the selection of the partner of an agent, the observed proximity matrix P^O is seen as a match probability matrix. Thus, the probability $Prob$ of agent e_j is chosen as partner of agent e_i is given by eq. (10) (Note that P^O is a symmetric matrix):

  $$Prob\left(e_j = partner \,|e_i\right) = P_{ij}^O = P_{ji}^O \quad (10)$$

  After the partner e_j is chosen it is checked if it is already performing some activity somewhere. If not, the environment places it in the same e_i and asks him to choose an activity that is consistent with that of e_i. If e_j is already allocated, but their situation is consistent with the situation of e_i, then the environment asks the e_j to fit to the situation of e_i. However, if e_j is already in some situation contradictory to the one e_i is proposing, a new partner is chosen. The process continues until some partner is willing to e_i following its proposal of activity and location. If no partner is found, the environment asks the e_i to choose a new activity and location suggestion.
- **Comments:** Note that the choices of future states performed by agents follow the probabilities given by the observed profiles and that the matches between the agents follow the probabilities provided by the observed proximity matrix. Thus, these behaviors in the simulated environment reflect the behaviors observed in the real environment.
- **Coherence rules:** An important point for the validation of the states suggested by the agents is the coherence of the suggested states, whether it is related to the state itself or to the social relations of

the simulated environment. As an example of the lack of coherence of the state, one can imagine a situation where the agent is engaged in the activity of eating in a place that is known to have no food. As an example of the lack of coherence of states related to social relationships, one may think of a situation in which one agent is engaging in the activity of alienation while his partner is engaging in another incompatible activity as fighting. Such rules of consistency are established by the environment that does not allow this type of situation.

- **The transmission of knowledge:** In the simulated environment whenever the agent e_i is with another agent e_j shall happens the transmission of the knowledge of the agent that has the greater knowledge in a specific competence to the other. Knowledge is transmitted by considering the activities that the agents are performing at the time of the encounter and is given by eq. (1) having the agents e_i and e_j like the subjects S_i and S_j, respectively. Note that the matrix M, which stores the coefficients used by the knowledge transmission equation is expressed by eq. (1), although it is only used by the simulator at this moment. This matrix is of fundamental importance in determining the flow that knowledge will have through the agents.

Initially, examples of the initial and final knowledge, the profile and the proximity data that are obtained through the data observed for each of the agents involved in the simulation are presented. These data are the rules of coherence that establish the constraints that will be used by the simulator to probabilistically select the location, activity and partner at each instant of time of the simulation. As the encounter between the two agents occurs, the equation of the knowledge transmission model that seeks the knowledge transmission coefficient in the hypothetical matrix is applied according to the activities being carried out by the transmitting and receiving agents, thus modifying, the knowledge of the receiving agent. The virtual observer, mentioned in the next section, (seen in Fig. 2) has the task of reading the data of the simulated environment to verify if they agree with the observed real environment.

2.7. Simulated Environment Observation

During the execution of the simulation, some information are accumulated as if observers were collecting data from the virtual environment. Such information follows the same logic of information observed in the real environment and is used to govern the virtual environment. The meetings between the agents in the virtual environment are annotated to compose the matrix according to eq. (11) given:

$$P^S \in [0,1] \quad (11)$$

The simulated proximity matrix P^S stores the proportion of encounters between two agents during the simulation, with P^O as its observational counterpart. The simulated profiles of each agent are also read from the simulated environment following the eq. (12):

$$B_i^S = \{U_i^S, V_i^S\} \text{ for } i = 1,..,N_S \quad (12)$$

At where U_i^S is the simulated activity profile and V_i^S is the simulated site profile. Thus, B_i^O has B_i^S as their simulated counterparts. Also, the set of all simulated profiles is given by eq. (13):

$$B^S = \{B_i^S\} \text{ for } i = 1,..,N_S \quad (13)$$

The simulated proximity data P^S and simulated profiles B^S are used in the investigation of the coherence of the simulator with the social model that governs it (if it agrees with the real environment). This inquiry is important to know how much the consistency rules implemented in the simulator deviates from the real environment.

In the simulated environment, the observed initial proficiencies CI^O are used to configure the initial conditions of the agents. The final proficiencies observed CF^O refers to the observed knowledge of subjects at the end of the observed period. The virtual parallel of CF^O is represented by eq. (14):

$$CF^S \in [0,1] \quad (14)$$

CF^S can be obtained at any instant of time t of the simulation by eq. (15):

$$CF_i^S = c_i(t) \text{ for } i = 1,..,N_S \quad (15)$$

Where,

$c_i(t)$: is the knowledge accumulated by the agent e_i at moment t, as described in eq. (7).

2.8. Exposure of the problem

The observed data are used to build the social model W which together with the knowledge transmission model M constitute the basis for the simulation behavior in the virtual environment. While W is used by the simulator to maintain the coherence of the simulation with the observed data, M is used to implement the knowledge flow between agents. Note that W and M remain constant throughout the simulation. The simulator's decisions to iterate the virtual environment use probabilities and non-linearities (coherence rules), and therefore, different executions generate different results, even with identical initial parameters and conditions. In this way, each execution of the simulator is represented separately. The eq. (16) presents these parameters.

$$C_r^M(t) = R_r(M, W, t) \text{ for } t = 1,..,N_t, r = 1,..,N_r \quad (16)$$

Where,

N_t : is the number of iterations for each run of the simulator;

N_r : is the number of times the simulator is executed with the same parameters M and W;

$C_r^M(t) \subset [0,1]^{N_S}$: are the proficiencies of the agents at the instant of time t of the execution r of

the simulator when the transmission model M is it used.

The model M that is being elaborated should be such that the simulator runs R_r find the simulated proficiencies $C_r^M(t)$ close to the final proficiency observed CF^O consistently in relation to the r simulator runs. For this, the set $H^M = \{H_r^M\}_{N_r}$, $H_r^M \in [0,1]^{Ns}$ containing the proficiencies that are closest to those expected in each execution, being eq. (17):

$$H_r^M = min_{C_r^M(t)} \| C_r^M(t) - CF^O \|, \text{ for } r = 1,..,N_r \quad (17)$$

At this point, we have the set of results H^M which represents the best results found by the simulator using the model M. The comparison between the elements of H^M and the expected result CF^O is made using two criteria:

The first criterion is the distance δ^M between CF^O and the midpoint of H^M:$\mu^M = \frac{1}{N_r}\sum^{N_r} H_r^M$

$$\delta^M = \| \mu^M - CF^O \| \quad (18)$$

Where,

$\mu^M \in [0,1]^{Ns}$: is the midpoint of H^M.

The second criterion is the standard deviation σ^M of the H^M given by eq. (19):

$$\sigma^M = \sqrt{\frac{1}{N_r}\left[\sum_r^{N_r} \| H_r^M - \mu^M \|^2\right]} \quad (19)$$

Having established the two criteria for the evaluation of the knowledge transmission model, the problem that establishes the conditions to find the solution model, $Erro$, is defined by eq. (20):

$$Erro(p,M) = min_M\{p\sigma^M + (1-p)\delta^M\} \quad (20)$$

Where,

$p \in [0,1]$: is the influence of the standard deviation.

Note that eq. (20) allows us to use the distance and standard deviation criteria in a weighted way in determining the solution $Erro(p,M)$. The higher the value of p, the greater the convergence requirement of the different simulator executions in relation to the importance of the convergence to the observed result.

2.9. Problem resolution

The problem of finding the matrix with the coefficients for the transmission of a specific knowledge defined by eq. (20) makes use of a non-deterministic computational simulator that makes difficult, if not impracticable, the use of an analytical approach for its resolution. Therefore, we chose to use the computational technique of GA. Because it is a known and widely used technique, the theoretical details of the GA will not be described. In the next section, it is presented how the algorithm was used in this work and the main decisions taken to adapt it to the problem in question.

2.10. Genetic Algorithm (GA)

The adaptation of the problem to the GA is performed by the evaluation procedure, described later. Also, the procedures selected, crossover, mutation and local mutation, although they are common to GA, will be detailed so that one can perceive the adopted conceptual decisions.

2.11. Adapting the Problem to the GA

Adapting the problem to GA implies knowing how the matrix M is generated from a genotype and on how the simulator is used to find a scalar value that evaluates M to be used by GA to represent the genotype capacity that generated M. The genotype is represented by the vector $g \in [0,1]^{N_G}$ with the number of genotypes $N_G = N_A^2$. The transposition of g for the knowledge transmission model M is performed directly in eq. (21):

$$M_{ij}(g) = g_K, \text{ where } k = (i-1)N_A + j, \ \forall\ i,j \in [1,N_A] \quad (21)$$

Theevaluationorfitnessofgenotype performance g is performed by the training function: Cap: $[0,1]^{N_G+1} \mapsto [0,1]$. This function uses eq. (17) to obtain the conformance *error* of: $M(g)$ with the observed data, given by eq. (22):

$$Cap(g,p) = 2 - Erro\big(p, M(g)\big) - \frac{1}{N_A^2}\sum M_{ij}(g) \quad (22)$$

In conclusion, the algorithm for implementing the genotype evaluation procedure, which is implemented by eq. (22), sets a value for the weight of the standard deviation p. Fig. 3 summarizes the weighting criteria used for the fitness assessment of the knowledge transmission matrix to verify the lowest error found.

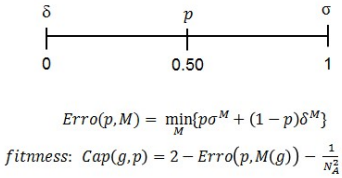

Figure 3. Weighting of the evaluation criteria to find the smallest error

Before describing the genetic operators, it is important to emphasize that they use the $r : \mathbb{R}^2 \mapsto \mathbb{R}$ to represent the generation of random numbers with uniform distribution. It is ensured that $x \in [a,b)$ to any $x = r(a,b)$. Note that $r(a,b)$ results in a different value each time it is referenced.

Selects: This procedure is responsible for selecting a genotype from a set of genotypes (population) with probabilities proportional to their abilities. There are $P = \{g_i\}_{N_P}$, the set of genotypes $g_i \in \mathbb{R}^{N_G}$ population size N_P and $f : \mathbb{R}^{N_G} \mapsto \mathbb{R}$ the function that indicates the qualification of the genotypes. First, we calculate the vector $q \in \mathbb{R}^{N_G}$ with the pertinence intervals, using:

$$q_i = 0,\ q_{i+1} = q_i + f(g_i),\ \text{for}\ i = 1,..,N_P$$

Then the random value is chosen $r_x = r(0, q_{N_P+1})$ which is used to select the genotype $s \in g$ according to eq. (23):

$s = g_i$, where i is such that $q_i < r_x < q_{i+1}$ (23)

Crossover: this genetic operator is responsible for generating a genotype resulting from the mixing of two others. In the adopted crossover procedure, the uniform crossover was used. Be the genotypes $u, w \in \mathbb{R}^m$, the genotype $g \in \mathbb{R}^{N_G}$ resulting from the application of the crossover operator is given by eq. (24):

$$g_i = \begin{cases} u_i, & \text{if } r(-1,+1) < 0 \\ w_i, & otherwise \end{cases} \text{ for } i = 1,..,N_G \quad (24)$$

The vector *r[ ]* was generated randomly with the purpose of serving the choice of the elements that will compose the final vector. If the *r[ ]* is negative, then the corresponding element of *u[ ]* is copied to *g[ ]*, if it is positive, then the corresponding element of *w[ ]* is the one copied to *g[ ]*.

Mutation: this genetic operator is responsible for generating the genotype resulting from the mutation of one or more genes of the original genotype. Be the genotype $u \in \mathbb{R}^{N_G}$ and the probability of mutation of a specific gene, $p_g \in [0,1]$, then the genotype $g \in \mathbb{R}^{N_G}$ resulting from the application of the mutation given by eq. (25):

$$g_i = \begin{cases} g_i & ,if\ r(0,1) < p_g \\ r(0,1), & \text{otherwise} \end{cases} \text{,for } i = 1,..,N_G \quad (25)$$

Local mutation: this genetic operator is responsible for generating the genotype resulting from the local mutation of one or more genes of the original genotype. While the single mutation produces new genes using the whole domain of the gene, in this case the interval [0,1], the local mutation produces mutations restricted to a range around the current value of the modified gene. Be the genotype $u \in \mathbb{R}^{N_G}$, the probability of mutation of a specific gene $p_g \in [0,1]$, and the size of the range $s \in [0,1]$. Then the genotype resulting from the application of the local mutation, $g \in \mathbb{R}^{N_G}$ is given by eq. (26):

$$g_i = \begin{cases} g_i, if\ r(0,1) < p_g \\ r(max\{0, g_i - s\}, min\{1, g_i + s\}), otherwise \end{cases}, \text{for i} = 1,..,\text{N}_\text{G} \quad (26)$$

3. Analysis of results and conclusion

Data used as a reference for the simulation were captured from December 2, 2011 to April 5, 2012. A more detailed description can be found at [8]. Initially a hypothetical situation will be presented with the observed data generated by the simulation for a better understanding of the data and the simulation procedure. Then, a situation will be shown using real observed data, although intermediate, with few activities and few subjects, and therefore it becomes hypothetical. Finally, the situation generated by the simulator will be presented with all the data observed. The computer used for the tests was an Asus notebook with an Intel i3 - 236m processor, CPU 1.40 GHz, 2GB memory and 64 bit Windows 8.1 Pro operating system. The programming language used for the development of the Horacio and Cuzco platform was Java and the programming environment was NetBeans. Cuzco software used the MySql database and workbench development environment. All software used for development is free.

3.1. First experiment

The first hypothetical experiment sought to be as simple as possible to test the Horacio platform and explain the operation of the proposed model to validate the tool. Four agents and three activities were used. The parameters used for the experiment were: 200 steps (simulation interval), number of rounds: 20 (number of times the simulator will run the experiment), number of genotypes: 40 (individuals from each population), number of generations: 1000 (Population), standard deviation weight: 0.600000, range: 0.10, lowest error found for each round: 0.001677, fitness: 1.960530 (best individual from each population), the standard deviation was: 0.009072, the mean at: 0.034226, and the running time of the experiment lasted 5 minutes and 24.716 seconds (this data can be seen in the listing of the experiment generated by the simulator in [8]).

The simulation uses the (e_1, e_2, e_3, e_4) with the initial knowledge observed (CI^O) given by (*Ca* 1,0000, *Cp* 0,0000, *Me* 0,0000, *Ml* 0,0000), respectively and final knowledge observed (CF^O) given by (*Ca* 1,0000, *Cp* 0,8000, *Me* 0,0000, *Ml* 0,4000). The matrix of activities (*U*) given by (*BS*, *IO*, *L*) and profiles (*V*) of each agent given by V_1 (0.0, 0.0, 1.0), V_2 (1.0, 0.0, 0.0), V_3 (0.0, 0.0, 1.0) and finally V_4 (0.0, 1.0, 0.0). Note that by the proximity matrix, e_1 only meets e_2; e_2 meets with e_1 and e_4; e_4 meets with e_2 and e_3; and e_3 only meets e_4. Also, e_1 only performs the activity *L*, e_2 only performs the activity *BS*, e_3 only realizes *L*, and e_4 only realizes *IO*. As at the beginning of the simulation only e_1 has the knowledge and it only encounters e_2, in this possible meeting, e_1 can only be performing *L* and e_2 can only be performing *BS*. Thus, the transmission coefficient of the knowledge of *BS* for *L* should be significant (it should allow e_2 obtain 0.123 of the knowledge coefficient of e_1). Following the same reasoning, between the e_2 and e_4, we have that the coefficient of e_1 passed to e_2, is also transmitted to the e_4 obtaining 0.185 knowledge of the e_2. Finally, the relationship between e_4 and e_3 shows that the transmission coefficient of *IO* for *L* does not transmit anything to e_3 (final knowledge: 0.000) and therefore should be null.

The experiment tested in the Horacio platform through several simulations shows that the GA converged to find the knowledge transmission matrix

and the simulation found the best matrix (the best individual in a population) when it reached knowledge CF^S next to CF^O. This was already expected because the simulator was built for this. Given a coherent input and output data the simulator must find an array that leads from one to the other. The validation is done by the convergence of the simulated data to the real ones. The execution time was the necessary to observe such convergence. This data (observed final knowledge) was provided by the LEC of IPUSP.

The proposal is to verify if the model allows such convergence, that is, if the simulation leads to the finding of the final observed knowledge. In this case, we present the coefficients (found by the GA) of the knowledge transmission matrix that highlight the peripheral activities that can influence the knowledge exchange of a specific activity (breaking coconuts, in this case).

After finding these coefficients of the transmission matrix, the equation of the knowledge transmission model is applied to quantify the knowledge of the agents. Thus, the best knowledge transmission matrix (using the GA) was found through the tests performed by the simulations, to be used in the equation of the proposed knowledge transmission model to validate the Horacio platform.

The first hypothetical situation presented can validate the proposed platform, which was confirmed by the experiment. The tool can propose a scenario consistent with the observed data and presents a way in which a certain knowledge can have its transmission aided by (peripheral) activities that apparently have no relation to specific knowledge (breaking coconuts).

3.2. Second experiment

The second intermediate experiment used the actual data observed from the PET island to test the Horacio platform. We used 6 agents and 4 activities, thus, it also becomes a hypothetical experiment (partial set). The parameters used for the experiment were: 1500 steps (simulation interval), number of rounds: 10 (number of times the simulator will run the experiment), number of genotypes: 40 (individuals from each population), number of generations: 5000 (populations), standard deviation weight: 0.600000, scale: 0.10, lowest error found for each round: 0.112615, fitness: 1.927509 (best individual from each population), standard deviation was: 0.020858, the mean at: 0.020418, and the run time of the experiment lasted 2 hours, 56 minutes and 36.363 seconds. (The data can be seen in the listing of the experiment generated by the simulator in [8]).

The simulation uses the agents (*Cz*, *Nd*, *Nl*, *Nr* and Sg) with the initial knowledge observed (CI^O) given by (*Cp* 1,0000 *Cz* 0,0000 *Nd* 0,0000 *Nl* 0,0000 *Nr* 0,0000*Sg*0,0000),respectivelyandfinal knowledge observed (CF^O) given by (*Cp* 1,0000 *Cz* 0,5000 *Nd* 0,0000 *Nl* 0,7000 *Nr* 1,0000 *Sg* 1,0000). The simulated final (*Cp* 1,0000 *Cz* 0,4985 *Nd* 0,1720 *Nl* 0,4406 *Nr* 0,9585 *Sg* 0,9883). The knowledge transmission matrix found is presented below. The activities (*U*) and profiles (*V*) of each agent can also be observed [8] generated by the simulator.

The value found for the element of the knowledge transmission matrix, referring to the agent who performed the "*Eat Provided Food*" (transmitting) activity while meeting another who performed the "*Hug*" (receiver) activity, doing both part of the same social network (having a link through the proximity matrix), was *0.013*. This means that this was the value used in the simulation by the equation of the proposed knowledge transmission model, at each occurrence of this situation. The same can be applied to the agents that performed the "*Social Play*" activity (sender) when they found other agents that were performing the "*Alert*" activity (receiver) using a coefficient of 0.295 at each encounter. The coefficient of *0.019* was also used for the transmission of knowledge of the agents that carried out the "*Social Play*" activity (sender) for the agents that performed the "*Eating Supplied Food*" (receiver) activity.

The second experiment tested on the Horacio platform through several simulations showed that the GA converged to find the best knowledge transmission matrix when it reached the knowledge CF^S next to CF^O. This second hypothetical situation using real data with few subjects and few activities can also validate the proposed platform as presented in the first hypothetical experiment.

3.3. Third experiment

In this third experiment, data from the PET, which consists of 38 robust capuchin monkeys observed in their daily activities with a repertoire contemplating 26 activities were considered. The parameters used for the experiment were: 800 steps (simulation interval), number of rounds: 10 (number of times the simulator will run the experiment), number of genotypes: 20 (individuals from each population), number of generations: 3000, weight of standard deviation: 0.600000, range: 0.10, lowest error found for each round: 0.2771753, fitness: 1.879682 (best individual from each population), standard deviation was: 0, 001145, the mean at: 0.010941, and the runtime of the experiment lasted 7 hours, 38 minutes and 6.819 seconds. All data can be seen in the listing of the experiment generated by the simulator in [8].

Naturally, convergence depends heavily on input data. In this case the dimension of the problem has increased considerably, impacting the time spent by the simulator in the search for convergence. Other factors such as the diversity of relationships established in the social network, and the diversity of

activities, characterize the scenario as complex. In this case, the search space for the solution reaches such a magnitude that it becomes difficult, if not highly improbable, to find a solution. In addition, it is worth emphasizing that there are many variables that were not considered in the model, which can also contribute to its imprecision and consequently to be able to find an appropriate matrix. That is, the phenomenon of social transmission of knowledge of the real environment is much more complex. An intermediate experiment may present better results by working with fewer subjects and fewer activities.

Since the simulator uses probabilities, each execution generates different results, even with identical initial parameters and conditions. In these cases, the ideal is to repeat the experiment several times and analyze the coherence (proximity) of the values found. If the standard deviation is small, then the mean serves as a good result. Therefore, we conclude that we need more input data related to the monkey profile (such as gender, breaking coconut activity, age, affiliation, among others) in order to obtain a better result when the analyzed number is larger.

4. Conclusions and Future Work

The developed simulator creates a virtual environment composed of virtual agents that represent the observed subjects of a real society that one wishes to analyze. The simulator functions as a virtual experiment lab where situations that have not yet occurred or are difficult to observe could be tested. The purpose of the work was to propose and analyze a social model of knowledge transmission that uses as base the activities that the subjects are performing daily. The social model assists in the transmission of knowledge from one subject to the other only when they are close, thus establishing a kind of relationship (connecting link) between the subjects. For the simulation to be coherent with the real environment, the behavior of the simulated agents is based on statistical rules using the data observed from the RE.

It was presented a hypothetical first situation with few subjects and few activities to debug and validate the proposed platform, which was confirmed by experiments 1 and 2. The tool can propose a scenario consistent with the observed data and presents a way of a certain knowledge may have its transmission aided by (peripheral) activities that apparently have no relation to specific knowledge (breaking coconut).

Although the profile used considered only the subjects' activities, the results showed the validity of the proposed concept. This occurred, even though other information about the profile, such as age and gender, influences the transmission of the knowledge in question. It is worth mentioning that the input data (observed) used in the simulation determine the results, so that any inconsistencies in the input data compromise the process. Likewise, a small number of observations also seems to be critical to the success of the simulation procedure. A greater amount of data also imposes problems as it becomes more difficult to converge the simulation to an adequate result (since the search space for the solution reaches such a magnitude that it becomes difficult, if not highly unlikely to find a solution) and the decision-making of the Horacio platform requires more profile data for a better result (variables that were not considered in the model, for example, gender, age, etc.) as shown by the third experiment with all subjects and all activities.

It is important to note that we are only considering the transfer of knowledge from one subject to another through the proximity of the two. In the system there is no increase (quantification) of knowledge (specific capacity) when the subject performs an activity alone.

For the Horacio platform, knowledge transfer occurs only when the activities performed are dependent on social situations related to the encounters. More explicitly, when the social relationship is considered to be important for the transmission of a knowledge (which is related to the profile of the agents based on the observations of the RE), however, it is not any knowledge that makes sense to be transmitted; depends on the situation in which it is being simulated.

In practice, the view that the "contagion" of knowledge is linked to the meetings made possible by social dynamics allows us to think better about the relations between the social structures we characterize and how (and what) is transmitted; which would otherwise only be perceived by other subjects through some type of observation tool, for example, social networks of friends used on the internet or television (in the case of human societies).

The understanding of how knowledge is transmitted is directly related to the understanding of social and environmental relations. In other words, social relations and environmental situations facilitate or not the transmission of knowledge, although not all relations or situations can influence it.

Considering the influence of the activities on the transfer of knowledge and the bibliographical survey in primatology, it was noticed that other data should be considered by the platform, related to the profile of the subject like gender, the age group, the affinity with other subjects, etc. The current profiles of the agents in this work contains observed data about the activities performed daily that are used to govern the system according to the RE, while were disregarded, in this implementation, the other mentioned factors. This is something that should be incorporated into an upcoming revision of this platform.

According to [9] research at the PET island, two hypotheses may influence the exchange of

knowledge transmission: (1) the social tolerance that presents the intimate relations between two subjects using the vertex degree (representing the number of subjects to which each monkey is connected). The AGM finds only the strongest relationships. This measure reflects how much the subject is tolerated, or can also be seen as the most popular subject; and (2) the observation of proficient coconut breakers, which according to [10], the choice of the subject's target of observation in the coconut break activity is associated with a series of links between the social group (e.g., affiliation and affinity), but research also shows the strong relationship between the observation of proficient subjects and the consequent ability to acquire such proficiency [11].

In the results of this work, we did not observe the activity of coconut breaks observed in the data observed in the field, which, according to the bibliographic survey, is one of the factors that influence and a lot, in the transmission of knowledge.

Finally, the Cuzco tool developed for the LEC of IPUSP allows researchers to replace the currently manual processes (recorded using spreadsheets) and in papers, by electronic information flows. Considering the current process, this results in data redundancy, inflexibility, low level of security and difficulty in sharing between softwares (spreadsheet and software that generates AGM). It has a set of interrelated components that collects (or retrieve), process, store and distribute information to support decision making, for example, the generation of dynamic multi-filter FGM, of various epochs and regardless of the focus of the work or the search, the system stores the information centrally, keeping a history of the information for later analysis.

It is worth emphasizing that the two softwares can and should be used together, being the first (Cuzco) the identification of a subgroup of study and the second (Horacio) the use of this same subgroup to analyze the influences established between the several competences as presented in the hypothetical experiment 2.

In general, we believe that the contribution of this work can be adjusted to apply not only in primatology, as the case of this study, but also to other scenarios in which social situations can serve as support for their elements to the development or improvement of skills. The proposed knowledge transmission matrix correlates the different competencies and influences that one has over others and the way they develop. It was considered only a specific knowledge (coconuts break), but nothing prevents other knowledge from being considered, since the software can be expanded to cover any amount of knowledge. This, however, can be implemented in future work.

5. References

[1] IZAR, P. Female social relationships of Cebus apella nigritus in a southeast atlantic forest: an analysis through ecological models of primate social evolution. Behavior, vol. 141, p. 71-99, 2004.

[2] ALTMAN, Jeane. Observational Study of Behavior: Sampling Methods. Behavior, v 49, Issue 3. 1974, p. 227-266. ISSN: 0005-7959.

[3] RESENDE, B. D. Ontogenia de comportamentos manipulativos em um grupo de macacos-prego em situação de semiliberdade (Tese de Doutorado), USP, 2004.

[4] RESENDE, B. D.; IZAR, P.; OTTONI, E. B. Social play and spatial tolerance in tufted capuchin monkeys (Cebus apella). Revista de Etologia, v. 6(1), p. 55-61, 2004.

[5] LOULA, A. C. Comunicação simbólica entre criaturas artificiais: um experimento em vida artificial. 2004. Dissertação (Mestrado), Universidade Estadual de Campinas, Engenharia Elétrica, Campinas, São Paulo, 2004.

[6] COUSSI-KORBEL, S.; FRAGASZY, D. M. On the relation between social dynamics and social learning, Animal behaviors, v. 50, p. 1441-1453, 1995.

[7] RINALDI, L. C. A.; MOREIRA, R. B. T.; RESENDE, B. D. de; NETTO, M. L. A System for Social Network Analysis. In: The 8th International Conference for Internet Technology and Secured Transactions (ICITST-2013), 2013, London, UK, 2013. p. 126-131.

[8] RINALDI, L. C. A. Transmissão de Conhecimento Coerente com Comportamentos Observados: Análise de Redes Sociais. Saarbücken, Deutschland: Verlag, Novas Edições Acadêmicas. 2017. v1. p. 168.

[9] CARVALHO, M. P. Influência social sobre a aprendizagem do uso de ferramentas em um grupo de macacos prego que apresenta o uso espontâneo de quebra de cocos (Trabalho de Iniciação Científica). Orientadora: Briseida Dogo de Resene. Instituto de Psicologia da USP. São Paulo, 2011.

[10] COELHO, G. C. Observação por co-específicos e influências sociais na aprendizagem do uso de ferramentas para quebrar cocos por macacos prego (cebus sp) em semi-liberdade. Dissertação de Mestrado. Instituto de Psicologia da USP, São Paulo, 2009. 147 p.

[11] OTTONI, E. B.; RESENDE, B. D.; IZAR, P. Watching the Best nutcrackers: wat capuchin monkey (cebus apella) know about others tool-using skills. Journal of Animal Cognition, v. 8, p. 115-219, 2005.

6. Acknowledgements

We would like to thank Continental A. G. and CNPq for the scholarship through the program FUSP to Luciene Cristina Alves Rinaldi. We also thank Marcio Lobo Netto, Mauro Enrique de Souza Muñoz, Rodrigo Bossini Tavares Moreira, Briseida Dogô de Resende, Dinamérico Alonso Gaspar and Ronaldo Gonçalves dos Santos.

