UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

DANIEL SCAIN FARENZENA

The cost of search and evaluation in problem-solving social networks: an experimental study

Thesis presented in partial fulfillment of the requirements for the degree of Doctor of Computer Science

Advisor: Prof. Dr. Luís C. Lamb
Coadvisor: Prof. Dr. Ricardo M. Araújo

Porto Alegre
June 2016

CIP – CATALOGING-IN-PUBLICATION

Farenzena, Daniel Scain

The cost of search and evaluation in problem-solving social networks: an experimental study / Daniel Scain Farenzena. – Porto Alegre: PPGC da UFRGS, 2016.

96 f.: il.

Thesis (Ph.D.) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2016. Advisor: Luís C. Lamb; Coadvisor: Ricardo M. Araújo.

1. Social Computing. 2. Human-Computer Interface. 3. Game Theory. 4. Social Experiments. I. Lamb, Luís C.. II. Araújo, Ricardo M.. III. Título.

"For a successful technology,
reality must take precedence over public relations,
for Nature cannot be fooled."
— RICHARD FEYNMAN

AGRADECIMENTOS

Gostaria de agradecer, primeiramente, ao meu orientador, Prof. Luís da Cunha Lamb. Para a realização da presente tese, foram oferecidas à mim as melhores oportunidades que um pesquisador poderia ter. Estas oportunidades não são fruto apenas da dedicação do Prof. Lamb, como orientador, ao longo do meu doutorado, mas também de sua dedicação para com o Instituto de Informática e o seu comprometimento com ensino e pesquisa de excelência. Esta tese não teria sido possível sem a dedicação do Prof. Lamb. *If I have seen further, it is by standing on the shoulders of giants.*

Também agradeço ao meu co-orientador Prof. Ricardo Matsumura de Araújo, que abriu as portas do conhecimento para o estudo do comportamento humano através das Redes Meméticas Artificais. As contribuições do Prof. Ricardo para os artigos e para o desenvolvimento da tese foram fundamentais para o desenvolvimento do presente trabalho.

Agradeço à ProfªArpita Ghosh e ao Prof. Jon Kleinberg, que me receberam em seu grupo de pesquisa e compartilharam do seu conhecimento para que o presente trabalho pudesse ser aprimorado e realizado.

Aos professores e funcionários do Insituto de Informática que, junto com o Prof. Lamb, fizeram do Instituto de Informática um motivo de muito orgulho. Os professores e funcionários do Instituto de Informática honram suas carreiras e são exemplos de profissionais e servidores públicos.

Desejo agradecer também à minha família e amigos, que compreenderam o tamanho do desafio e tiveram a paciêcia de esperar até a conclusão desta longa jornada.

Finalmente, agradeço às agências de fomento, CNPq e CAPES, pelo suporte financeiro provido, permitindo que esta tese pudesse ser realizada com maior dedicação.

ABSTRACT

Online networks of individuals have been used to solve a number of problems in a scale that would not be possible if not within a connected, virtual and social environment such as the internet. However, the quality of solutions provided by individuals of an online network can vary significantly thus making work quality unreliable.

This dissertation investigates factors that can influence the quality of the work output of individuals in online social networks. Specifically, we show that when solving tasks with small duration (under 5 minutes), also known as microtasks, individuals decision making will be strongly biased by costs of searching (and evaluating) options rather than financial or non-financial incentives. Indeed, we are able to show that we can influence individuals decisions, when solving problems, by rearranging elements visually to modify an the search sequence of an individual, be it by designing the virtual work environment or manipulating which options are first shown in non-controlled environments such as the Amazon Mechanical Turk labor market.

We performed several experiments in online networks where individuals are invited to work on tasks with varying degrees of difficulty within three settings: mathematical games with objective truth (Sudoku and SAT instances), surveys with subjective evaluation (public policy polling) and labor markets (Amazon Mechanical Turk).

We show that the time spent solving problems and the user interface are more relevant to the quality of work output than previous research have assumed and that individuals do not change this behavior while solving the sets of problems. Finally, to complement our study of online problem-solving, we present additional experiments in an online labor market (Amazon Mechanical Turk) that agrees with our networked experiments, shedding new light on how and why people solve problems.

Keywords: Social Computing. Human-Computer Interface. Game Theory. Social Experiments.

LIST OF ABBREVIATIONS AND ACRONYMS

AGT	Algorithmic Game Theory
AI	Artificial Intelligence
AMT	Amazon Mechanical Turk
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
CPD	Computational Protein Design
GPU	Graphics Processor Unit
HC	Human Computing
HIT	Human Intelligence Task
QA	Question and Answer
OCPS	Online Collaborative Problem-Solving
OCR	Optical Character Recognition
PF	Protein Folding
AMN	Artificial Memetic Networks
SC	Social Computing
SIMD	Single instruction, multiple data
SoC	System-of-Chip

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

In the past decade, humans are increasingly spending more time in online, networked environments, such as online social networks, collaborative work websites or playing multiplayer games (DAVID; JON, 2010). Therefore, given the amount and relevance of human activities performed in online environments, we need to be able to design information systems capable of optimally harvesting human resources available online. However, are there good models to explain how humans behave in online environments? What are important factors to consider when designing systems where humans interact and work? This dissertation aims to investigate important factors that can influence human decision making in online environments and that can can be used to explain online human behavior and possibly rules on the design of information systems.

While the change in behavior from the offline to the online world is happening in networked environments, provided by an immense infrastructure composed of connected electronic devices (ABOWD; MYNATT, 2000), data from human activity is being collected and made available to the network hosts, becoming inexpensive to observe human behavior on a larger scale. As a consequence, we can take advantage of the this scenario and collect empirical evidence on the decision making process of individuals. In this dissertation, as a methodology to achieve our goal of understanding online decision making, we performed several experiments and collected data in non-controlled environments to provide empirical evidence about how humans behave in online environments and a new model based on *Artificial Memetic Networks*.

Within this scenario of massive, cheap and available data, new behavioral phenomena has been observed and online applications are becoming as successful as to affect a number of individuals comparable to the total human population, raising the interest from academia, industry(AHN; CATHCART, 2009) and government. Examples of successful applications are reCAPTCHA, FoldIt and Amazon Mechanical Turk (AMT). Another example: *Quora* is a web site where individuals can ask questions and have them answered, edited and organized by the other users of the website.(PAUL; HONG; CHI, 2012) As much as artificial intelligence have progressed when it was first envisioned a few decades ago , no artificial systems based on AI can properly read natural language questions and answer them in a satisfactory way as to surpass the popularity of Quora or other Question and Answer (QA) websites. Therefore, online human behavior studies are receiving special attention and we will model some of our experiments after applications where humans solve tasks, specifically Boolean Satisfiability Problems (SAT) problems and Sudoku game instances. Both types of tasks have equivalent mathematical descriptions and can be used to compare solving performance and which traits influence how they decide to solve the tasks among many possible ways.

There are many research areas dedicated to understand online human behavior. One way to understand how humans can achieve online problem-solving is to look at human cooperative behavior. It is the process where individuals work or act together for common or mutual benefits (FEHR; FISCHBACHER, 2003). For a more detailed description of what cooperative behavior is, we could choose one among several fields which will provide their own definitions, whether with rigorous mathematical models like in Economics (OSBORNE; RUBINSTEIN, 1994; OHTSUKI et al., 2006) or more subjective statements found in the humanities (RICHERSON; BOYD, 2004). The study of cooperative behavior is overlapped by several classic fields of study, such as Economics, Psychology and Social Sciences. Additionally, in recent years, new denominations were created trying to better describe the ongoing interdisciplinary work being made, for instance *Social Computing*, *Human Computing* and *Social Dynamics*. There is no fair agreement to what properties, phenomena or objects receive special emphasis in each of the above mentioned areas because, as we will see in the next chapters, theories and models about cooperative work are still under development or are subjective and comparisons are not straightforward. However, we can more easily distinguish each area with the above definition of cooperative behavior if we add context and methodology. For instance, the game theoretic approach[1], although shared with other fields, is more commonly found in Economics (KREPS, 1990) and makes use of complex mathematical models (BREDIN; PARKES, 2012). We are going to use this method, of highlighting important traits about each field, to tell apart each different approach when necessary.

Nonetheless, as we present the results from the first phase of our experiments (to be described in section 4), we gathered evidence to show that people will decide to cooperate with any source of information they can connect to, and not only humans (even machines) as they were communicating with humans, thus making our model less specific to networks of individuals cooperating but more generals terms of nodes of information exchanging information. Put another way, if machines or information systems are not considered cooperative peers, our results applies to humans working in non-social, disconnected environments. Therefore, we can understand how people behave at the group level by using characteristics that are present even when individuals are not in a group. With this new hypothesis in hand, we perform a second phase of experiments to collect empirical data to prove our hypothesis where individuals are asked to participate in a controlled public policy poll and, after that, a non-controlled

[1] Also found as *game-theoretic* approach.

experiment in a labor market, the Amazon Mechanical Turk.

In this work, we have investigated human behavior in online environments from our own model standpoint, the *Artificial Memetic Networks* (AMN). AMN were originally inspired by the naturally occurring spread of ideas (DAWKINS, 1990) and was successfully applied on computer simulations to investigate the problem-solving performance of networks of artificial agents (ARAUJO; LAMB, 2008)(ARAUJO, 2010). The present work modifies AMN and use it to design, perform and analyze experiments with individuals in problem-solving online networks scenarios. Our model provides enough mathematical rigor to allow us to model experiments and tell apart each object under investigation with precision, while keeping certain generality of the assertions to allow comparison between other models and theories. This goes along with our goal of making possible a better information system design where people perform activities online.

The AMN model was first introduced in (ARAUJO, 2010)(ARAUJO; LAMB, 2008) and can be used, as we will explain in the coming chapters, to understand human behavior in both connected and disconnected environments by modeling how memes (or ideas) spread and mutate as they are exchanged between individuals. As we developed the concepts of AMN, we will also perform hundreds of experiments to provide empirical evidence that our model can correctly pinpoint behavioral traits that can explain observed phenomena.

We show that when solving tasks with very small duration (under 5 minutes), also known as microtasks, individuals will act less rationally as expected in classical game theoretic predictions, but strongly biased by costs of searching (and evaluating) options rather than financial or non-financial incentives. Therefore, the search and evaluation process, modeled more commonly in algorithms for combinatorial optimization, plays an important role in human decision making and constrains on the options available to humans are required for a realistic modeling of human decision making in social network. A brief introduction to human decision making can be found in section 2, where we present an introduction to literature that sets the context for applications and experiments that mark the beginning of highly available, massive data and low cost social experimentation

The constraints to solutions search and evaluation when solving problems, as suggested in our study, limit the number of available options in microtasks solving. Moreover, influence from one individual to another, as show in this work, when solving problems, can be created by rearranging elements visually to modify search sequence order of an individual, be it by designing the layout of the virtual work environment or by manipulating which options are first shown in non-controlled environments such as the Amazon Mechanical Turk labor market.

Individuals, therefore, rely more on the search conditions then on exogenous incentives (MANSKI, 1993). Our evidence is based on several experiments in online social networks we performed where individuals work on tasks with varying degrees of difficulty within three settings: mathematical games with objective truth (Sudoku and SAT instances), surveys with subjective evaluation (public policy polling) and labor markets (Amazon Mechanical Turk).

A description of the model we used and the experiments are detailed in section 4. We provide empirical evidence experimenting with thousands of subjects, from controlled experiments to naturally occurring phenomena observed in a labor market and in a social network (Facebook). Section 5 describes the results and the data analysis that support our conclusions. Section 6 presents a discussion, conclusions and future work.

2 AN INTRODUCTION TO THE LITERATURE

The performance of human cooperative work has been studied extensively in many contexts (MALONE; CROWSTON, 1994). However, research by means of the methods of computer science did not come before the dawn of successful online applications whose value is based on the work of millions of users coordinated by electronic infrastructure (BRAN et al., 2009).

In section 2.1, we will show examples of applications that were able to prove the power that human cooperative work is able to achieve in online environments in an era where Artificial Intelligence with Machine Learning was believed to solve any tractable problem with powerful hardware alone. By showing not only the success of applications but also that they were based on incomplete or unverified theories, we will argue that a stronger, more reliable theory is needed, one that can be backed by experimental data and have properties that can be applied generally to applications that rely online human work.

In section 2.2 we will present examples of candidate theories constructed to close the gap between the unforeseen applications and existing, classic theories. Although they may provide clues about how people behave, they oversee several naturally occurring phenomena in experiments or cannot be applied to the design of experiments, as they are still under development and assume properties that are not realistic to what has been observed so far.

In the final section, we will present studies that show experiments specifically designed to provide evidence for new theories on human online collaborative work.

2.1 Examples of Internet Applications

2.1.1 reCAPTCHA

A lush of computer science papers have been published in the past years that presents faster ways of processing images by making use of Graphic Processing Units (GPU) which, with the impressive economic growth of the electronic game industry, have developed from simple processing cores with special Single Instruction, Multiple Data (SIMD) instructions to hundreds of specialized, stream processors integrated into System-on-Chips (SoC). The increase in image processing power hinted at an era where image segmentation and object recognition could be performed on a huge scale. However, it was found that the quality of the extracted information from images could not be paralleled by a new way of automatically recognizing images: a network of motivated individuals tagging images, known as reCAPTCHA(AHN et al., 2008).

A CAPTCHA (*Completely Automated Public Turing test to tell Computers and Humans*

Apart) is a method to tell humans and machines apart by using Hard AI problems (AHN MANUEL BLUM; LANGFORD, 2003):

> A captcha is a program that can generate and grade tests that: (A) most humans can pass, but (B) current computer programs can't pass.

A known successful application of captcha is the reCAPTCHA (AHN et al., 2008), which is more commonly known as the wobbly word pairs that one has to identify as a requirement to requesting information on a website, such as survey forms or e-mail access when providing credentials (see figure 2.1).

Figure 2.1 – The reCAPTCHA as an example of a CAPTCHA

In reCAPTCHA, an Internet user is required to type two words before submitting his or her credentials to a web form, aiming to identify if the user is a machine or not. One of the images of a word was generated by a computer and thus the system knows what word is contained in the image. The other word was captured from a real situation, e.g. a scanned book or a photo, thus not generated artificially by a computer. If the user types both words and the generated image does not match the word used to generate the image, the system will assume the user is a machine and his interaction with the web application ignored. Therefore, the user, when typing both words, does not know which one was generated and thus is required to type both words correctly to pass the test. Therefore, by correctly typing both words, the user is not only providing evidence that he or she is a human, but also performing character recognition for the non-artificially generated image.

With reCAPTCHA, words in images are correctly identified as users are required to match the image with the correct words, otherwise they cannot continue interacting with the web site the way they wish. Therefore, there is a motivational component (users want to interact with a website), usually explained in the Game Theory context of *Mechanism Design* or in Algorithmic Game Theory (AGT), and the outcome of the interaction produces Optical Character Recognition (OCR) work with higher quality then previously available from any other OCR algorithm performed by machines. Therefore, with a CAPTCHA, which is a generalization of reCAPTCHA (AHN et al., 2008):

> [...] either the problems remain unsolved and there is a way to differentiate

> humans from computers, or the problems are solved and there is a way to communicate covertly on some channels.

By far the most successful example of Captcha, the reCAPTCHA was acquired by *Google* for an estimated $400 million dollars, proving quality OCR services for Google goals of extracting as much information from the web as possible and for digitizing books in the *Google Books* project.

The theory that explains how and why reCAPTCHA works comes from a paper from the reCAPTCHA authors (AHN MANUEL BLUM; LANGFORD, 2003). However, this work was published 10 years after the original patent for OCR with CAPTCHAs (see (LILLIBRIDGE et al., 2001)) and it does not provide a clear way on how to design new, similar applications. It is not surprising at all that applications may precede theory and this pattern will repeatedly happen with online applications and today we are yet to have a reliable theory that can be used to actually design a Captcha-like application (considered a reference application of networks). What is demonstrated in (AHN MANUEL BLUM; LANGFORD, 2003; AHN; DABBISH, 2008) is that, assuming each user has to match two words (see figure 2.1), it is possible to compute the probability that the information provided is true, assuming an acceptable error. Therefore, it does not model why people can perform better than machines when performing OCR, and it that case we cannot predict the performance of a newly designed system.

2.1.2 Wikipedia

Wikipedia is a cooperative effort to build the most up-to-date, well-reputed encyclopedias. Currently, Wikipedia is the 7th most accessed web site in the world (WikipediaRanking, 2016) and is able to be as reliable other non-internet encyclopedias (GILES, 2005).

To create Wikipedia, generally any user can edit articles and contribute to their content and organization. Each article, together with their respective *talk pages* where users can discuss and share information, is constructed with parts of information each user can contain. Moreover, users can edit each other contributions, making corrections or adding to existing information on each article. Therefore, Wikipedia is an online application of networked individuals sharing pieces of information and aggregating them into articles. As a reward, volunteers are essentially co-authors of any article they contribute to. Another incentive is to be able to altruistically contribute to the advance of knowledge and have other share their knowledge. It is important to observe at this point that there are no monetary incentives in Wikipedia or reCAPTCHA.

2.1.3 GalaxyZoo

Beyond interstellar dust, the project GalaxyZoo (GalaxyZoo, 2015) proposes the classification of galaxy morphology from images taken from satellites. In this case, from an expected crowd of 20,000 to 30,000 participants, the project quickly (and unexpectedly) gathered 100,000 active participants (LINTOTT et al., 2011) (see figure 2.2).

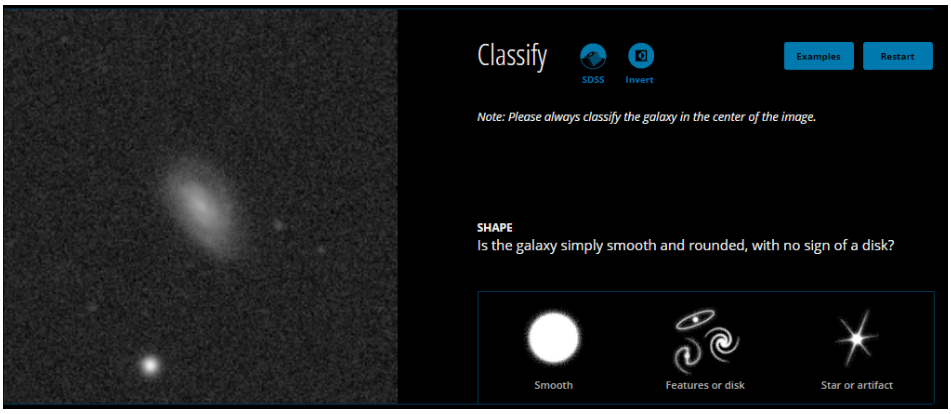


Figure 2.2 – GalaxyZoo interface

Volunteers are requested to classify the object at the center of the image. As the object is classified between classes *Smooth*, *Features or disk* and *Star or artifact*, more questions might be asked to the user.

We have been emphasizing the interconnected nature of some online applications aiming to point that, on the Stardust@Home project or on the GalaxyZoo project, both have a central server that has data from physics experiments. The data is distributed to volunteers and the output of the work is sent back to the server. The output is thus aggregated and stored in the central server. One could argue that the volunteers are individually working and do not cooperate as they are not communicating directly between them, however no work on such scale can be performed without a (machine) central point of interaction (in this case, an Internet server). This is very important to consider when comparing online applications to other networks studied in Economics and Social Sciences, as previous works did not consider situations where work sharing would have such low cost and where tasks could be performed so quickly. This could be argued as a difference between predicted results from classic fields and what has been observed in online behavior.

Additionally, we observe that no individual task performed by any volunteer described so far took more than a few seconds to perform. As we will see in some experiments, this type of work is sometimes classified as a *microtask* and have interesting, reproducible properties on their own that will be useful when analyzing online behavior in the next chapters.

2.1.4 FoldIt

So far, we discussed online applications that provide high quality output work when matching individuals work (reCAPTCHA), applications that predefined work is distributed from a server (Stardust@Home, GalaxyZoo) and now we are going to show an example of an application where individuals have to create solution to problems for which no solution has been created so far, thus highlighting a creative component of the system.

In the *FoldIt* project (FoldIt, 2015), participants are given 3D molecule structures and are asked to find geometric configurations that will maximize (or minimize) the energy potential of the molecule. Finding the minimum energy of a molecule and the corresponding geometry is a central problem in Bioinformatics and Protein Folding (PF) (ANDRADES et al., 2013):

> [Computational Protein Design (CPD)] design can be considered as the inverse of the protein folding (PF) problem (OSGUTHORPE, 2000) because it starts with the structure rather than the sequence and looks for all sequences that will fold into such 3-D structure. Considering that there are 20 naturally occurring amino acids for each position, the combinatorial complexity of the problem amounts to 20^{110} or 10^{130} (FLOUDAS et al., 2006).

Therefore, PF is an intractable problem and the use of heuristics is required. By challenging volunteers to fold a protein to the lowest energy state, FoldIt aims at leveraging the intelligence advantage over machines that humans have and for long has been sought by AI researchers. It has never been proven that humans are more or less capable of processing data than computers, however no robot or machine can match human cognitive skills in absolutely all types of activities, and thus it is still believed that humans have special cognitive skills. In this sense, the project FoldIt tries to harness this human cognitive potential by allowing individuals to work on proteins through a graphical interface (figure 2.3).

It has been shown that individuals are indeed capable of finding interesting structures, thus validating the hypothesis that humans can provide solutions machines are not able so far. The evidence for this is a series of solutions for folding geometries provided by humans that were not found previously by any system that searches for solutions only using machines (FoldIt, 2015).

2.2 Relevant fields and methods

In this section, we will show examples of studies in classical fields (Algorithmic Game Theory and Evolutionary Biology), new fields (Networks Science and Behavioral Economics)

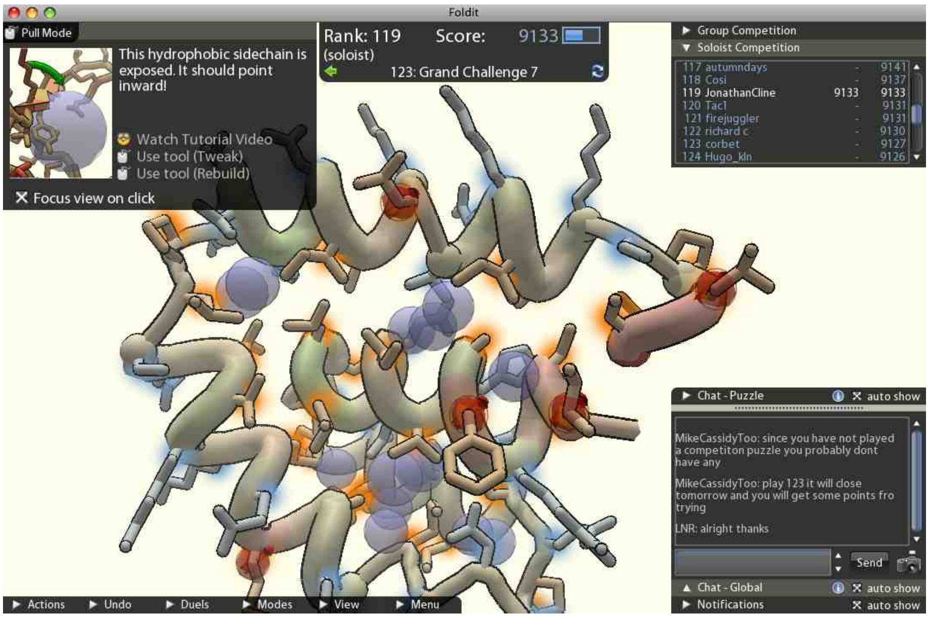


Figure 2.3 – FoldIt Project: volunteers manipulate molecules to find the lowest energy state possible of the molecule.

In the FoldIt project, volunteers manipulate molecules to find the lowest energy state possible of the molecule. Although easy to compute the energy of the geometric configuration, finding the lowest energy is an intractable problem.

and experiments that were performed with the current successful applications of social networks in mind (KEARNS; SURI; MONTFORT, 2006).

The mentioned fields do overlap and we will aim less at addressing broadly how each field tries to explain human behavior but more at highlighting specific emphasis each field gives to specific methods.

2.2.1 Algorithmic Game Theory

Rigorous mathematical treatment of group behavior is originally found in Economic Theory and later in Algorithmic Game Theory, and extensive references are available to introduce the topic. We start by quickly exposing the basic rational agent model that is generally used in academia and then showing how it is being applied in the setting of online social applications.

A game theoretic analysis generally starts from the following model: there exists an entity *agent* that is deemed as *rational* for being capable of, given a set of actions, choosing the action that is the *best*, i. e. one that can increase a given *fitness* quality of the agent. The actions available to the agent and the change in the agent fitness according to his or her actions is given by a *payoff function*. The sequence of actions an agent takes is her strategy. If all agents know the history of actions of all other agents, we can say this is a perfect information game.

Let us practice the definitions above applied to the Facebook social networking web site. The agent (human individual, or user) can take actions (e. g. post text and images, comment, like, scroll the page, leave the website, turn off the computer) that is expected to improve his or her fitness. A payoff function maps the current user context (e. g. set of peers, time of the day, native language, etc) and set of actions to a change in the agent fitness. Moreover, a sequence of actions can be grouped into a strategy of an agent e.g. first post a message, then comment, then turn off the computer, then turn on the computer, then post again.

In the example above, we can easily trace the analogy between agents and users and between posts and actions. However, we cannot easily determine the payoff function of users. Indeed, there is no generally accepted methodology to model a user payoff function and each field has its own ways with pros and cons, and this is a motivation to research new models.

In Algorithmic Game Theory, we want to provide assumptions about the payoff functions of users and restrict their actions so we can describe the system as a mathematical model that can have a known analytic solution in game theory, also known as a *Solution Concept*. Having a model without an analytic solution is possible but avoided as it can introduce more parameters and weaken the model.

We will take from (GHOSH; HUMMEL, 2014) as a reference for social behavior modeling in a Question and Answer (QA) forum. In a QA forum, each agent i can choose among an infinite, continuous set of actions:

- to not contribute with an answer;
- to contribute with an answer with quality q_i;

On the last type of action, the agent also can choose the quality of the contribution, from 0 to infinity, then making the number of possible actions infinite as well. We will base part of the above model in our explanation of results combined with a fitted function to our experimental model. Moreover, work done in (JAIN; PARKES, 2013) will also be used to show our point that AGT does not yet cover enough cases found in our experiments to have them used in as a tool for predicting or fully analyzing online behavior.

A recent approach to study human behavior by means of the network structure can be found on (DAVID; JON, 2010) and became known after identifying the naturally occurring scale-free network (also used in our experiments) (BARABáSI, 2002). The cited work should be used, in this dissertation, to describe what kind of effect should have been detected in our experiments.

2.2.2 Solving the coloring problem with a network of human subjects

One experiment that has been used extensively as a reference to design the first phase of our experiments is (KEARNS; SURI; MONTFORT, 2006). In this work, the author proposes an experiment where individuals are asked to solve collaboratively the Coloring-Problem. Each individual is given a color and is positioned inside a virtual social network implemented in a lab experiment. Other nodes in the network are individuals as well and the action allowed for the individual is to change their own color aiming that, as neighbors in the network see each other's colors changing, to solve the classic Graph Coloring Problem.

The paper is a case where financial incentives are provided. Reports written by the participants indicate that, while only being allowed to communicate by changing colors in a timely manner, individuals established a signalling protocol. However, the methodology for this analysis was not described, as well as the motivation for the way behavioral traits were analyzed.

The experiments were performed at the university lab with each individual in a single computer terminal.

2.2.3 Theoretical predictions against empirical evidence

Around 2011, Duncan Watts and Siddhartha Suri published a paper where they perform experiments in a social network and (confirming experimental results we found and that we will introduce later in section 5.1), they found that network topology did not affect their outcome (SURI; WATTS, 2011):

> Surprisingly, we did not find significant differences in the average contributions in each round for the various topologies. Moreover, we did not find large differences in the distribution of individual contributions over each round, or in the average contributions of groups defined by the network topologies over each round. Thus we conclude that topology does not exert a noticeable impact on contributions at any level: individual, group, or aggregate.

Meanwhile, they were able to expose the balance of the contributions between theoretical work and what has been found experimentally, mainly pointing to the fact that only near end game situations is where individuals can more closely relate to theoretical predictions.

They have performed around 50 experiments, around the same number of experiments we did within a lab environment (opposing later the open, dynamic nature of the Amazon Mechanical Turk experiments we performed).

3 ARTIFICIAL MEMETIC NETWORKS

We will first introduce a more formal concept of culture so we can then talk about memes. There are a number of models for culture, some of which wish to present a mathematical definition of culture. An interesting survey may be found in Birukou et al. (2013) and from there we borrow and adapt a formal definition of culture.

3.1 Model

We denote each agent as a_i and the m-th group of agents as $A_m = \{a_1, a_2, ...\}$. When no index is specified, A represents all agents that may exist in our context (Internet, social network, etc.).

Each agent a_i possesses *cultural traits*, which are (BIRUKOU et al., 2013)

> characteristics of human societies that are potentially transmitted by non-genetic means and can be owned by an agent

However with different names, trait is an element that can be found in most theories about humans, societies and cultures. For example, traits may be found in other works in the form of *ideas*, *memes* or even as *behaviors*, such "The Earth is round", a description on how to build a computer or even just a number. All such instances of traits are based in informal, subjective definitions and often too broad concepts, making it difficult to compare and construct a more universally accepted description.

Nevertheless, existing definitions of traits share a common characteristic: **traits can be transmitted and learned by individuals** (BIRUKOU et al., 2013). We can use this common characteristic to define traits.

Definition 3.1.1. *Agent a_i possesses zero or more (cultural) traits τ_j and the set of traits of a_i is T_{a_i}, itself contained in the set of all traits $T_{a_i} \subseteq T$. Traits can be shared and learned by agents. Agents can share traits that have never been learned by them, i. e. agents can create new traits and share them*[1]*. Thus, we also present the function:*

$$T_{a_i} : t \to T_{a_i,t} \tag{3.1}$$

to map trait sets at different instant of time as they change due to learning or trait creation.

[1] In our model, the creation of new traits (or memes) is the result of a mutation processes (ARAUJO; LAMB, 2008).

In our model, T is a totally ordered set and elements can be indexed uniquely with j and $j \in \mathbb{N}_+$. This is a reasonable assumption because traits can be described uniquely with natural language descriptions (such as "The Earth is round.") that can be, at least, ordered alphabetically (see 3.1.3 for a deeper discussion).

Finally, we will start with the following informal definition of culture:

> The culture of any society consists of the sum total of ideas, conditioned emotional responses, and patterns of habitual behavior which the members of that society have acquired through instruction or imitation and which they share to a greater or less degree (LINTON, 1936)

Culture is the set of traits of all agents in a group. Thus, the culture of group $A_m = \{a_1, a_2, ...\}$ is $C_{A_m} = \{T_{a_1} \cup T_{a_2} \cup ...\}$.

Definition 3.1.2. (Definition of Culture) *The culture C_{A_m} of the m-th group of agents A_m is*

$$C_{A_m} = \bigcup_{i=1}^{|A_m|} T_{a_i}. \tag{3.2}$$

If group m has only one agent a_1 then the culture of the group is the agent's traits $C_{A_m} = C_{a_1} = T_1$ and we can talk about an *agent's culture*.

An active field of study is how cultures evolve and, to model the culture's capability to evolve, we will define a culture in different points of time t. For this, τ_i contains a timestamp that informs when agent a_i acquired t_i and C is equipped with a total order.

$$C_{A_m} : t \to C_{A_m,t} \tag{3.3}$$

For instance, if one wish to know the culture at times t_0 and t_1, one computes $C(t_0)$ and $C(t_1)$, respectively.

Definition 3.1.3. (Evolution of Culture) *The evolution (or difference) of culture of group A_m from time t' to t'' is*

$$\Delta_{t',t''} C_{A_m} = C_{A_m}(t') \cap C_{A_m}(t''). \tag{3.4}$$

Generally, we may be interested in the evolution of culture between the beginning and the end of an activity of agents to solve problem instance p_k. If t' and t'' are the time instants when problem instance p_k begins and ends, respectively,

$$\Delta C_{A_m,p_k}(t) = \begin{cases} \emptyset & \text{if } t \leq t' \\ \Delta_{t',t'+t} C_{A_m} & \text{if } t' < t < t'' \\ \Delta_{t',t''} C_{A_m} & \text{if } t \geq t'' \end{cases} \tag{3.5}$$

In our work, traits are mainly limited to constitute problem's solutions, e. g. each trait is a SAT solution or a Sudoku solution. Therefore, agents in our social network may have only one trait type [2].

3.1.1 Capability to forget

Differently from (BIRUKOU et al., 2013), in our model agents may forget traits, thus we will not take into account for our model definitions and axioms related to causality and sharing of traits presented in our reference work. The possibility of an agent forgetting a trait is necessary to justify why sometimes agents try to solve problems with solutions they already tried and failed in the past, a situation that is not possible in the (BIRUKOU et al., 2013) model. Still, we use an adapted form of *snapshots* (BIRUKOU et al., 2013) when we presented eq. 3.1 and 3.3, allowing the causality definitions to be included back as long as proper notation is presented.

3.1.2 A Game Theoretic Description

Here we present a Game Theoretic description of meme evolution that will be useful to discuss gains and costs on our experiments. Here we will try to provide a description of meme evolution based on basic Game Theory concepts.

Each agent a_i strategically chooses an action (also traditionally known as a *move*) $\alpha_i(t) \in \{E, I, N\}$, where $\alpha_i = E$ indicates that agent i chooses to act extrovertedly by copying another agent's trait. In this case, we may also specify from which agent a' the trait was copied by with $\alpha_i = E(a')$ notation. Additionally, $\alpha_i = I$ indicates that agent i act introvertedly by adapting his set of traits T_i into a new trait(ARAUJO; LAMB, 2008). When necessary, we may specify the new trait created τ' as $\alpha_i = I = \tau'$. Finally, $\alpha_i = N$ indicates that agent i does not act.

If we assume agents are solving a Sudoku game, for instance, then the game instances being played are finitely repeated games and where necessary we will specify the time each action was

[2]We performed experiments where agents were able to communicate with natural language, and not only problem's solutions. However, those cases were excluded from our analysis in this paper as the amount of data needed to was not sufficient. Additional experiments are planned.

taken with the t parameter. All first actions of each agent happen in $t = t_0$ in 3.5. However, not showing the time instant does not imply that time steps have the same duration between agent moves, although we can always add as many $\alpha_i = N$ to make all agents actions synchronized with constant time steps.

The strategy of agent i is denoted as $s_i = \{\alpha_i(t_0), \alpha_i(t_1), ..., \alpha_i(t_{p_k})\}$ or s_{i,p_k} when we specifically refer to the strategy taken by agent i while solving the k-nth problem instance p_k. The strategy profiles are denoted as $S_{p_k} = \{s_0, s_1, ...\}$ for the k-th problem instance p_k. Since our game is finite (discrete steps and time bound), we define a new $s_i(t'') = s_i(t') \cup \{\alpha_i(t')\}$ when each agent take an action individually. Each next step is indexed sequentially, e. g. $t_1, t_2, ..., t_{p_k}$ and t_{p_k} is the duration of problem p_k.

The N action is important to transform the asynchronous actions of agent a_i into actions that happen at the same time, by adding a $\alpha_i = N$ action to s_i when a_i did not played while A_{-i} played.

We are already able to formalize specific statements about agent interaction in our experiments. For instance, an agent a_3 copying another agent's a_5 solution τ_7 at time $t_1 1$ is formalized as $\alpha_3(t_{11}) = E(a_5, t_{11})$.

3.1.3 Continuity of Traits (or Ideas)

We declared previously that traits can always be described with a sequence of characters in natural language. Even if phrases or words can be ambiguous, we can always add more description to the trait to resolve the ambiguity. For instance, the Wikipedia article about "Neighborhood" define this word as part of a city. As a consequence, whenever we want the article about the mathematical notion of "Neighborhood" (from Set Theory), we add the context to the Wikipedia article's title in parentheses ("Neighborhood (mathematics)"). In this way, we can uniquely address every trait, or idea, and set a relation such as the alphabetical order.

We define the metric space (T, l) where l is the Levenshtein Distance (LD) between the string description of two traits l : description$(T) \times$ description$(T) \to \mathbb{N}$. This metric is going to be used extensively to represent graphically T.

For any experiment or observation, we need to force agents to change traits in a way that there are discrete, unitary changes from the past solution. Enforcing this behavior, we guarantee that we trace the agent's introspective line of though since it represents the agent's strategy while not sharing his or her solution with other agents. We assume that an agent cannot hold a long number of traits in her memory and thus need to use the interface to keep track of the current solution.

Lemma 3.1.1. (Continuity of Traits in a Culture) *For all agents $a_i \in A$, any sequence of actions $\alpha_i \in \{I, N\}$ forms a topological space with neighborhood*

$$B(\tau') = \{\tau'' \in T | l(\tau', \tau'') \leq 1\}. \tag{3.6}$$

Proof. Users are forced by the UI to take actions that modify only one variable of a SAT solution or only one Sudoku tile each time they take actions I. Also, the encoding of the SAT solution is an array of variables encoded as their value with the 'T' character for *true* values and 'F' for *false* on each variable of the problem, and for the Sudoku as the concatenation of all numbers on the board, replacing the blank space with '0' (zero). With this encoding, each time an agent changes one variable (or one tile in the Sudoku example), the Levenshtein Distance (LD) is at most 1. Finally, the action $\alpha_i = N$ have Levenshtein Distance zero and abide Lemma 3.1.1. $\square$

The continuity of agents introspective behavior showed in Lemma 3.6 will also be used to identify whenever an agent acted extroverted by copying from other agent, i. e. when there is any break on the continuity established in Lemma 3.1.1. We can identify this clearly from our experiments, but we should not assume this information will be available all the times, as non-controlled observations cannot be certain if agents were copying or thinking introspectively.

3.1.4 Visualizing Cultural Evolution

Representing a dataset of online work, constructed from collecting data from observations of human behavior, can be very subjective. If we divide behavior into two classes, for instance *cooperative* and *non-cooperative*, we have to group the data points from our dataset into those two classes. Initially, when first observing humans cooperating, we do this classification by visual inspection, picking each observed agent and adding to one of the two groups. However, when choosing the a specific activity (e.g. playing a game, doing a business or negotiating price), it becomes clear that there is no objective way of telling apart cooperating and non-cooperating agents.

How can one say that a specific individual has been cooperating or acting selfishly in the long term? From Game Theory and Animal Sociobiology it is well accepted that agents are always acting selfishly, but can, in the short term, show cooperative behavior, i.e. accept temporarily a situation of less gain so in the future an overall gain is justified.

To classify an organism to be cooperating or not is straightforward in Game Theoretic experiments with well defined rules, or within life and death situations that living beings might face. In the case of Artificial Memetic Networks, at least simulated experiments, we will provide a similar, objective way of telling whether the agent is cooperating or not. However, when looking at social experiments, we are biased by our opinions and beliefs about who is cooperating or not.

Aiming to find in Artificial Memetic Networksa compromise between definitions provided by evolutionary theory and the well developed mathematical theory behind Game Theory, we need a visual aid to not only better explain Artificial Memetic Networks, but also to receive hints about the how can we can chose a classification method. In this sense, it is important to first provide a way for researchers to intuitively agree with our proposed model, in the sense that one should be able to recover, under the layers of mathematical models and data analysis, what we usually accept *cooperation* to be. Our goal in this section is, therefore, to provide a way of expressing behavioral patterns in a visual, intuitive way.

We will now take the proposed model and offer a visual representation for each of the described elements, mainly memes and meme transitions, as a intuitive way of understanding the model.

3.1.5 Timeline of the Actions of Agents

In this section we model agents solving a problem as an optimization problem and relate the solving process with the cultural dynamics of a group of agents. Through action $\alpha_i(t') = I$, agent a_i generates a trait τ' based on other traits in T_i (see section 3.1.2)(ARAUJO; LAMB, 2008). We name this process *generation* for two reasons.

First, when acting $\alpha'_i = I$ all information needed by the agent to present the trait τ' is endogenous to the agent. Thus, no information was used from external resources, such as other peers A_{-i} in the social network. For an external observer, trait τ' was *created* or *generated*, since the output τ'_i came from agent a_i with on direct interaction with the environment. Secondly, if trait τ'' is subsequently generated at time t'' by $\alpha_i(t'') = \alpha_i(t') = \tau' \text{and} t'' = t'+1$, we expect the cardinality of T_i to remain unchanged as $|T_i(t')| = |T_i(t'')|$, since the set of traits is unchanged. However, because agents can forget (see 3.1.1), $T_i(t'')$ may be smaller than $T_i(t')$ with the loss of one or more traits. Nevertheless, in our work $|T_i(t)|$ is always monotonic and agents' capacity to forget will be mentioned only when an agent takes action $\alpha_i(t'') = \tau'', \tau'' \in T_i(t')$ and $t' < t''$, which means traits (also solutions to problems in our experiments) are being repeated by the same agent, even thou they do not represent the final solution to the problem being solved.

We name the generative function as follows:

Definition 3.1.4. *Agent a_i can generate traits with G_{a_i}:*

$$G_{a_i,\alpha_i,t} : (T_{a_i}, T_{a_{-1}}) \to \tau_t \tag{3.7}$$

and

$$\tau_t : \begin{cases} \tau_t \in T_{a_i} & \text{\textit{if} } \alpha_t = N \\ \tau_t \in T_{a_{-i}} & \text{\textit{if} } \alpha_t = E \\ \tau_t \in T & \text{\textit{if} } \alpha_t = I \end{cases} \tag{3.8}$$

Therefore, generating traits can produce a new trait not previously in T_{a_i} and $T_{a_{-i}}$.

The process of generating a trait is at the core of the gain (and cost) function. It is in this function that actions takes place and the agent has to decide whether to copy, to mutate a trait or to do nothing. Indeed, the decision process will be shown to have important costs if an individual decides to act extrovertedly e.g. $\alpha_i = E$ that we are able to modify an individual behavior when they are performing microtasks.

We may start now connecting our model to (ARAUJO; LAMB, 2008). From (ARAUJO; LAMB, 2008):

Definition 3.1.5. Aggregation by Copying the Best Neighbor*: Let A be the set of adjacent vertices to any vertex; let u = argmax$_{x \in A}$eval(x). If more than one vertex in A may satisfy this condition, u is chosen randomly among these. Then, make $v \Leftarrow u$, otherwise v is left unaltered.*

To translate 3.1.5 to our mathematical framework, we have to make $u = T_{a_{-i}}$ for connected peers, i. e. the agents that are connected and visible to a_i. Also, $x : \tau$.

Definition 3.1.5 states that agents act rationally by copying the *best* solution according to a evaluation *eval*, just like a game theoretic model. Naturally, both Artificial Memetic Networksand Game Theory extended in many ways this assumption.

There are more steps involved processing Artificial Memetic Networks. However, for the sake of our experiments, it is enough to introduce the above definition, as others steps, like the *Connection Step*, *Appropriation Step* and the remaining *Aggregation Step* models are out of the scope of this work.

If $\alpha_i(t') = I$ and agent's trait set $|T_i(t')| = |T_i(t'+1)|$, $\forall t' < t'', |s_i(t')| < |s_i(t'')|$ still holds and $|s_i(t)|$ is monotonically increasing and $s_i(t') = \alpha_i(t')$. It may be useful to present a graphical representation about how agents interact and how their traits and culture change as they take actions.

We present the diagram in Figure 3.1 to describe the timeline of agents taking actions during a game. Although we can describe the actions of agents with natural language, when we deal with multiple agents and hundreds and possibly thousands of actions in a single experiment run, the agent activity representation became cluttered and we should simplify.

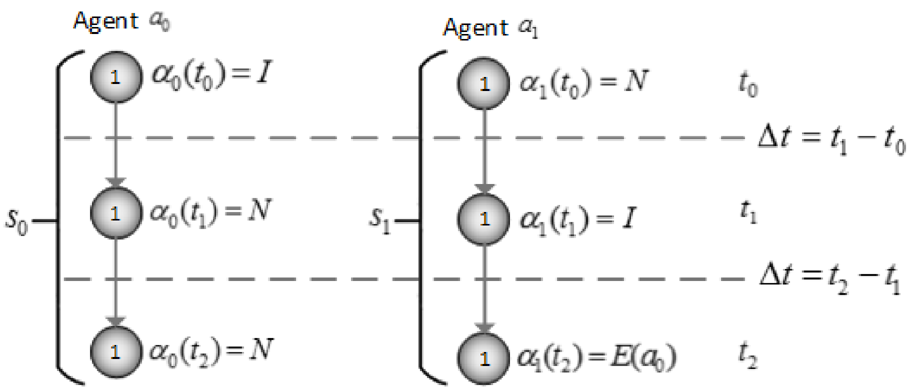


Figure 3.1 – Visual model for meme evolution

The diagram shows two columns, each representing agent a_0 and a_1 strategies s_0 and s_1 respectively. When agent each takes any action individually, time changes by Δt and $s_{-1} = N$ and a new last trait is selected. In the example, first agent a_0 takes action $\alpha_0 = I$, which means it does not copy from agent a_1 but generates trait τ_0 (indicated by the edge τ_0) based in his own set of traits T_1. Because agents most probably do not take actions simultaneously, we set $\alpha_1 = N$ to keep the game synchronous. During the last step t_1, agent a_1 copy from agent a_0. Only in action $\alpha_0(t_0)$ a trait τ_0 is generated. The number 1 inside each circle will be explained in Figure 3.2.

Using Figure 3.1 as a reference, we present a simplified diagram in Figure 3.2 that show graphical features in social activities that inspired our further analysis of agents interaction with cultural dynamics.

To simplify, agents identity is of no interest for culture dynamic analysis, so we drop labels (in Figure 3.1 we drop a_0 and a_1). It is implied that any path in this graph represents a strategy of one or more agents, so strategic information is not lost and can still be used to approximate payoff functions.

Even if agents perform actions at different times, there is no individual timeline representation in 3.2 and this will not be necessary to present our results. Additionally, the type of problem instances we used in our experiments make agents quickly take actions and displaying agents actions with time layers (as in Figure 3.1) does not change significantly the representation. The

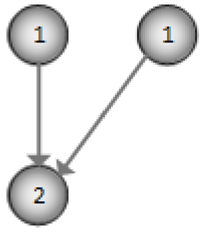


Figure 3.2 – Simplified visual model for meme evolution

Simplified visual model for meme evolution: the present diagram simplifies Figure 3.1 to introduce Culture Dynamics concepts on the light of Game Theoretic models. We drop the labels a_0 and a_1, exclude $\alpha = N$ actions and set node value as the number of agents that ever output t' Differentiation between $\alpha = I$ and $\alpha = E$ is guaranteed by continuity breaks (see Lemma 3.1.1). The diagram layout must be arranged with the *neato* algorithm so we can visually see continuity breaks.

timing information will be useful for the next step in our work where we study homophily with the aid of the traits interpretation.

While generating unique traits, there is a growth of culture C_{a_i} and therefore culture C_{p_k}. Each agent a_i selects a strategy s_i aiming to maximize it's utility function $u_i : S \to \mathbb{R}$ where S is the space of all possible actions an agent can choose from.

A first visualization can be found in Figure 3.3 and a typical interaction diagram is displayed in Figure 3.4.

Important features, such exchanging memes, strategy and the evolution of memes, are being represented and used to intuitively explain phenomena such as meme exchange in the context of cooperation and homophily. Our goal with the representation is not to directly show or verify cooperative behavior but to provide a grasp on the concepts of meme evolution and exchange that were presented in the previous sections.

We constructed a mathematical model for RMA that addresses important questions that arise in the context of social experiments. The model was also used to generate a visual representation. In the next section, we will apply and verify our model to online experiments.

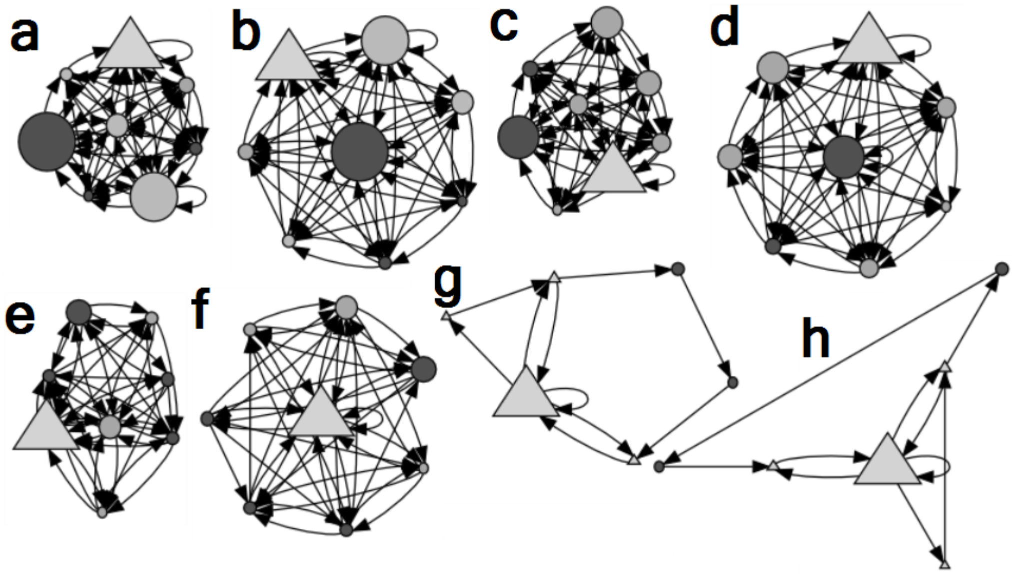


Figure 3.3 – Problem Instances with a low number of states cannot be visually represented with our visual model

Problem Instances with a low number of states cannot be visually represented with our visual model: the low number of states for the SAT instances decreased the value of the visualization. After generating this visualization, we planned working with SAT problem instances with a higher number of variables. Nonetheless, over 65 variables was already too much for human agents to solve any problem instance.

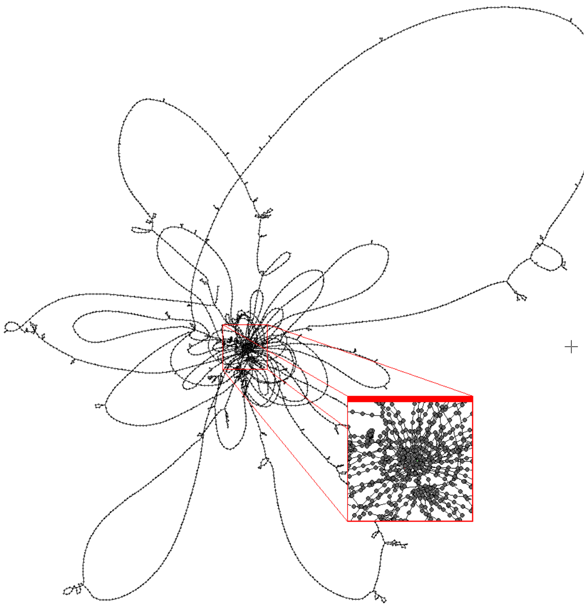


Figure 3.4 – Meme Evolution paths can be clearly seen in this visualization, but not on areas with popular memes.

Memetic evolution paths can be clearly seen in this visualization, but not on areas with popular memes: after finding that visualizing meme evolution on a problem instance with a low number of states was not possible, we devised a new set of experiments with Sudoku problem instances. The visualization of a single instance of a Sudoku problem can be seen above. Now, it is clear that there are agents that take isolated paths from other agents, while most of the agents try to copy each other and have similar ideas. However, because of the high number of states and the popularity of some memes (solutions), we required to zoom in. The next figures show a zoomed version of the central area.

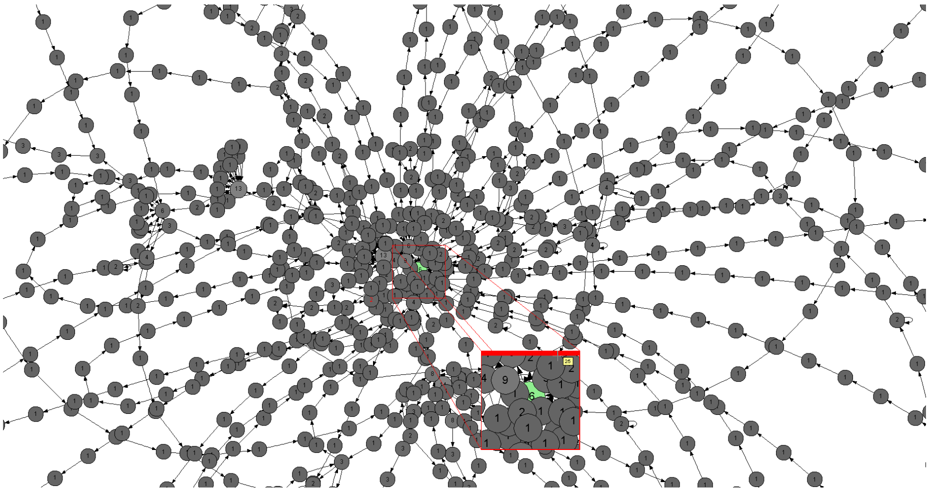


Figure 3.5 – Meme Evolution visualization with zoom 2x can already show when agents cooperate by exchanging memes (problem instance solutions).

Meme evolution visualization with zoom 2x can already show when agents cooperate by exchanging memes (problem instance solutions): even in areas with high density of memes, we can still see the particular strategies of each agent. When information exchange happens, an arrow merge two paths into one. In green, we can see the correct solution for this problem instance (zoomed in the detail box). In our experiments, we let agents to finish solving the problem and all agents were able to solve the problem instance.

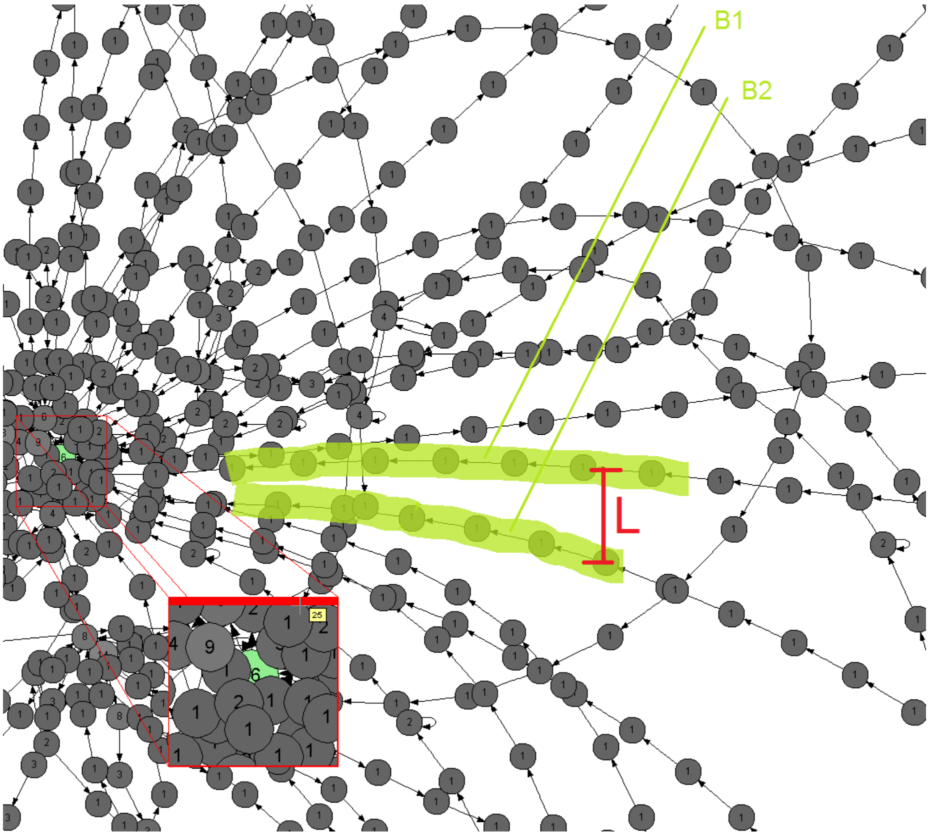


Figure 3.6 – Meme Evolution paths distant by L, but converging towards the solution. L is a measure for the meme space and represents homophilly intensity.

Meme Evolution paths distant by L, but converging towards the solution. L is a measure for the meme space and represents homophilly intensity: the symbol L represents homophilly between agents, measured by the distance between their strategies. In this visualization, homophilly is not dynamic over time. Highlight are paths that show visually the Continuity of Traits behavior (see Lemma 3.1.1) highlighted. In green, the solution of the problem, after a few minutes the most (and only) popular solution.

4 GOALS AND METHODOLOGY

In this chapter we will define our goals and present the methodology used in the research developed in this dissertation.

4.1 Goals

This dissertation aims at understanding factors that explain how individuals behave in online environments, specifically related to their performance when working on specific tasks. Previous work showed that Artificial Memetic Networks (AMN) can be used to model social behavior and to provide explanation for observed human cooperative problem solving (ARAUJO; LAMB, 2008)(ARAUJO, 2010). However, research on AMN has focused on computer simulations instead of human agents on a controlled environment or even in natural settings such as online social networks like Facebook or Google+. Adding to (ARAUJO, 2010), we aim at performing experiments modeled after AMN with the following objectives:

1. To model online human behavior experiments according to Artificial Memetic Networks (AMN).
2. To identify potential factors that affect human behavior in online environments using AMN and performing experiments to provide empirical evidence.
3. To apply the modeled factors to a new experiment and demonstrate how we can affect human behavior in online environments.

Therefore, our first goal is to provide a mathematical foundation for AMN in the context of online problem-solving social networks, one that can be unambiguously used and compared to other models and to design experiments that can be performed and tested. Our goal will be fulfilled when we can isolate a property, one that can be used to compare our model with others and one that can be found in the highest number of (experimental or observable) situations in social environments. Additionally, we want to design and perform experiments that are able to provide confidence around our model. We believe that our model is able to make more detailed behavioral descriptions that can pinpoint where theories diverge from observations and which phenomena can be attributed to social interaction. The methodology must comply with the scenario we exposed earlier: low cost experiments that are able to motivate individuals to act and generated larges amounts of data.

4.2 Methodology

To assess our model, based on AMN, we will perform experiments in connected environments in three different settings:

1. Experiment Phase I: we analyze what factors are important in objective truth problem solving. Individuals were challenge to solve SAT and Sudoku instances.
2. Experiment Phase II: we extend the results found during Phase I to problems with subjective truth, specifically to public policy polling, in a disconnected environment.
3. Experiment Phase III: we perform experiments that provide financial incentive to individuals and generalize results found in Phase I and II using Amazon Mechanical Turk.

Below, we provide details of each phase of experiments.

4.2.1 Experiment Phase I: Social Experiments

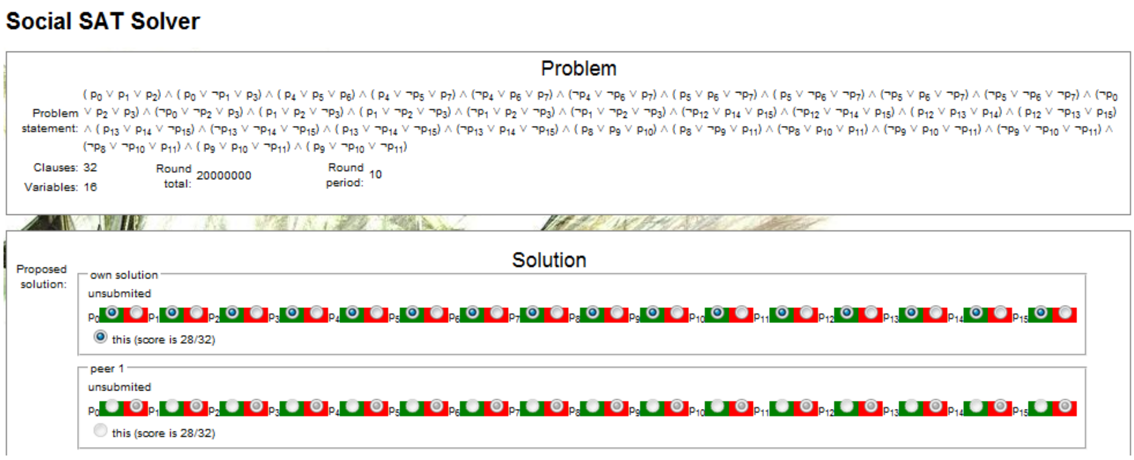
For Phase I, we planned several experiments where human agents are able to communicate in a controlled way and are asked to solve a mathematical challenge. A first group of individuals solved SAT instances (Boolean Satisfiability Problem instances). Instances of SAT were extensively studied and their solution space and equivalence to other problems is well known, making it easier to compare to agents solving other, equivalent problems. Another reason to have chosen SAT instances is that we wanted to analyze the performance of an Artificial Memetic Networks solving SAT instances individually or in group, aiming to possibly find new applications for collaboration.

A second group solved Sudoku instances. Sudoku is a game mathematically equivalent to SAT instances and thus suits our needs for a problem that can be more objectively compared regarding number of variables, difficulty and (intuitive) complexity or difficulty. However, it naturally provides motivation to individuals, diminishing effects of individuals having other interests but to win the game.

Both problems, SAT and Sudoku, should be used within a range of network settings, i.e. a varied number of agents, topology and problem instance characteristics, so we can reduce bias from our experimental setup. We will have met our goal if any behavioral pattern observed in the first experiment (SAT) can be also found in the second (Sudoku), showing that Artificial Memetic Networks offer predictability power while also providing results that can be generalized to other settings.

The experiments will be performed with two online web applications developed by our research group. Both applications allow collaborative solving of SAT and Sudoku instances. In each computer a subject will access the server through a web browser to participate in the experiment. The server is configured to allow users' solutions to be visualized by other users according to a selected topology and no other form of communication between participants was allowed. Moreover, the position of each neighbor's solution in this area is randomized before they were presented on the screen and then fixed until all instances were solved by the network members. When one subject submits a solution, it becomes available for her neighbors and it is shown on their respective screens, therefore it is not up to the user to select if he or she wants to share their solution. A button allows the user to copy any of her neighbor's solutions, making it her own as a way to make the copying process easier. The solutions can then be changed before being submitted again. A screenshot of the Social SAT Solver can be found in Figure 4.1.

Figure 4.1 – User interface of the Social SAT Solver (S3) experiment



For the experiment with Sudoku 4.2, users access a Facebook App where, unlike the SAT experiment, they can (potentially) join the experiment at will. Hence, the number of participants in each experiment may fluctuate.

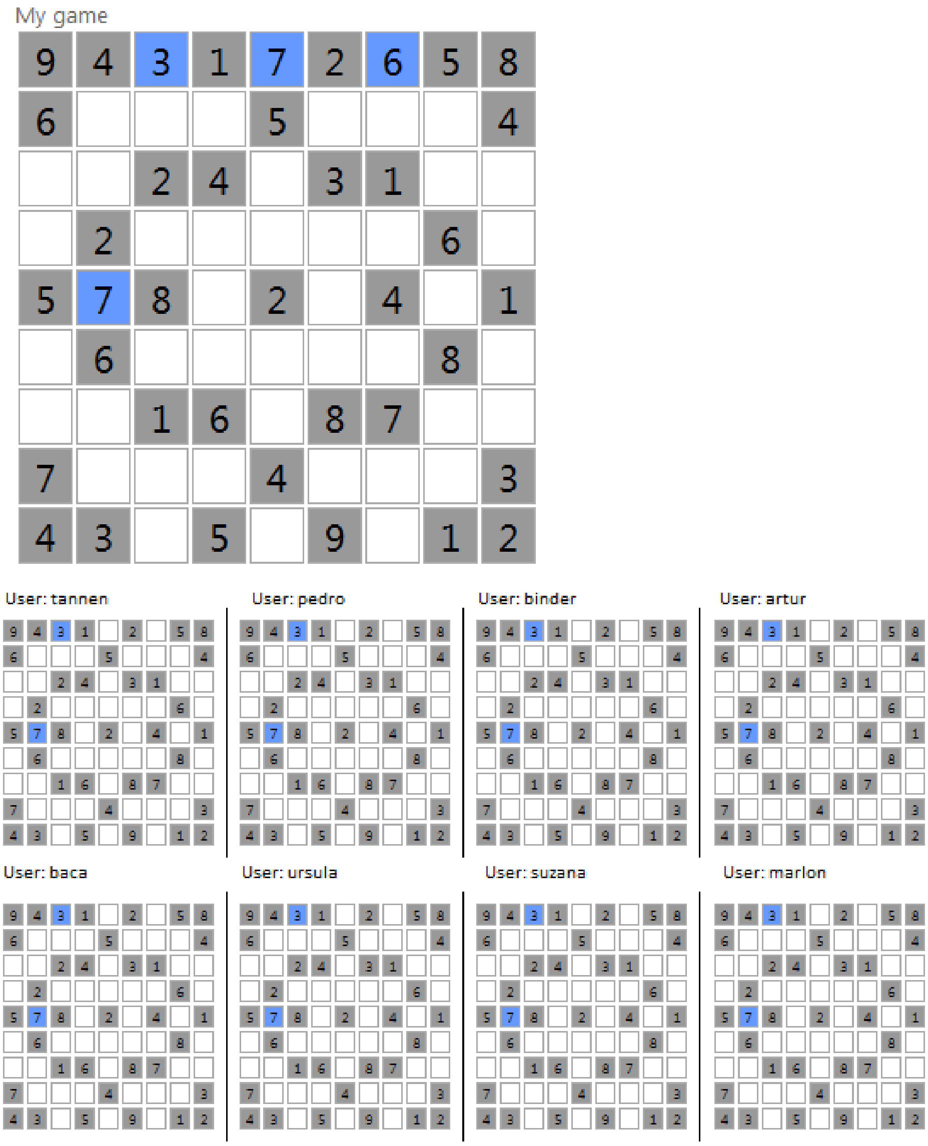
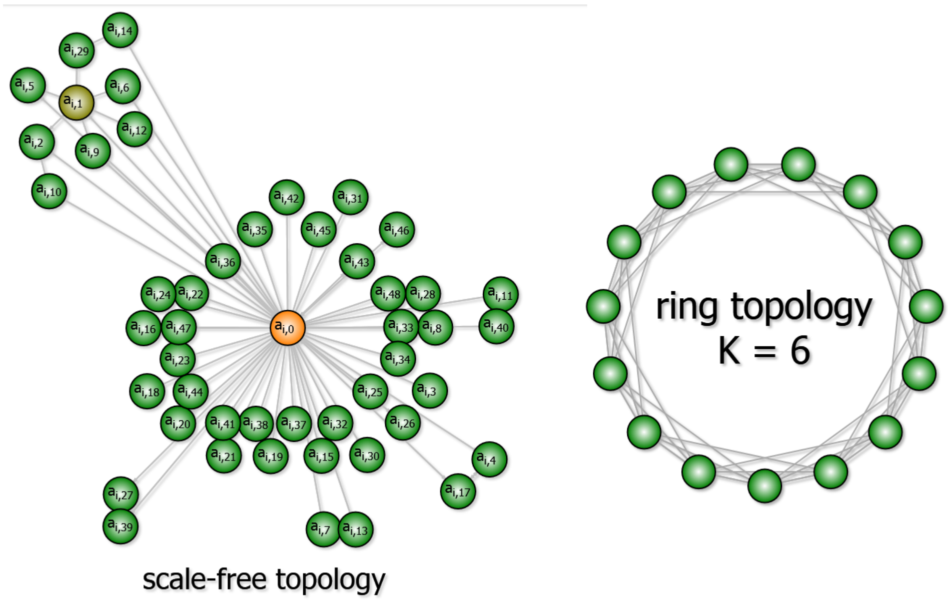


Figure 4.2 – User interface of the Facebook Sudoku experiment

It is assumed that the topology of social networks have a significant impact on the behavior of individuals and the group while solving problems (ARAUJO; LAMB, 2008; KEARNS; SURI; MONTFORT, 2006). In our work, we shall define a few topologies to perform the experiments. We use two ring networks (NEWMAN, 2010) with different neighborhood sizes, 4 and 6, where subjects are arranged in a regular ring. Moreover, a scale-free network (BARABáSI, 2002) is used, where the degree distribution follows a power-law with a fat tail, allowing for the existence of highly-connected subjects (*hubs*).



Two network topologies were used in our experiments. The left topology is a scale-free network and we will use a $\gamma = 1.65$, which is higher as hubs have a higher number of neighbours. The topology to the right is a a ring network where K sets the number of neighbour each peer will have. All peers have the same number of neighbours in the ring topology.

In order to perform a general analysis of the results, one has to consider variations in experiment runs, or, alternatively, accept that the analysis applies to fewer social networks setups. Also, variation in the number of human subjects participating at each experiment represents more closely what happens in non-controlled online social networks.

Experiments using the SAT and Sudoku problem will be performed within the laboratories at the universities and, before starting each session, a short explanation of the problem is given to the subjects. For the case of the Sudoku experiment, therefore, the App wil be only available to this restricted set of students. During the SAT experiment, we introduce the SAT problem, its relevance to Computer Science and artificial intelligence, and how subjects should use the interface of the application to create solutions. Instructions on how to visualize their peers solutions (when available) will also be presented.

In order to motivate their participation, we will announce that a global ranking will be published and a prize for the best solvers will be given, with the value of the prize increasing with the rank of the solver. The ranking mechanism can possibly provide the incentive needed for subjects that could not be interested in solving the instances alone. When all participants are ready, the system will allow them to start working on the selected problems. The experiment subjects consists of groups of first and second year Computer Science undergraduates. The common knowledge background of the students are basic acquaintance with propositional logic. No subject will have taken courses or is knowledgeable about search algorithms or any other SAT or Sudoku solving technique.

We planned a total of 30 meetings with students, totaling 30 experiments.

4.2.2 Experiment Phase II: Public Policy Poll

Our goal is to verify if the results from Phase I with Sudoku and SAT hold even in the context of non-objective problems of public policy polling.

In problem solving, it is often the case that the problem being solved has no objective answer and it often depends on the culture of the group. For instance, there is an objective answer for *What is the result of 1+1?*, but the answer to *What is the best football team?* will depend on the local culture, whether locality is defined by geographic features, a political map, ethnicity, etc.

In this new experiment, we share a hyperlink on Facebook where users could click and access a web tool to estimate how public policies affect the bus fare at the city of Porto Alegre[1]. As users click each policy item, an estimated value for the new bus fare was shown on the screen, as well as the individual value for each policy chosen. Therefore, the time to evaluate each option was greatly reduced as we showed the summed components of the bus fare. The set of options were chosen as to represent what was perceived to be the opinions of left and right wing individuals, as well as more centered positions. Therefore, we had each policy item have a monetary value and a subjective, political weight (from left to right wing, as perceived by us) associated with it. See Figure 4.3 for a screenshot of the web site.

As the Facebook web site currently works, users are allowed to share information with their social network, much like memes are exchanged as described in AMN. They do so by clicking a *Share* button or, in our experiment, on a similarly looking button *compartilhar a sua proposta*. All clicks and Facebook shares are recorded, along with geographical information of the users accessing the web site. The shared information can be used to separate between users that are exploring the options by clicking on each option, much like an introspective I action described in our model, or users that are exchanging information with the media in a extroverted action E.

[1]Porto Alegre is located on the Rio Grande do Sul state in Brazil.

Figure 4.3 – User interface of the Bus Fare Hike experiment

Bus Fare Hike experiment interface: users can hover with the mouse on top of each policy item to see the difference in price that each item would provide to the final bus fare. At the lower part of the page, the selected items would be summarized as a final bus fare. The bus fare selected could be shared on Facebook to stimulate more access to the web site, but also to control how the users who share information behave compared to the ones that do not share.

4.2.3 Experiment Phase III: Online Labor Market

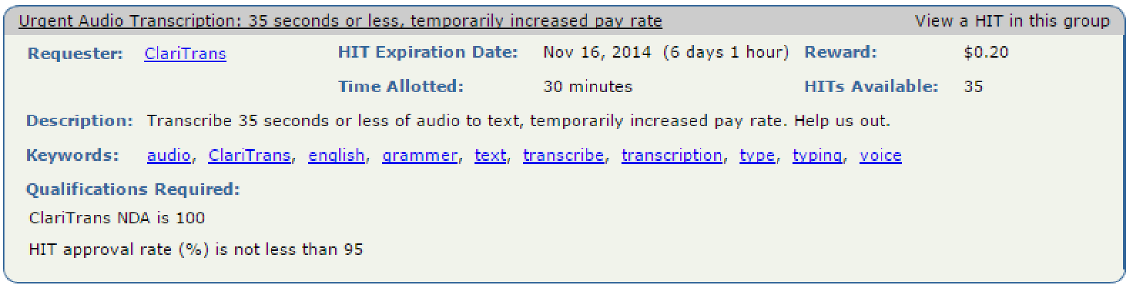
The experiments in Phase I and II provide incentives that are non-financial, i.e. the work users perform by seeking solutions and spending time working on problems are rewarded with gains other than money or other financial instrument. In this case, one could argue that our results cannot extend to the same scenarios from the mentioned applications.

To extend our results to internet applications that aggregate work, we analyzed data from experiments we performed in cooperation with Cornell University. The original intent of the project was to analyze how pay rate affects workers on Amazon Mechanical Turk, however we were able to use the generated data to our current goals. The original results of the project in cooperation are not directly related to the goals of this dissertation and are presented in Appendix B.3.

Amazon Mechanical Turk (AMT) is a web site where *requesters* post tasks such as *Transcribe an audio file* or *Tag five images*. Each task is named in AMT as a HIT (Human Intelligence Task) and individuals, known as *workers*, can agree to work on the tasks and receive a *reward* set by the requester if they complete the work. A requester can post as many HITs as he or she wants and the workers can accept as many tasks as they want. Is up to the requester to chose if a task was satisfactorily completed by a work. If so, the worker receives the reward, set by the requester and paid by AMT. The requester is responsible for buying credits for AMT and it has to do so before posting any task, aiming to assure the worker is going to get paid.

There are many web sites like AMT and to this class of application, where workers can get paid by working on tasks posted online, we call *Online Labor Market*. In the case of AMT, it is considered to be of microtasks, even though the tasks can take up to a few hours to work.

Before accepting to work on a task, the worker can have more details about the task conditions, such as amount of time they have to complete the task, after which they are not paid. Other details are present in Figure 4.4. The most important fields in our analysis will be title, description, reward and allotted time.


Urgent Audio Transcription: 35 seconds or less, temporarily increased pay rate — View a HIT in this group

Requester: ClariTrans **HIT Expiration Date:** Nov 16, 2014 (6 days 1 hour) **Reward:** $0.20

Time Allotted: 30 minutes **HITs Available:** 35

Description: Transcribe 35 seconds or less of audio to text, temporarily increased pay rate. Help us out.

Keywords: audio, ClariTrans, english, grammer, text, transcribe, transcription, type, typing, voice

Qualifications Required:

ClariTrans NDA is 100

HIT approval rate (%) is not less than 95

Figure 4.4 – Details of a task on Amazon Mechanical Turk (HIT)

A single HIT detailed: HITs are described by a *Title* (top line with text), the requester name, the expiration date (after which the HIT becomes unavailable for workers), a reward (in United States Dollars), a time allotted, which is the amount of time a worker has to complete the task after agreeing to work on it, the numbers of HITs available (of similar tasks), a description of the task, keywords and qualifications that the worker needs to have to accept the task.

Upon accessing AMT's web site, the worker is presented with a list of all available tasks with summary descriptions (Figure 4.5).

As the number of HITs for each task is shown, we can sample periodically each row of the presented list in Figure 4.5 to measure how many HITs have been worked by workers per unit of time, indicating the preference of workers and, aligning with other parameters, what are the

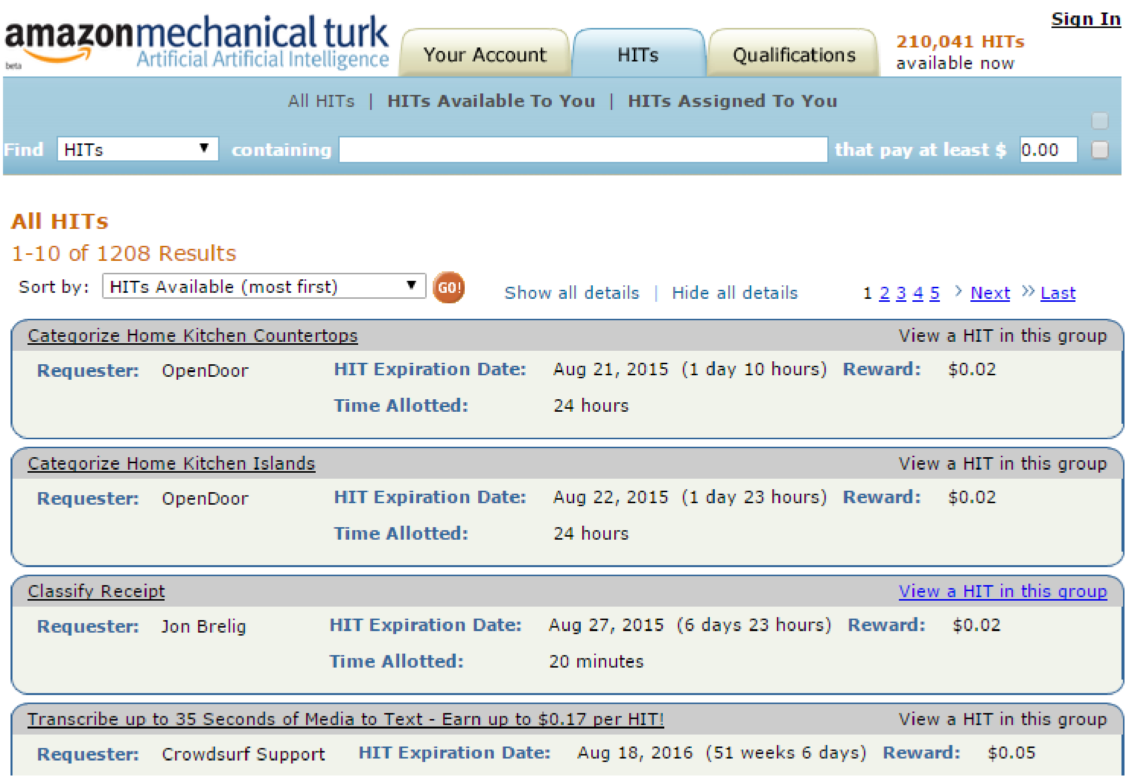


Figure 4.5 – Amazon Mechanical Turk web site interface

Amazon Mechanical Turk web site interface: the initial screen of AMT shows a list of HITs that can be accepted by workers. Before accepting, the worker can look into the details of the HIT. It is very important to notice that the default sorting for HITs is *HITs Available (most first)* (on the top, left corner of the screen). This default sorting has an important impact on the dynamics of workers on AMT. Also, the term *HIT Group* is used to aggregate tasks with the same description, but might have small differences between each other.

factors that affect the preference of workers on AMT.

5 RESULTS

We were able to identify a common pattern of aggregation behavior across all experiments. Surprisingly, agents were, on a first analysis, more importantly affected by the interface then by the quality of solutions that could be exchanged with their peers. In the following sections, we will explain the results achieved in each experiment and how have we arrived at our conclusions. This result was published across 2 abstracts and 1 paper (FARENZENA; LAMB; ARAUJO, 2010),(FARENZENA; ARAUJO; LAMB, 2011b) and (FARENZENA; ARAUJO; LAMB, 2011a) (abstracts can be found in Appendix I) and a paper submitted. In the following section, we will describe the social experiments scheduled and executed, along with analysis in the first step of our series of experiments.

5.1 Experiment Phase I: Social experiments

During Phase I, published in (FARENZENA; ARAUJO; LAMB, 2011a), we ran 42 experiments with the participation of 160 human subjects. At the time of performing the experiments, the exact number of participants was unknown and thus not informed at the methodology section of this dissertation.

As the time to collectively solve each instance was different in each group of agents (from 2 to 35 simultaneous agents in the network, with total 125 different agents throughout all the experiments) and in each topology, each participant did not solve the same number of instances in all experiment runs.

For the SAT experiment, solutions are presented in a row, while for Sudoku they are presented in a grid (Fig. 4.2). It is known that the topology of social networks have a significant impact on the behavior of individuals and the group while solving problems (ARAUJO; LAMB, 2008; KEARNS; SURI; MONTFORT, 2006). We use two ring networks (NEWMAN, 2010) with different neighborhood sizes, 4 and 6, where subjects are arranged in a regular ring. Moreover, a scale-free network (BARABáSI, 2002) is used, where the degree distribution follows a power-law with a fat tail, allowing for the existence of highly-connected subjects (*hubs*).

The SAT experiments were performed along a planned 30 experiment sessions, detailed in table 5.1. Out of 30 planned experiments, only 14 were scheduled and 8 were considered successful. Planned experiments would not be scheduled mainly due to overlapping academic activities, such as exams and fairs. Scheduled experiments failed mainly due to software bugs, hardware or infrastructure failure or a poor set of SAT problem instances available (something only found while performing the first set of experiments).

Table 5.1 – Schedule for the Collaborative Problem-Solving SAT Experiments

Date	*Time*	*Type*	*Result*	*Comment*
9/23/2009	10:30	Individual	Negative	Server hardware failure.
9/30/2009	10:30	Individual	Positive	SAT instances proved to be too easy.
10/6/2009	15:30	Individual	Positive	Individuals asked to fill report.
10/7/2009	10:30	Individual	Negative	
10/8/2009	15:30	Individual	Negative	Additional SAT instances were not ready.
11/9/2009	8:30	Individual	Positive	Network hardware failure, but still worked.
11/9/2009	10:30	Individual	Positive	
11/10/2009	15:30	Group	Negative	Software bug.
11/11/2009	8:30	Group	Negative	Software bug.
11/11/2009	10:30	Group	Negative	Software bug.
11/12/2009	15:30	Group	Positive	Ring N=26; $K_{ring} = 4$ and $K_{ring} = 6$.
11/18/2009	10:30	Group	Positive	
12/8/2009	10:30	Group	Positive	Software bug, but still worked.
12/14/2009	10:30	Group	Positive	Ring k=4.

Positive result express that data was collected successfully in the desired experiment setup. Negative result express that due to several possible reasons the data collected had to be discarded. Among important reasons for a failure are: a software bug, a hardware failure, data corruption or inconsistency or the parameters for the experiment were not correctly set.

Next, we show the summary for data collected for the SAT experiments (table 5.2) and Sudoku experiments (table 5.3).

In column *topology* we classify experiments by the topology we have used. In column *experiments* we represent the number of instances submitted for each topology, and the number of different problem instances used. We have used the same instances a few times to assess if subjects could recognize repeated instances; only 2% were able to do so. The *Agents* column shows the number of human subjects that participated in the experiments (there is a range of subjects instead of a fixed number, because the number of subjects changed from one problem instance to another due to subjects' availability). The *Solutions* column accounts for the number of distinct solutions subjects provided or copied from neighbors during all experiments for each topology, while in brackets we present only the percentage of solutions copied. The *Variables* and *Clauses* columns are specific to Table 5.2 (SAT experiment).

Our hypothesis was that the position of a peer's solution would hardly matter, as agents would seek copying the best solutions shown evaluating each solution. In order to test this hypothesis, we plotted the number of times a neighbor solution was copied as a function of its position in the screen for the SAT problem (Fig. 5.1). We can see this plot as the probability of the k_{th} solution being copied by a neighbor. Considering two possible outputs (copy or not),

Table 5.2 – Experiment runs for SAT instances

Topology	Experiments (instances)	Variables	Clauses	Agents	Solutions (copied)
Disconnected	16 (16)	3 to 26	5 to 63	125	-
Ring ($K_{ring} = 4$)	4 (3)	4 to 8	16 to 16	16 to 19	335 (13%)
Ring ($K_{ring} = 6$)	3 (2)	8 to 10	12 to 16	22 to 31	268 (14%)
Scale free ($\gamma = 1.65$)	11 (9)	4 to 26	8 to 63	1 to 32	1179 (10%)
Total	34 (24)	3 to 26	5 to 63	1 to 125	7630 (11%)

* Total exchanged solutions ratio excludes solutions from disconnected topology

Table 5.3 – Experiment runs for Sudoku instances

Topology	Experiments	Agents	Solutions (% copied)
Ring ($K_{ring} = 4$)	2 (2)	2 to 5	533 (25%)
Ring ($K_{ring} = 6$)	2 (2)	17 to 20	11114 (18%)
Scale free ($\gamma = 1.65$)	4 (3)	14 to 35	6012 (87%)
Total	8 (6)	2 to 35	17659 (42%)

and a discrete distribution, then we can model it as a geometric distribution. We fitted this function as a probability mass function of the geometric distribution where $\langle X(k) \rangle$ denotes the probability of an agent copying the k_{th} neighbor solution, with parameter $p = 0.5479$:

$$\langle X(k) \rangle = (1 - p)^{k-1} p \tag{5.1}$$

As it turned out, the frequency that a solution from a neighbor is copied decreases with how far to the right it is in the solution's row. Therefore, agents were picking closest solution on the interface instead of using the best solutions available to make their choices. Since we connected agents randomly, if they were to pick the best solution, the distribution $X(k)$ would be uniform, which is not the case in our experiments. This suggests that *subjects select the most readily available solution, which may not be the best solution of the given problem instance*. Also, in Table 5.2 we show that the number of neighbors could be as high as 20 (for Scale-Free networks), which implies that agents had up to 20 solutions available. The limitation to 20 was set because of the total number of agents (up to 35), divided in two nodes with high number of connections, representing the power law distribution of nodes. However, almost all solutions copied at least once were along the first 6 positions, even though agents had access to far more solutions than those in these positions. That is, even when participants were not able to improve their solution and chose to copy a neighbor's instead, they still would not go over all neighbors' solutions to try and find the best solution, choosing instead to copy some seemingly random solution from the pool.

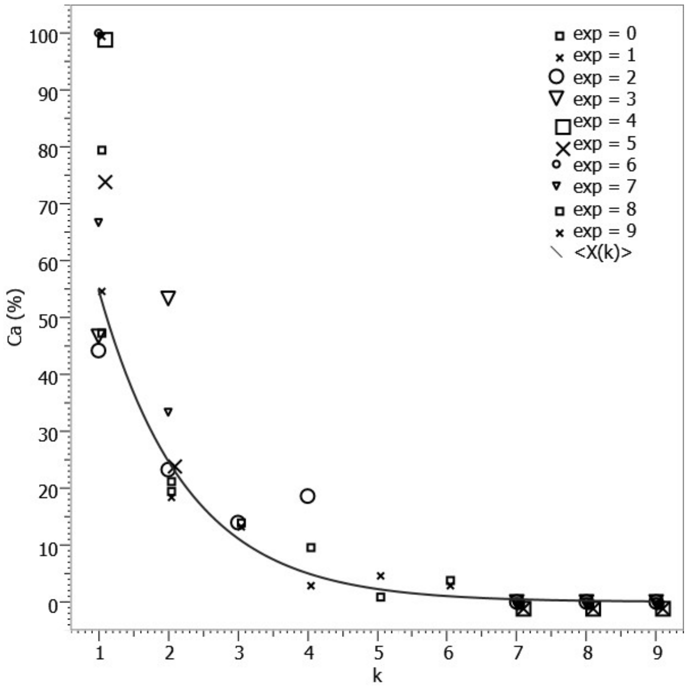


Figure 5.1 – Positional influence on user strategy

Each point shows the percentage of times Ca a solution on position k on the interface was copied. When the solution was copied, the agent that copied the solution had solved exp instances prior to that instance. $\langle X(k) \rangle$ is the probability mass function of the geometric distribution that fitted the data with 5% significance level.

In the sequence, we performed this same analysis for the Sudoku experiment (the estimation of the probability of a neighbor's solution to be copied as a function of its position on the interface) and found a similar pattern. We fitted the function as a negative binomial distribution. To assess fitness, we tested the collected data for the null hypothesis of a normal distribution with *Anderson-Darling test* and rejected the null hypothesis with 5% significance level. Comparing the modeled geometric distribution with the sampled data in a *two-sample Kolmogorov-Smirnov test* we did reject the null hypothesis of the distributions, with the exception of the vertical axis of the Sudoku experiment, where the null hypothesis was not rejected.

However, in a AMNmodel that follows principles of natural selection, it is not really necessary for agents to evaluate each neighbor solution. This apparently random behavior explains cooperation with the following proposed *procedure for human cooperation*. An agent has to solve his/her own problem, and it tries to do so by proposing solutions and getting feedback from the environment. The environment is a virtual problem-solving activity where solutions are objectively evaluated by a computer. Thus an agent can “navigate” through the problem solutions using only this feedback without the need for external sources, such as peers’ shared solutions. Even with varying degrees of problem solving skills, all agents can solve alone the problem instances, as verified by the first part of the experiment where all agents correctly solved problem instances without being allowed to cooperate.

In that sense, with varying skills, all agents individually create solutions aiming at reaching the correct solution. If, for instance, an agent opts to copy a peer’s solution, then her own solution is completely replaced by the peer’s and the agent will continue to improve upon this solution, even if the copied solution was not strictly better then the original. If we take the average solution distance of all agents to the correct solution, we will find that it decreases with time, as all agents are able to solve the problem. Thus, even with a non-deterministic choice of neighbor for cooperation, an agent achieves the correct solution sooner or later as it modifies the solution towards the correct solution.

In the case where a specific agent is surrounded by agents that share the very same solution, when the agent chooses to cooperate (i. e. exchange solutions), the probability that this solution will be copied is 1, which explains with AMN the *conformist* behavior already discussed and proposed in (EFFERSON et al., 2008). In a non-virtual environment, the feedback of the environment can cause an agent to stop communicating, e.g. in the case of injury or death. In this case, the agent cannot communicate his solution to the problem and the probability of this agent’s solution getting copied is effectively zero. In conclusion, copying is a justified behavior and necessary for survival of the group, as it only happens between agents able to do so (presumably the ones that are still alive to communicate their solution). Although this has been argued in many fields for a long time, a direct verification of such ideas through experimentation was not possible as a result of insufficient experimental data and a model that could be used as a general representation of human behavior.

Finally, for problem-solving social networks, cooperative behavior cannot be described in a model with deterministic rational agents, our results evidence that solutions have not been evaluated. Indeed, if we had the case of a real environment such as the one described above, agents would not need to be rational at all, since the existence of many agents would mean many random solutions being tested against the environment; bad solutions could result in an agent stopping communicating while good solutions would still be around, providing further chances for other agents to copy them and survive the environment. Thus, our results show that in our settings *cooperation works because agents can copy each other*.

However, even concluding that random acting agents can survive in a real environment, in our experiments they do not change their solutions randomly, since we know that agents can solve problems by themselves. Another evidence that agents do not act randomly is that the feedback in our experiments is not like in a real world, i.e. there is no possibility that a bad solution will cause the agent to stop communicating the solution. Therefore agents need other mechanism to evolve alone their own solutions, such as reasoning.

Mapping to our original mathematical framework and referencing equation 3.1.5, it is initially assumed the existence of function eval and for other modified versions of the *Aggregation Step*. As each option, in our model regarded as a τ, is evaluated, the chance of being copied decreases. We hypothesize that Phase I provides evidence that eval has an internal state influenced by previous evaluations. We could think about time influence, however the every time an individual choose to evaluate peers solutions, even after several repeated experiments, the evaluation for the first solution remained at the same level. Therefore, we assume that the internal state is not time, but the cost of having evaluated other solutions τ_{-i} before.

$$\text{eval}(\tau_{-i}) : (\tau_{-1}, \text{eval}_{t-1}) \Rightarrow \mathbb{R} \tag{5.2}$$

We have mapped the $eval$ function to real numbers $\mathbb{R}$ for simplicity. We can chose any function with an order defined over the T domain, i.e. the domain of all traits.

5.1.1 A limit for the number of connected peers

There is a quantitative limit for the number of peers with which a subject will interact with in computational social problem-solving. This number was close to 6 in our experiments. This result suggests that, when choosing a topology for designing a problem-solving social network, one may limit the degree of nodes to a maximum number that is relatively small, leading to a sparse network that may be easier to handle. Although we do not generalize the limits found in our experiments to other problem-solving social networks, it is possible to use the social network experimentation methodology used in this work to find this maximum number for any other given problem. For instance, in a (hierarchical) organization, each working group probably has a maximum number of members, which may be determined by the methodology used in our work, i.e. one can find out how to collect communication or solution data exchanged between participants and model the data according to a geometric distribution. This is relevant in collaborative work and collective intelligence applications (BRAN et al., 2009; NOWAK, 2006; WOOLLEY et al., 2010).

5.1.2 Capability to forget

Around $5.31\% \pm 6.93$ of all solutions used by all agents were, at some point, already provided by those same agents. Consequently, agents do forget some of their previous attempts to solve the problem and we cannot model the problem as if agents were unable to lose traits.

We presented a model that provides information on how humans interact when performing computation in a social network through AMN. From extensive experimentation with human subjects networks we provide empirical evidence of cooperation emergence as a result of AMN.

5.2 Experiment Phase II: public policy poll and internet wide protests

Following our model of agent interaction from eq. 5.1, we plotted the same graph where we analyze how people are affected by the position of options on the application interface. Each element k on the screen receives consecutively less interaction from the user, as seen in Figure 5.2. The type of *interaction* a user may perform is rather ample and is specific to what users are allowed to do in experiments.

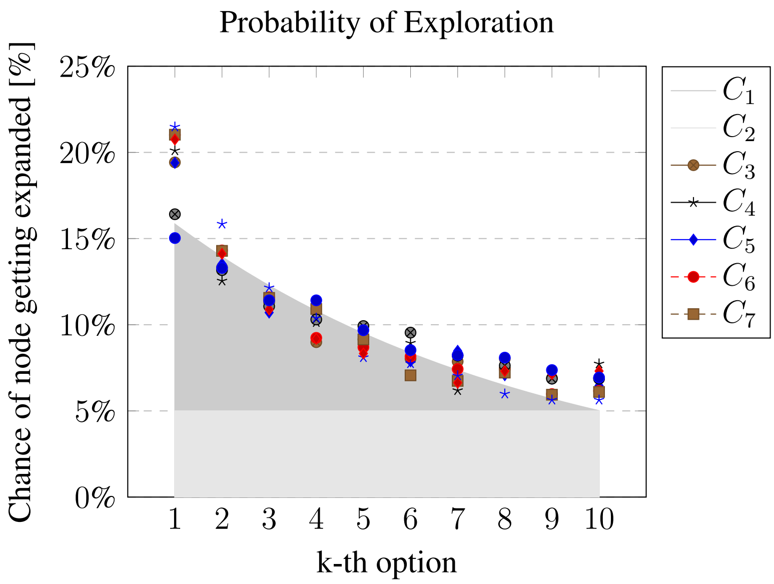


Figure 5.2 – Positional influence of public policy options

Positional influence of public policy options: each layout presents the above distributing of element exploration by the users. The elements are linearly ordered, vertically.

$$\langle X(0) \rangle < \langle X(1) \rangle < ... < \langle X(N) \rangle \tag{5.3}$$

As a consequence, no matter what is first element of the screen is shown, the user will interact more with it. As we observed in experiments, if the experiment is a poll, then the user has a high probability of picking the first option of the poll, and close to zero of picking the last one. Therefore, choosing the order of elements on the screen can determine the result of poll, assuming $r(k)$ is sufficiently small. To avoid influence by the layout, and thus removing the undesired behavior that pollutes the collected data with type 2 users, with propose randomizing the elements of the layout of the application.

When we randomize the layout, let us suppose that each layout will be displayed $\xi(0) = a$ times with order of elements A, $\xi(1) = b$ times with order of elements B and so on, where $a + b + ...$ and M is the total number of users interacting.

To verify our models, we applied our method to the poll experiment. Our result on Figure 5.2 shows the distribution of interaction when only the distribution of interaction over one layout is applied. All layouts showed the same pattern. However, if we do randomize uniformly each order of options (thus a different layout) we see results of Figure 5.3, which represents the voting without the influence from non-motivated agents, thus fulfilling our promise of delivering only high quality data from users.

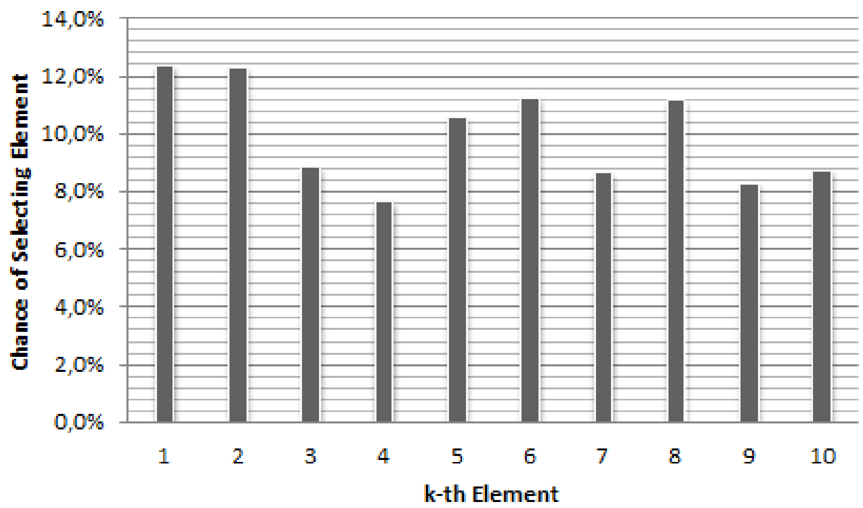


Figure 5.3 – Randomizing the options order on the layout for policy options

Randomizing the elements on the layout we cancel out the effect of data provided by users that selected options rationally, but according to a distribution not related to the value of the options as solutions to the problem domain.

Table 5.4 – Data Collected from Amazon Mechanical Turk

	Our study	Chilton†	Ratio
Setup	All pages	3 pages	
Total pages	240	36	15.00%
Total time (hours)	1,440	32	2.22%
Distinct HIT Groups	67,181	2,040	3.04%
Total HITs	3,696,730	2,207,548	59.72%
Sum of reward (USD)	$531,313.18	$232,051.15	43.68%

†For reference, see (CHILTON et al., 2010).

5.3 Experiment Phase III: the case with financial incentives

We have performed a data sampling of each task present on AMT each 5 minutes on average along 2 months uninterruptedly [1]. Table 5.4 summarizes our collected data and compares to a paper where data from AMT was also analyzed. The comparison has been made for an unrelated discussion to this chapter's goal, found in Section B.3.

To compare AMT with our experiments, we have to compare the decay rate of total number of tasks per HIT Group with the number of exchanged solutions on our experiments. Our goal is to prove that no other factor is more important to worker activity than the order at which items (τ_i) are evaluated. Even considering the amount of money rewarded to workers or the amount of time they spent working, the position in the screen, and thus the order in which their are presented, is the most important factor. This is an outcome of AMT choosing as a default sorting the number of HITs available per group, once again allowing us to identify that searching is expensive to individuals and they might chose not to be picky or to evaluate each option individually.

For our investigation, we performed a linear regression with ridge regularization over our data and with 5 explanatory variables:

- **Number of HITs**: it is the total number of HITs per HIT Group. This is the field that is sorted by default on AMT.
- **Alphabetical order**: what is the influence of the order of title alphabetically.
- **Expiration time**: how close the HIT Group is close to expiration and thus being removed from AMT list of available HITs.
- **Reward**: the amount of money the worker gets upon completing and having the work approved by the requester.

[1] Except for a few hours when computer infrastructure was not available.

Table 5.5 – Linear Regression Analysis of AMT

	Estimate	Standard Error	t-Student	p-Value
(Intercept)	-345.46	29.462	-11.726	$1.1553e{-31}$
Rank by number of HITs	1.2442	0.11187	11.122	$1.1555e{-28}$
Rank by alphabetical order	0.061772	0.011668	5.2943	$1.2055e{-07}$
Rank by expiration time	-0.016155	0.0097449	-1.6578	0.097375
Rank by reward	-0.0059581	0.0089948	-0.66239	0.50773
Rank by time allotted	0.046183	0.0083148	5.5543	$2.8178e{-08}$

Number of observations: 23460, Error degrees of freedom: 23454
Root Mean Squared Error: $1.25e03$
R-squared: 0.0103, Adjusted R-squared 0.0101
F-statistic vs. constant model: 48.7, p-value = $2.61e{-50}$

The initial screen of AMT shows a list of HITs that can be accepted by workers. Before accepting, the worker can look into the details of the HIT. It is very important to notice that the default sorting for HITs is *HITs Available (most first)* (on the top, left corner of the screen). This default sorting has an important impact on the dynamics of workers on AMT. Also, the term *HIT Group* is used to aggregate tasks with the same description, but might have small differences between each other.

- **Time allotted**: the amount of time a worker has to complete the task.

Our results, shown in Table 5.5, do confirm that the HITs are more influenced by any other factor in our analysis However, the statistic R^2 was very low, pointing that our model is incomplete.

Assuming that other requesters are also aware of the fact that being on the first page of AMT is a key factor to attract workers, we can state that AMT is being gamed by other requesters and that the first page do not actually represent the tasks with the highest number of real tasks available. Most tasks are fake and only created to inflate the number of total HITs available for a given task.

The idea that being the amongst the first HIT groups shown on AMT first page will increase the *consumption rate* of HITs, also known as *uptake rate*, might not be new. Indeed, if we assume that, we should adjust our linear regression analysis by removing any page that is at the first positions in any sorting method. We performed a new analysis without the first position HIT Groups and found the results shown in Table 5.6.

All significance levels improved and now expiration time and time allotted explanatory variables have confidence levels $p < 0.05$. However, we still have low R^2 levels.

We have shown that AMT does behave in similar ways to our previous experiments by sampling data from AMT web site and performing a linear regression over the collected data. We have shown once again that variables that are usually taken as the most important factors,

Table 5.6 – Linear Regression Analysis of AMT, corrected

	Estimate	Standard Error	t-Student	p-Value
(Intercept)	-183.87	13.825	-13.299	$3.2764e{-}40$
Rank by number of HITs	0.78528	0.052312	15.012	$1.069e{-}50$
Rank by alphabetical order	0.030653	0.005438	5.6368	$1.7521e{-}08$
Rank by expiration time	-0.016715	0.0045583	-3.6669	0.0002461
Rank by reward	-0.0081355	0.0042198	-1.9279	0.53878
Rank by time allotted	0.022	0.0038769	5.6745	$1.4078e{-}08$

Number of observations: 23142, Error degrees of freedom: 23136
Root Mean Squared Error: 579
R-squared: 0.0167, Adjusted R-squared 0.0165
F-statistic vs. constant model: 78.8, p-value = $2.95e{-}82$

First results excluded : removing the first results improved R^2 and the significance levels for the reward and time allotted explanatory variables.

such as reward and time allotted, are not as important as interface factors such as the position of the tasks in the screen. However, our data analysis, although with safe p values, still had a low R^2 and thus there are more factors that have to be taken into consideration. On section B.3 we present ongoing research where we offer an explanation on why linear regression methods fail to provide a complete model for our data and that a few articles that analyzed AMT data must be reviewed under the light of our discoveries.

6 CONCLUSION, DISCUSSION AND FUTURE WORK

This dissertation assumed initially that the connected nature of online applications could be used as a base to explain online applications that make the use of human work. As our first experiments were conducted, described in Chapter 5, we realized that when individuals are solving microtasks, then a strong component of the cost of searching and evaluation affects how people make decisions in connected environments. Indeed, the experiments provided evidence that cooperative behavior is strongly biased to the high costs of searching and evaluating compared to the gains of evaluating each peer solution individually. We showed that individuals, when performing tasks in online environments, have search and evaluation costs comparable to gains of cooperating and thus need to reduce costs by searching less and spending less time evaluating options. The result holds even as individuals gain experience by performing the experiments repeatedly along a few hours, showing that they are not learning and modifying their behavior, or that their behavior is not a consequence of a purely exploratory use of the applications they used to solve tasks.

Our model of cooperation implies individuals copying solutions from each other. There are other cooperation models where individuals need to simultaneously agree to connect to each other, as in a dynamic topology network, or where they need to agree in sharing information with connected nodes. New experiments should be performed within other cooperation models found in literature to extend our results.

As it happens with similar experiments where individuals are challenged to solve mathematical problems (JUDD; KEARNS; VOROBEYCHIK, 2010), our results could be possibly explained by specific interface features, like the way we organize elements on the application layout, or even a particular phenomena that only happens in problems that can be modeled as SAT instances.

To improve our investigation and clarify if our experiments were biased in some way, we extended our study with a new round of in an application of public policy polling, which became known as Experiment II. This time, the task had no objective truth, although each option had an intrinsic financial value that could potentially impact of saving the participant could have in the future if the public policy passed. Moreover, the layout was displayed vertically. Once again, we were able to identify the same behavior when deciding upon one or another option, i. e. the chance to vote for a specific public policy would decrease as more options were shown.

Experiments I and II were performed with voluntary participation. One could argue that the lack of financial incentive could decrease so much the gains for the efforts of each participant that the cost of evaluation would easily overcome the gains. With the aim of investigating that this is the case, we needed to experiment with an application where individuals received financial incentive, which is the case of the online labor market of Amazon Mechanical Turk. This experiment was not controlled and we only collected data from Mechanical Turk. First, we were able to show that the individuals are heavily affected by search and evaluation costs, even when they could be able to find tasks with higher rewards, by searching and evaluating properly using search tools that are less costly. Most importantly, by gaming the market and moving our tasks to the first positions of the default task display, we were able to show that we can affect, using our model, the decision making process of individuals and have the value they attribute to each task go very high, indeed higher than financial incentives.

Therefore, we make the point of how important it is to consider search and evaluation costs when humans perform microtasks and we extended results to objective and subjective problems and to non-financial and financial incentives. This dissertation presented a mathematical description of traits transmission based on Artificial Memetic Networks and showed how the evaluation function of the *Aggregation Step* behaves in several settings, shedding light on online problem-solving social networks.

The costs of search and evaluation during problem-solving or task performing obey the same power law across all experiments, although with different parameters. For the case of financial incentives, the model did show how optimizing the search costs is more important than financial incentives, but the error indicated the model needs more explanatory variables or a improved model to provide a more complete explanation of gains and costs in a labor market.

The experiments with Mechanical Turk showed a low R^2 value, which directed us to exploring more about properties of AMT (Appendix II). Additionally, also found in Appendix II, a spin-off of the research that points to a method that can be used to filter the quality of the work output of workers in the future.

Although we can point to consistent results across experiments, but not without limitations. First, we applied the *meme* concept to small SAT and Sudoku solutions and then we suggest that they can be compared between each other with a given function, such as Levenshtein distance. However, this comparison would not be tractable for other non-continuum spaces of problems, such as valid binary representation of programs, or if we chose a distance for which the computational complexity grows quickly with the size of the input. Therefore, we have considered, in our experiments, cases where the solution has a small input. For instance, a SAT solution with 64 variables can be represented in a 64 bits (or 8 bytes in most architectures), while a Sudoku solution can have up to to 299 bits. If we tried to apply our *meme* model to natural language problems, such as interpreting human conversation or books, the input size could span from bytes to several megabytes.

Our research was limited to microtasks and we have not extended our results to long-running problems. Although we have provided evidence that individuals do not change their behaviour after repeated experiments, the total time a given individual worked from first problem instance to the last did not span more than 1 hour. Experiments that span days or maybe even months are required to understand the behavior of individuals in long-running problems.

REFERENCES

ABOWD, G. D.; MYNATT, E. D. Charting past, present, and future research in ubiquitous computing. **ACM Transactions on Computer-Human Interaction (TOCHI)**, ACM, v. 7, n. 1, p. 29–58, 2000.

AHN, L. V.; DABBISH, L. Designing games with a purpose. **Communications of the ACM**, ACM, v. 51, n. 8, p. 58–67, 2008.

AHN, L. von; CATHCART, W. **Teaching computers to read: Google acquires reCAPTCHA**. 2009. Disponível em: <http://googleblog.blogspot.com.br/2009/09/teaching-computers-to-read-google.html>.

AHN, L. von et al. reCAPTCHA: human-based character recognition via Web security measures. **Science (New York, N.Y.)**, v. 321, n. 5895, p. 1465–8, set. 2008. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/18703711>.

AHN MANUEL BLUM, N. J. H. Luis von; LANGFORD, J. CAPTCHA: Using Hard AI Problems For Security. **Proceeding EUROCRYPT'03 Proceedings of the 22nd international conference on Theory and applications of cryptographic techniques**, 2003.

ANDRADES, R. K. de et al. A cluster-DEE-based strategy to empower protein design. **Expert Systems with Applications**, Elsevier, v. 40, n. 13, p. 5210–5218, 2013.

ARAUJO, R. M.; LAMB, L. C. Memetic networks: analyzing the effects of network properties in multi-agent performance. In: **AAAI'08: Proceedings of the 23rd AAAI Conference on Artificial Intelligence**. [S.l.]: AAAI Press, 2008. p. 3–8. ISBN 978-1-57735-368-3.

ARAUJO, R. M. d. **Memetic networks : problem-solving with social network models**. Tese (Doutorado) — Universidade Federal do Rio Grande do Sul, 2010.

BARABáSI, A. **Linked: The New Science of Networks**. [S.l.]: Basic Books, 2002.

BIRUKOU, A. et al. A formal definition of culture. In: **Models for Intercultural Collaboration and Negotiation**. [S.l.]: Springer, 2013. p. 1–26.

BRAN, C. A. et al. The knowledge worker of the future. In: **Proc. OOPSLA'09 Companion**. [S.l.]: ACM, 2009. p. 851–852.

BREDIN, J.; PARKES, D. C. Models for truthful online double auctions. **arXiv preprint arXiv:1207.1360**, 2012.

CHILTON, L. B. et al. Task search in a human computation market. In: ACM. **Proceedings of the ACM SIGKDD workshop on human computation**. [S.l.], 2010. p. 1–9.

CSIKSZENTMIHALYI, M.; CSIKZENTMIHALY, M. **Flow: The psychology of optimal experience**. [S.l.]: HarperPerennial New York, 1991. v. 41.

DAVID, E.; JON, K. **Networks, Crowds, and Markets: Reasoning About a Highly Connected World**. New York, NY, USA: Cambridge University Press, 2010. ISBN 0521195330, 9780521195331.

DAWKINS, R. **The Selfish Gene**. [S.l.]: Oxford University Press, 1990.

EFFERSON, C. et al. Conformists and mavericks: the empirics of frequency-dependent cultural transmission. **Evolution and Human Behavior**, v. 29, n. 1, p. 56–64, 2008.

FARENZENA, D. S.; ARAUJO, R. M.; LAMB, L. C. Collaboration Emergence in Social Networks with Informational Natural Selection. In: **2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing**. [S.l.]: IEEE, 2011. p. 555–559. ISBN 978-1-4577-1931-8.

FARENZENA, D. S.; ARAUJO, R. M.; LAMB, L. C. Towards social problem-solving with human subjects. In: **IJCAI Proceedings-International Joint Conference on Artificial Intelligence**. [S.l.: s.n.], 2011. v. 22, n. 3, p. 2798.

FARENZENA, D. S.; LAMB, L. C.; ARAUJO, R. M. Combining human reasoning and machine computation: Towards a memetic network solution to satisfiability. In: **AAAI**. [S.l.: s.n.], 2010.

FEHR, E.; FISCHBACHER, U. The nature of human altruism. **Nature**, Nature Publishing Group, v. 425, n. 6960, p. 785–791, 2003.

FLOUDAS, C. et al. Advances in protein structure prediction and de novo protein design: A review. **Chemical Engineering Science**, Elsevier, v. 61, n. 3, p. 966–988, 2006.

FoldIt. 2015. Disponível em: <https://fold.it/portal/>.

GalaxyZoo. 2015. Disponível em: <http://www.galaxyzoo.org/>.

GHOSH, A.; HUMMEL, P. A game-theoretic analysis of rank-order mechanisms for user-generated content. **Journal of Economic Theory**, Elsevier, v. 154, p. 349–374, 2014.

GILES, J. Internet encyclopaedias go head to head. **Nature**, Nature Publishing Group, v. 438, n. 7070, p. 900–901, 2005.

IPEIROTIS, P. G.; PROVOST, F.; WANG, J. Quality management on amazon mechanical turk. In: ACM. **Proceedings of the ACM SIGKDD workshop on human computation**. [S.l.], 2010. p. 64–67.

JAIN, S.; PARKES, D. C. A game-theoretic analysis of the esp game. **ACM Transactions on Economics and Computation**, ACM, v. 1, n. 1, p. 3, 2013.

JUDD, S.; KEARNS, M.; VOROBEYCHIK, Y. Behavioral dynamics and influence in networked coloring and consensus. **Proc. Natl. Acad. Sci. U.S.A.**, v. 107, n. 34, p. 14978–14982, 2010.

KEARNS, M.; SURI, S.; MONTFORT, N. An experimental study of the coloring problem on human subject networks. **Science**, v. 313, n. 5788, p. 824–7, 2006.

KREPS, D. M. **A course in microeconomic theory**. [S.l.]: Harvester Wheatsheaf New York, 1990.

LILLIBRIDGE, M. et al. **Method for selectively restricting access to computer systems**. Google Patents, 2001. US Patent 6,195,698. Disponível em: <http://www.google.com/patents/US6195698>.

LINTON, R. The study of man: An introduction. 1936. Disponível em: <http://www.getcited.org/pub/101177595>.

LINTOTT, C. et al. Galaxy zoo 1: data release of morphological classifications for nearly 900 000 galaxies. **Monthly Notices of the Royal Astronomical Society**, Oxford University Press, v. 410, n. 1, p. 166–178, 2011.

MALONE, T.; CROWSTON, K. The interdisciplinary study of coordination. **ACM Comput. Surv.**, v. 26, n. 1, p. 87–119, 1994.

MANSKI, C. F. Identification of endogenous social effects: The reflection problem. **The review of economic studies**, Oxford University Press, v. 60, n. 3, p. 531–542, 1993.

MASON, W.; SURI, S. Conducting behavioral research on amazon's mechanical turk. **Behavior research methods**, Springer, v. 44, n. 1, p. 1–23, 2012.

NEWMAN, M. **Networks: An Introduction**. [S.l.]: Cambridge U. Press, 2010.

NOWAK, M. A. Five Rules for the Evolution of Cooperation. **Science**, v. 314, n. 5805, p. 1560–1563, 2006.

OHTSUKI, H. et al. A simple rule for the evolution of cooperation on graphs and social networks. **Nature**, n. 441, p. 502–5, 2006.

OSBORNE, M. J.; RUBINSTEIN, A. **A Course in Game Theory**. [S.l.]: MIT Press, 1994.

OSGUTHORPE, D. Ab initio protein folding. **Current Opinion in Structural Biology**, v. 10(2), p. 146–152, 2000.

PAUL, S. A.; HONG, L.; CHI, E. H. Who is authoritative? understanding reputation mechanisms in quora. **arXiv preprint arXiv:1204.3724**, 2012.

RICHERSON, P. J.; BOYD, R. **Not By Genes Alone: How Culture Transformed Human Evolution**. University Of Chicago Press, 2004. 342 p. ISBN 0226712842. Disponível em: <http://www.amazon.com/Not-Genes-Alone-Transformed-Evolution/dp/0226712842>.

ROGSTADIUS, J. et al. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In: **ICWSM**. [S.l.: s.n.], 2011.

SHAW, A. D.; HORTON, J. J.; CHEN, D. L. Designing incentives for inexpert human raters. In: ACM. **Proceedings of the ACM 2011 conference on Computer supported cooperative work**. [S.l.], 2011. p. 275–284.

SURI, S.; WATTS, D. J. Cooperation and contagion in web-based, networked public goods experiments. **ACM SIGecom Exchanges**, ACM, v. 10, n. 2, p. 3–8, 2011.

Turkopticon. 2015. Disponível em: <https://turkopticon.ucsd.edu/>.

WikipediaRanking. 2016. Disponível em: <http://www.alexa.com/siteinfo/wikipedia.org>.

WOOLLEY, A. W. et al. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. **Science**, v. 330, n. 6004, p. 686–688, 2010.

APPENDIX I: PUBLISHED AND SUBMITTED WORKS

AAAI 2010 - Combining Human Reasoning and Machine Computation (abstract)

Abstract: In this work we propose a framework where humans and computers can collaborate seamlessly to solve problems. We are currently developing and applying a network model, namely Memenets, where human knowledge and reasoning are combined with machine computation to achieve problem solving. The development of the Memenet is done in three steps: first, we simulate a machine-only network. Previous results show that the Memenet is an efficient problem solving model when compared to known search algorithms. In the second step, we perform an experiment with human agents organized in a virtual network. The investigation about human behavior while solving problems in a social network lead to postulate principles of communication in Memenets. These postulates describe an initial theory to enable human-computer interaction inside social networks. In the third stage, these principles combines human and computer computing to propose a complete Memenet based problem-solving model. (FARENZENA; LAMB; ARAUJO, 2010)

IJCAI 2011 - Towards Social Problem-Solving with Human Subjects (abstract)

Abstract: Recently, the use of social and human computing has witnessed increasing interest in the AI community. However, in order to harness the true potential of social computing, human subjects must play an active role in achieving computation in social networks and related media. Our work proposes an initial desiderata for effective social computing, drawing inspiration from artificial intelligence. Extensive experimentation reveals that several open issues and research questions have to be answered before the true potential of social and human computing is achieved. We, however, take a somewhat bold approach, by implementing a social networks environment where human subjects cooperate towards computational problem solving. In our social environment, human and artificial agents cooperate in their computation tasks, which may lead to a single problem-solving social network that potentially allows seamless cooperation among human and machine agents. (FARENZENA; ARAUJO; LAMB, 2011b)

SocialCom 2011 - Collaboration Emergence in Social Networks with Informational Natural Selection (full paper)

Abstract: Social collaboration can benefit individuals by avoiding efforts and risks inherent of trial-and-error learning. However, social collaboration may demand considerable effort and time. We present a new model of social collaboration based on Informational Natural Selection in order to investigate social problem-solving. We performed a set of social network experiments in which individuals solved problems in a virtual environment. Results show that collaboration can be viewed as a complex system's emergence promoted by a agent's behavior that results from Informational Natural Selection.(FARENZENA; ARAUJO; LAMB, 2011a)

First Monday 2016 - The cost of search and evaluation in online problem-solving social networks with financial and non-financial incentives (full paper submitted)

Abstract: Online networks of individuals have been used to solve a number of problems in a scale that would not be possible if not within a connected, virtual and social environment such as the Internet.

In this paper, we show that when solving tasks with small duration (under 5 minutes), also known as microtasks, individuals decision making will be strongly biased by costs of searching (and evaluating) options rather than financial or non-financial incentives. Indeed, we are able to show that we can influence individuals decisions, when solving problems, by rearranging elements visually to modify an individual search sequence, be it by designing the virtual work environment or manipulating which options are first shown in non-controlled environments such as the Amazon Mechanical Turk labor market.

We performed almost 50 experiments in online networks where individuals are invited to work on tasks with varying degrees of difficulty within three settings: mathematical games with objective truth (Sudoku and SAT instances), surveys with subjective evaluation (public policy polling) and labor markets (Amazon Mechanical Turk).

APPENDIX II: THE INTERLEAVING PROJECT

As part of a collaboration with Cornell University, we developed *The Interleaving Project*. The first goal of this project is to answer the following question: *Is it possible to improve the performance of individuals when solving a set of problems by re-ordering the set of problems according to their difficulty?* The inspiration for this effect is from the Flow theory found originally in Psychology studies (CSIKSZENTMIHALYI; CSIKZENTMIHALY, 1991).

By improving performance, we should be able to get higher quality work output from problem-solvers. Researchers have been investigating ways to improve AMT quality output by several means and our goal is to show that quality can be improved by re-ordering a set of problem instances (IPEIROTIS; PROVOST; WANG, 2010; MASON; SURI, 2012). This is a relevant question that has been raised before and with the availability of quick experimentation techniques, we are able to investigate several properties of online problem-solving tied to this question. Specifically, we wanted to investigate if, by interleaving problems with hard and easy difficulty, we could improve the quality of the work compared to other problem sets (e.g. all easy, all hard and others).

To answer the question, we setup a series of experiments involving around hundreds of individuals on Amazon Mechanical Turk (AMT). We posted several tasks on AMT where individuals were paid to count images of blood cells. The individual were unaware that the images were computer generated by our group and had an expectation that they were contributing to laboratory exams, which is an incentive that runs in parallel with the monetary incentive. A experimental task was performed in (ROGSTADIUS et al., 2011), however the focus of the study was not improving quality by reordering problem sets but looking to understand the differences between intrinsic and extrinsic motivation.

The interface of the task and the explanation given to the user show in Figure B.1.

Workers were given 10 images to tag, each one with a give difficulty level:

1. Easy: sparse cells; small number of cells;
2. Hard: dense, overlapping cells; high number of cells compared to the easy case;

For a set of 10 images to count cells, we will denote easy tasks as E_i and H_i where E denotes and easy problem and H a hard one, where i uniquely identifies the instance of a problem in the 10 problem instances set. For instance, for an all hard problem set we would denote $P = \{H_0, H_1, H_2, ..., H_9\}$, whereas a set with problems interleaved problems we would have $P = \{H_0, E_0, H_1, E_1, ..., E_4, H_4\}$.

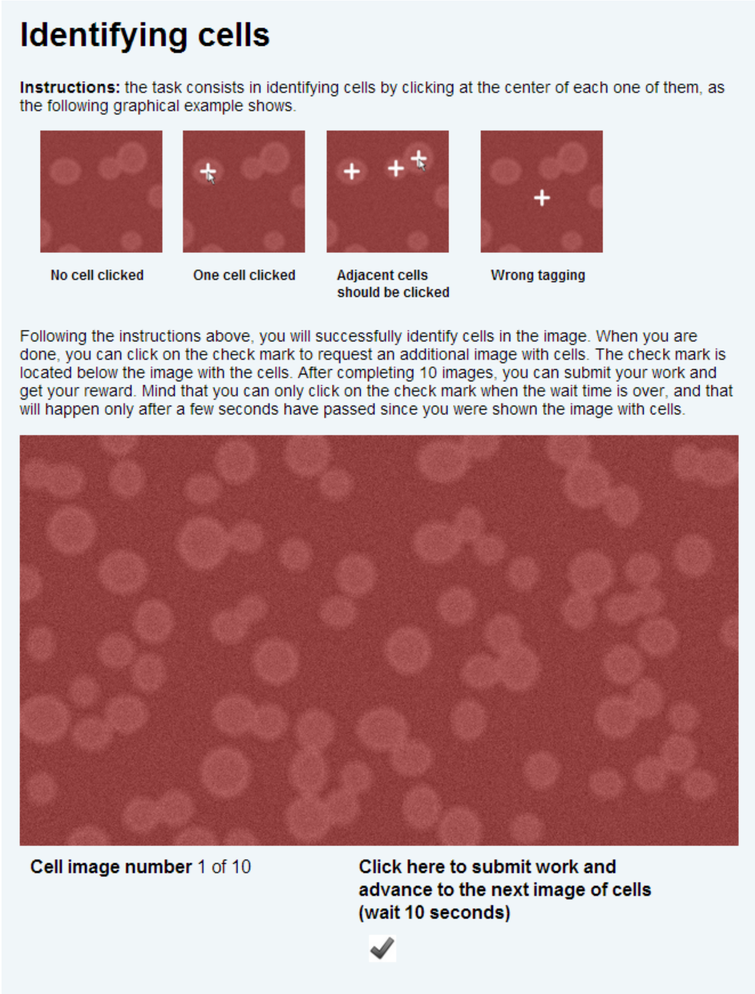


Figure B.1 – Interleaving Experiment interface

Workers were presented with examples on how to count cells with a general explanation of the activity, including that they should wait for a timer to time out and they can only participate in this task once. Allowing workers to participate once was required to make sure workers had the same experience level and did not learn by working on multiple tasks.

The experiments had four differents settings:

1. All hard: $\{H0, H1, ..., H9\}$
2. Interleaved: $\{H0, E0, ..., H4, E4\}$
3. One easy: $\{H0, E0, H1, H2, ..., H7, H8\}$
4. All easy: $\{E0, E1, ..., E8, E9\}$

We also we added a control (SHAW; HORTON; CHEN, 2011) to disallow workers to submit their work without waiting for at least few seconds, aiming to avoid having on the experiment individuals that shirk throughout all images. To measure the quality of the cell counting, we would give a score based on the proportion of cells that were correctly identified in the image.

The goal of having the images computer generated is that we wanted to know the exact number of cells on each image and being able to quantitatively vary the density and number of cells to have a fine grained control of the problem instance difficulty.

The experiment runs are described in Table B.1.

Table B.1 – Participation for the Interleaving Experiment

		All Hard†			Interleaved			Easy		
Exp.	Pay per HIT	Acc. HITs	Compl. HITs		Acc. HITs	Compl. HITs		Acc. HITs	Compl. HITs	
4	$0.10	181	92	50.8%	206	117	56.8%	-	-	-
5	$0.05	370	121	32.7%	351	142	40.4%	-	-	-
6	$0.01	649	353	54.3%	685	419	61.1%	-	-	-
7	$0.00	317	202	63.7%	331	216	65.2%	-	-	-
9‡	$0.05	449	236	52.5%	-	-	-	676	427	63.1%
10	$0.01	655	391	59.6%	676	427	63.1%	757	502	66.3%
11	$0.05	285	158	55.4%	-	-	-	364	242	66.4%
12	$0.10	118	80	67.8%	-	-	-	110	76	69.0%

†"Acc.": accepted, "Compl.": completed;
‡Only one easy task was inserted in this experiment.

Initially, our findings seemed to show that indeed the output quality is higher in experiment 4 (see Figure B.2).

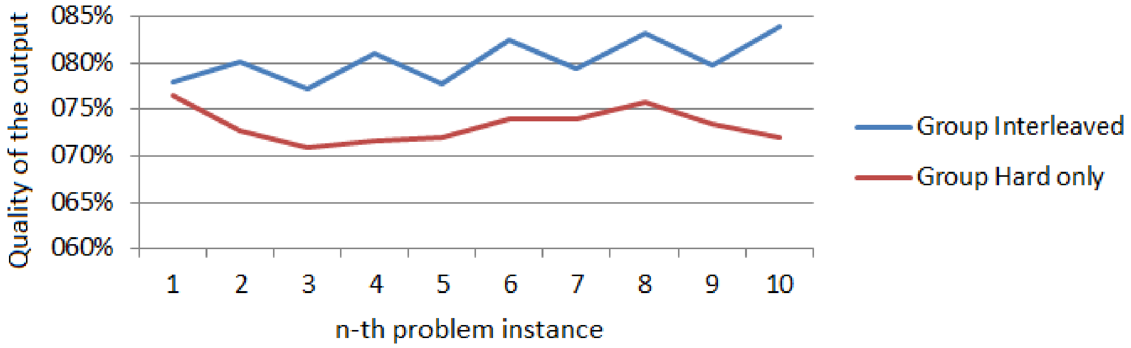


Figure B.2 – Average quality output of Experiment 4

Vertical axis shows quality output and horizontal axis the n-th problem from the 10 problems set the worker was paid to work. The interleaved group had an average increase in quality output whereas the hard only group had an inconsistent increase. To confirm workers started at the same level, we can see that in both problem sets they start with the same quality output, which is expected if they were solving the same problem.

To verify the validity of our experiments, we checked if the quality output from the first problem instance, which were the same for all problem sets (a H_0 cell image) were the same.

However, not all experiments were valid according to this criteria (see Figure B.3).

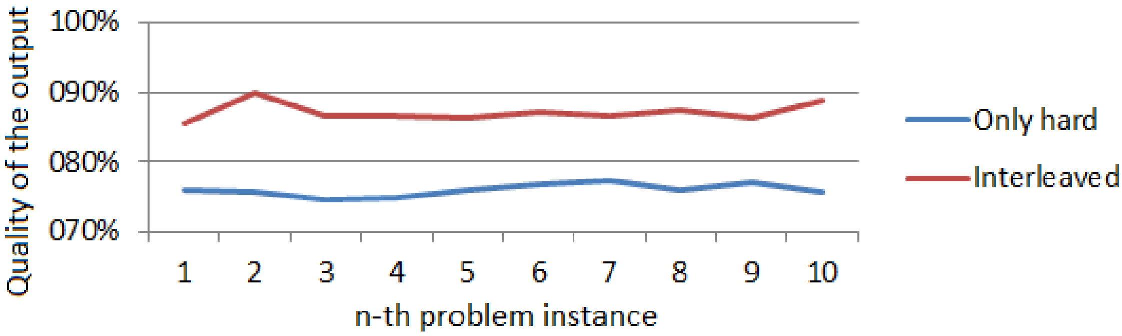


Figure B.3 – Average quality output of Experiment 6

In this case, however, the initial quality output was different, meaning that individuals were not actually behaving the same way since the beginning of the experiment. This was an indication that a different effect was in place. Lately, we explain that the results are not indeed due to interleaving, but as a side effect of *selection*.

By analysing the dropout rate along each problem, we were able to identify that, indeed, the quality output difference between problem instance difficulties could be better explained by a *selection* process. In this process, workers drop the experiment as they face E_i or H_i, leaving only workers with a specific profile of problem solving, a profile that would output the quality seen in Figures B.2 and B.3. Two dropout graphs can be seen in Figure B.4 and B.5.

We can state that our experiments show that quality output is increased in the interleaved case, however not due to the flow effect, but to selection. Therefore, we proposed a new method of increase quality by posing a specific set of problems at beginning of the task that will cause the selection of the workers that will generally provide higher quality output.

Whether due to selection or not, by varying the reward value from \$0.10 to \$0.00, we are able to show that by paying less money, only workers that have an intrinsic motivation are left to work on the task, and they are the workers that can output the highest quality we logged (see Figure B.6).

B.2 Evaluating workers quality output by task completion time

Requesters from Amazon Mechanical Turk (AMT) have a challenge when deciding whether or not to pay workers in the following sense: as discussed in Section 1, there are types of problems that are better solved by humans. AMT became popular by offering an automated way of hiring individuals to work on those types of problems. However, due to the very reason

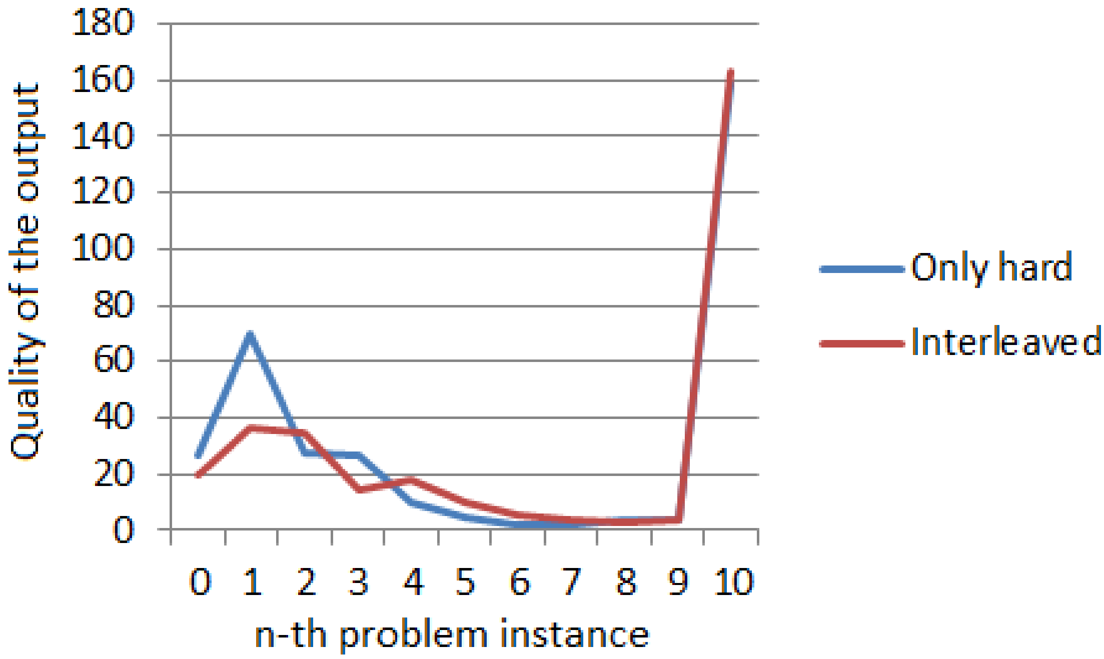


Figure B.4 – Dropout rate for Experiment 4

Vertical axis show the numbers of workers that completed the experiment up to the n-th problem (horizontal axis). For instance, the point (4,20) means that 20 workers completed 4 problem instances and then dropped from the task, thus opting to quit and get not reward instead of completing the problem. Therefore, at each problem, we filtered the workers and at the end we had only the workers that, on average, have the quality output profile seen in the quality output figures.

they hired individuals through AMT, requesters cannot automatically, by means of an automated system, evaluate if the worker provided any valid work at all. For instance, when answering subjective questions on AMT, a computer cannot tell if the answer makes sense or it is just a sentence made of aggregated random words generated by a computer. This is a problem because requesters might be getting poor quality answers without nowing or spending money on workers that do not provide good quality answers, increase the costs of AMT indirectly. The only way to qualify the answer of certainty is to have another individual to verify the answer, but this verification would also need human validation from an additional individual, *ad infinitum*.

Researchers have been looking for ways to indirectly qualify work from AMT. We propose a novel method of qualifying work from AMT without even accessing the contents of the task answer. After performing the experiments described in Section 9, we analyzed the distribution of work quality over time spent solving the problem to prove that there is a minimal amount of time that workers need to spend solving a problem so that the solution to the task has a minimal amount of quality (Figure B.7).

We propose that some tasks on AMT can have their quality assured with the following steps:

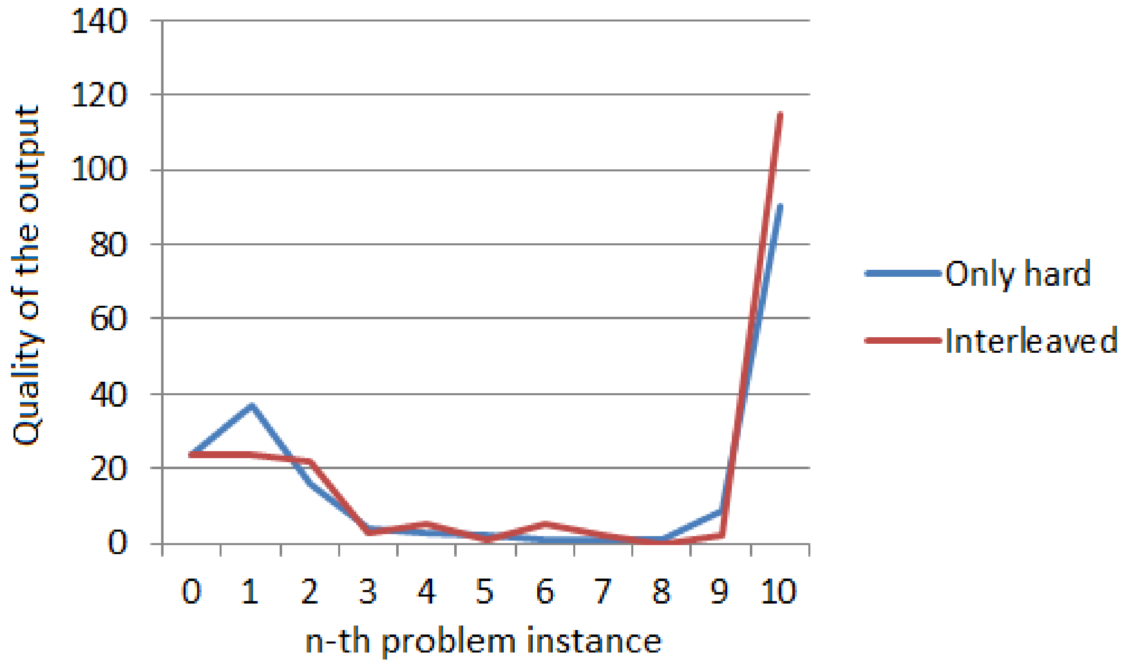


Figure B.5 – Dropout rate for Experiment 6

The dropout graphs are consistent with a behavior where individuals tend to either drop early or stay until the end.

1. Post an initial batch of tasks on AMT.
2. Evaluate, with the aid of a human, the quality output from each worked task to collect initial distribution parameters.
3. With the distribution in hand, the requester knows what is the minimal amount of time a worker will use to provide work with a quality threshold.
4. The task might be set, programatically, to inform workers that only work will only be accepted after a minimal amount of time as a form of control (SHAW; HORTON; CHEN, 2011). In this way, we let workers select if they want to work as much as specified, while keep quality of work high.

Not all tasks have clearly separable quality output by time. By not having a way to separated into two distributions, one increases the chance of denying payment for quality work. In this case, the requester is not being fair and can be flagged as bad requester and have his requester license dropped. See Figure B.8 for examples.

We showed a new method of evaluating quality output from workers on AMT. It cannot be applied in all cases and data must be collected to determined the distributions parameters before deciding whether the method is suitable for each specific HIT Group.

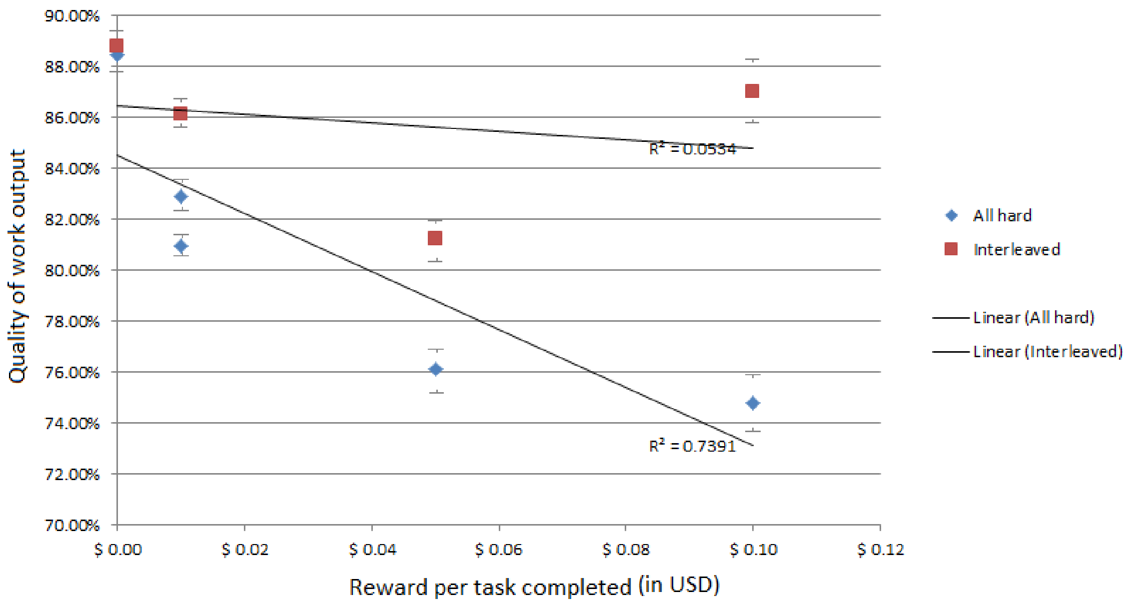


Figure B.6 – Quality Output versus Reward

In this graph, we show the correlation between Quality Output (vertical axis) and Reward (horizontal axis). At Reward \$0.00, the quality output was the highest due to the selection effect.

B.3 Amazon Mechanical Turk Economics

One of the main variables used to understand labor markets is the *pay rate*, or the amount of money a worker makes on a given period of time on average. In this chapter, we are going to investigate the importance of pay rate over uptake rate, i. e. the rate at which HITs are consumed or accepted and completed by workers. We have done a preliminary study that relates uptake to reward and time allotted, which are two values closely related to the pay rate.

Using the data we collected on our previous project (see table 5.4 for a summary), we present the correlation between Uptake Rate and Pay Rate on Figure B.9.

We do believe that pay rate has an influence on uptake rate. There are web sites specialized in extracting AMT information and publishing HITs by pay rate ((Turkopticon, 2015). Therefore, we will assume a more detailed exploration of the data is necessary.

To compute the pay rate, we assumed the following relation:

$$\text{Pay Rate} = \frac{\text{Reward}}{\text{Time Allotted}} \tag{B.1}$$

In reality, no worker will take exactly the time allotted to solve a task and our method of data collection by web scraping does not *allows* us to know exactly how much time workers

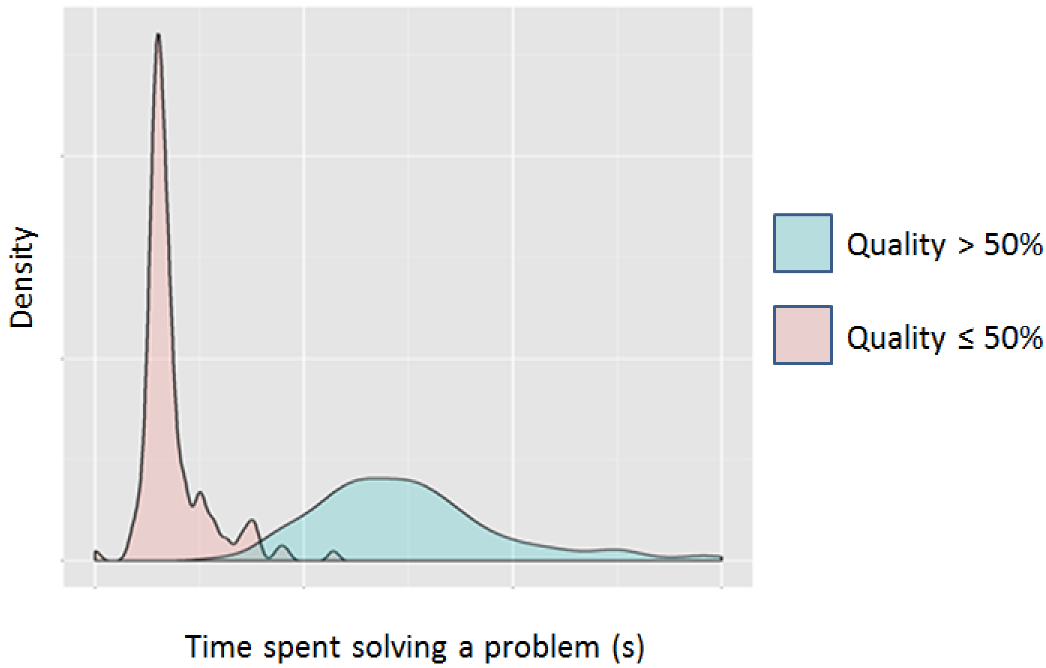


Figure B.7 – Relationship between time spent solving a problem and quality output

The figure shows two histograms each, where the bins represent time spent solving a given task.. The distribution with a high peak is the distribution of workers with quality output lower than 50%, whereas the other distribution shows the remaining output (>50%).

are spending to solve the tasks. However, more important then knowing what is their actual pay rate, is to know what is their *expected* pay rate, as this is how they evaluate whether to accept a task or not. This evaluation is partially done with the time allotted description, but more information is also available on the title and the description of the task (see Figure B.10).

As a solution, we added a Natural Language pre-processing step where we try to extract from the title and description of HITs more information about estimated payment and time resources consumed. There was no machine learning involved and we implemented a simple syntactic analyzer that would look for words like *money*, *seconds*, *minutes*, among others, to find sentences like *only 5 minutes* or *bonus of $1.00*. During this process, all time or financial information found on the title or description would replace the actual value of reward or time allotted in Equation B.1.

Processing the natural language descriptions of the HIT resulted in Figure B.11.

The identification of two groups provide new information about AMT. To identify what they mean, we hypothesized that it was due to the nature of the problem, this being classified according to the title and the description. We created a simple Bayesian classifier and added training data from a small sample we classified. The result classified grouped words by Reward, broadly,

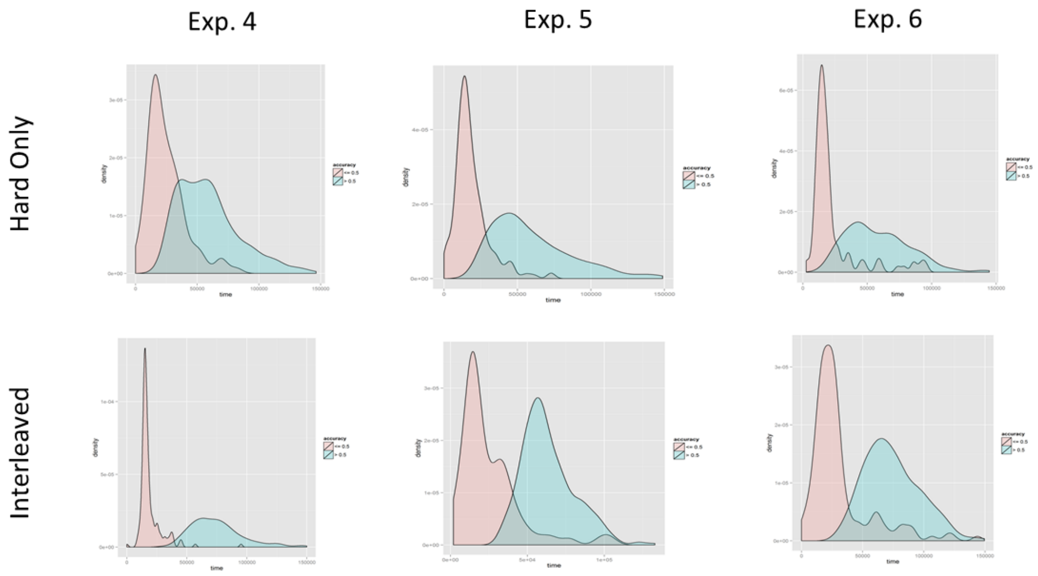


Figure B.8 – More examples of quality output histograms

All first four tasks can use our method to control the quality of provided work. The last two might cause good quality work to have payment denied and we should avoid to use the method.

into *Transcription, Video and Audio* and the rest of task types. By plotting the histogram of the pay rate, we are able to more clearly show that there are two groups of HITs (see Figure B.12).

Even knowing to which group a given HIT is contained, we are still, due to the variance present inside each group, unable to relate pay rate and uptake. Here, we will offer 4 explanations to the variance observed.

B.3.1 Gaming the market and detection

When requesters add HITs to inflate the total number of HITs available in their HIT group and reach the top position on AMT web site, we say that the requeter is *gaming the market*. This is an expression derived from Economics and can be used for others market as well to describe a situation where the participants of the market are acting strategically to manipulate the market.

It is possible to game AMT since after posting one or more HITs, it is also possible to remove those HITs without paying the reward value. No fee is charged from the requester from tasks that were not worked. Therefore, it is possible to keep adding HITs to a HIT Group until it reaches the first position on the AMT web site, which sorts HIT Groups by default according to HITs available.

Although we might assume that requesters do game the market, we cannot directly know

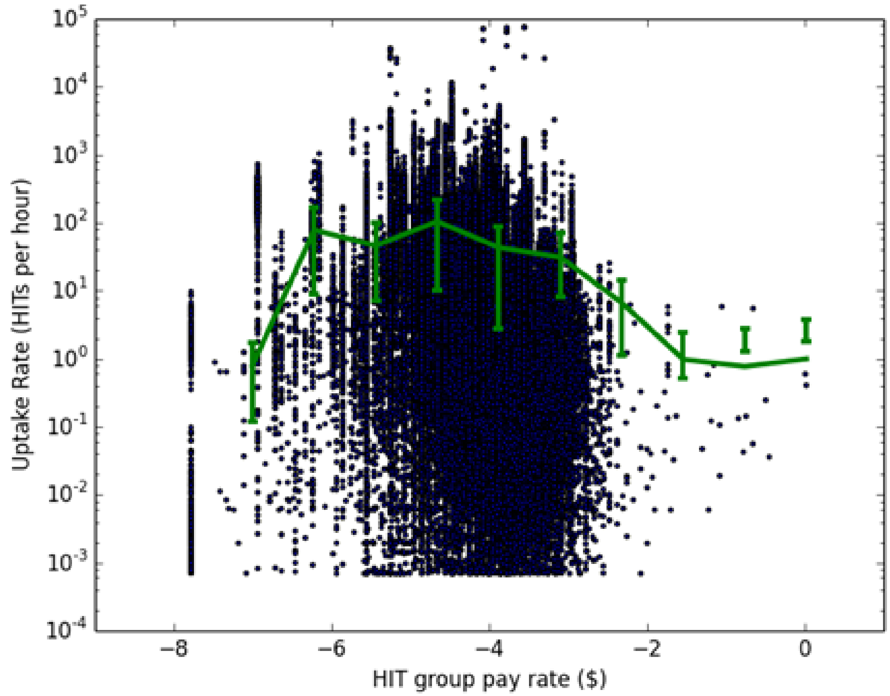


Figure B.9 – Pay Rate by Uptake Rate

Pay rate plotted against uptake rate on a logarithmic scale. As can be seen by the error bars, variance is too high and no conclusion can be drawn about pay rate influence on the market dynamic. However, this is in agreement with our previous linear regression analysis where R^2 was very low.

if they are gaming without direct access to AMT database, since externally we cannot tell if a HIT was consumed by a worker or removed by the requester. However, we devised a method that can possibly show when requesters are gaming the market, and therefore exclude that effect from our linear regression estimates.

Observing the first plot of Figure B.13, we can see the evolution of HITs available for a given HIT Group. Our expectation was that, as time progress, the uptake rate decreases, as the HIT Group is not anymore among the top 10 HIT Groups on the first page. However, what we see is that, after a certain threshold, the uptake increases and the HIT Group is removed completely from AMT. Looks like a secondary, discrete factor kicked off and started consuming the HITs. This is a pattern that can be seen in other HIT Groups, mainly the ones that are present around the first positions, which is the position where a gamer would like to manipulate the number of HITs.

We hypothesize that the abrupt difference change in uptake rate is due to a machine automatically removing HITs as fast as its infrastructure provides. To show the abrupt difference,

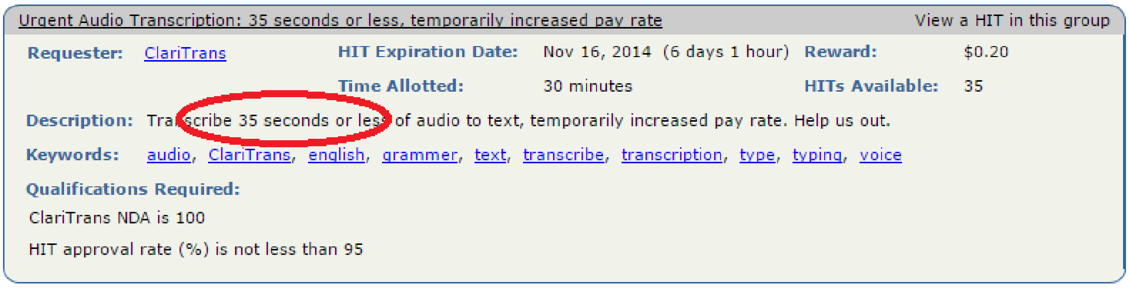


Figure B.10 – Time resources details on the description of a HIT

As highlighted with the circle, the requester provide a time estimate of how much time the worker is going to spend working on the task. Sometimes, HITs can contain bonuses as well, and thus the reward might be higher then indicated in the *Reward* field.

we plotted the histogram of uptakes and show that indeed there is more than one mode (see histogram in Figure B.13).

By having the parameters for the second normal distribution highlighted in Figure B.13, we should be able to remove the data represented by the distribution and thus remove the data collected and only present due to gaming, therefore reducing the error of our estimates.

B.3.2 Seasonal effects

Seasonality is known to affect precise AMT measurements and we have to take them into account as well.

B.3.3 Failure on Amazon Mechanical Turk Infrastructure

The large scale of managing 68 thousand HIT Groups simultaneously is not something simple to manage from the Information Technology point of view. Understandingly, with considerable frequency AMT infrastructure fails and both requesters and workers are unable to interact with AMT. Moreover, requesters sometimes host their own tasks and their infrastructure might fail as well.

During downtime, the collected data has a very non-linear nature and the errors introduced by downtime are hard to filter out of the analyzed dataset (See Figure B.15).

Differently from the other causes for variance, we did not currently applied any method to remove this source of error.

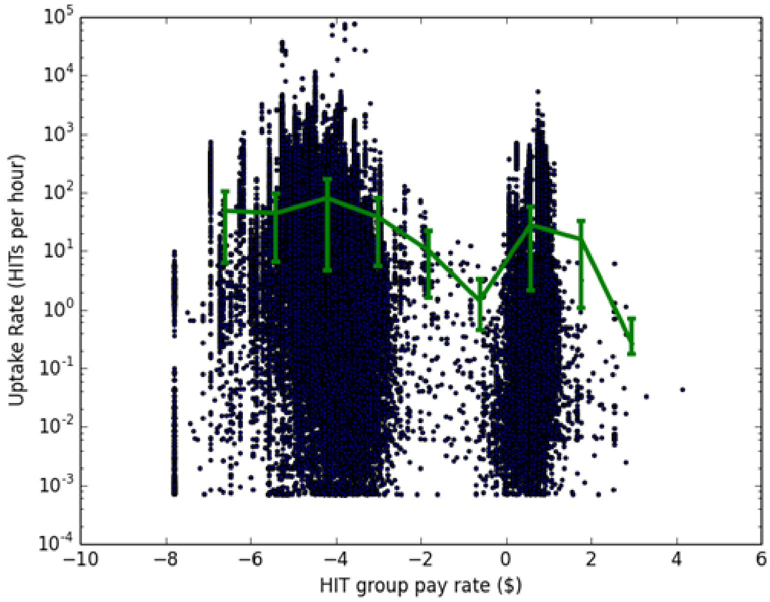


Figure B.11 – Pay Rate by Uptake Rate, with NLP

We processed the content of natural language descriptions on each HIT to provide new estimates about time and monetary information of the HIT. The error is still big, but evidence shows that indeed there are two big groups of HITs that need to be further analyzed.

B.3.4 Data Collection

When we compare the statistic of our results and analysis, we can identify important differences and we might ask ourselves if the dataset being analyzed is the same. With the motivation of showing that our dataset is complete and thus unbiased, we argue that our dataset presents better the error structure of AMT activity. A comparison of our dataset to a paper that performs a similar data collection can be found here 5.4 and proves that the referred dataset can have only 8% to 40% of the whole AMT market, thus is biased and might hide important sources of errors.

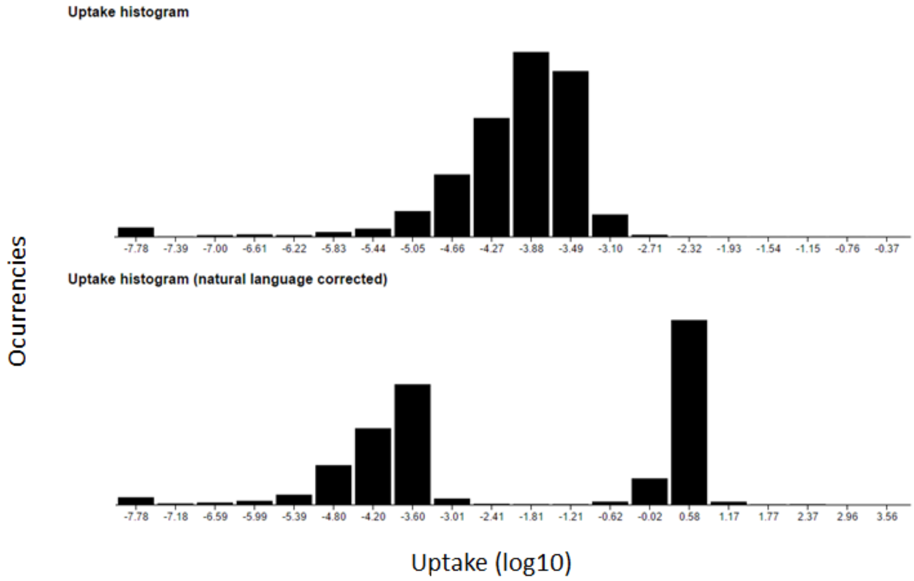


Figure B.12 – Uptake histogram

The uptake histogram with two groupings. The second graph shows the uptake histogram with Natural Language applied. Both graphs show the base 10 logarithm on the horizontal axis.

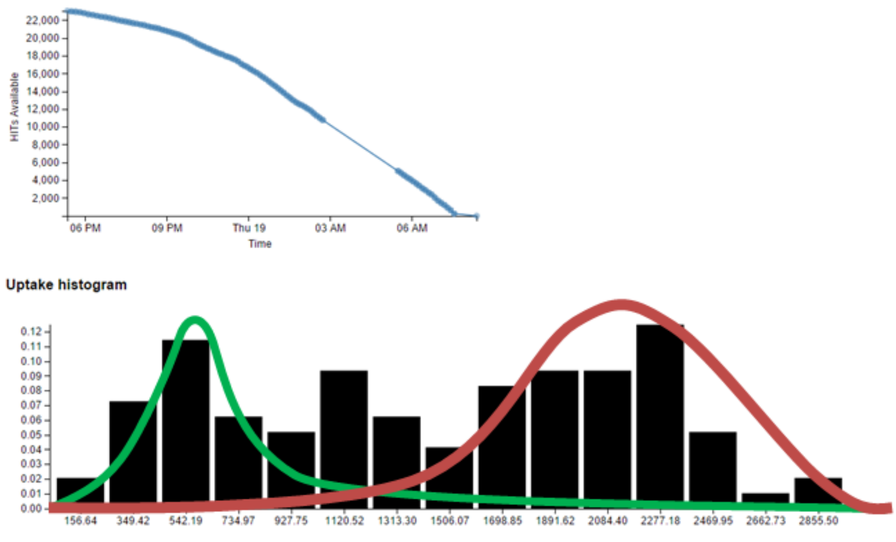


Figure B.13 – Gaming the market: number of HITs available and histogram of uptake rate

The number of HITs availabe from a HIT Groups from a top position is being shown here. The first graph shows the evolution of HITs for the HIT Group. We can see that around the second mark, a change on the uptake rate takes place. This can be confirmed by the histogram below, where we highlighted two modes with two normal distribution-like lines. The normal distribution that represents the highest uptake is what we hypothesize to be a requester gaming the market.

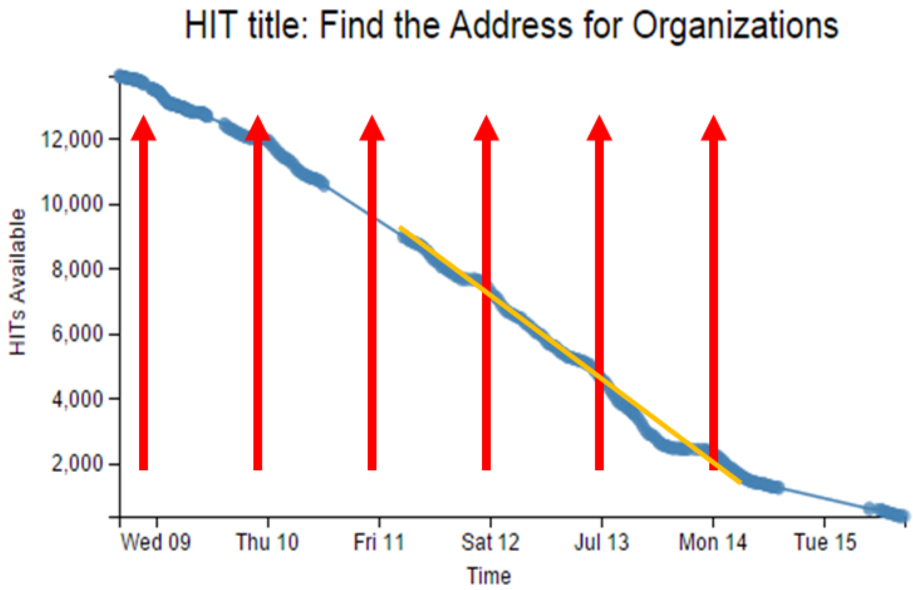


Figure B.14 – Daily seasonal effect

The upward arrows point the moment when uptake slows down due to a time in the day when workers are less active. We should point that 95% of the workers were divided evenly between United States and India, whose timezones are 12 hours apart.

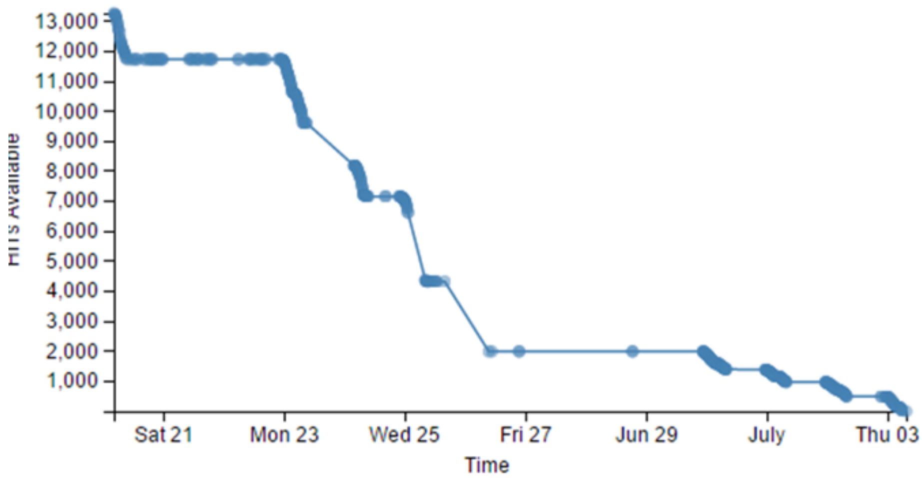


Figure B.15 – Information Technology (IT) infrastructure failure as a source of errors

Failure in computer infrastructure of AMT or the hosting requester can be an important source of errors. In this graph, we can see that the HIT count is non-linear and sometimes suddenly stops, indicating an infrastructure failure, often detected while performing experiments by not being able to communicate with AMT server.

APPENDIX III: RESUMO EM LÍNGUA PORTUGUESA

Nas últimas décadas, os humanos tem passado cada vez mais tempo em ambientes online e conectados, como por exemplo redes sociais online, web sites de trabalho colaborativos ou jogos multiplayer (DAVID; JON, 2010).

Portanto, considerando a quantidade e relevância das atividades que humanos realizam em ambientes online, é necessário desenvolver sistemas de informação que sejam capazes de agregar os recursos humanos disponíveis online. Porém, há bons modelos que explicam como humanos se comportam em ambientes online? Quais são relevantes fatores a considerar quando desenvolvendo sistemas em que humanos interagem e trabalham? A presente tese de doutorado investiga os importantes fatores que influenciam o processo de decisão humano em ambientes online e que podem ser usados para explicar o seu comportamento e possivelmente regras para desenvolver sistemas de informação que susteam as redes online.

Considerando a mudança de comportamento do mundo offline para o mundo online em um ambiente conectado, provido por uma imensa infraestrutura composta de dispositivos eletrônicos conectados (ABOWD; MYNATT, 2000), dados de atividade humana estão sendo coletados e tornando-se disponível para os detentores das redes, tornando pouco custoso armazenar e analisar informacões sobre comportamento humano em larga escala. Na presente tese, como uma metodologia para alcançar o nosso objetivo de entender o processo de tomada de decisão em ambientes online, nós realizamos diversos experimentos e coletamos dados em ambientes não controlados para prover evidências empíricas acerca de como humanos se comportam em ambientes online.

Neste trabalho, nós investigamos o comportamento humano em ambientes online da nossa perspectiva, que é utilizando *Redes Meméticas Artificiais* (AMN). O modelo AMN foi originalmente inspirado pelo modelo de transmissão, influência e assimulação de idéias presente em (DAWKINS, 1990) e aplicado com sucesso em simulações usadas para investigar a performance na solução de problemas em redes de agentes artificiais. O modelo proposto na presente tese modificará AMN para desenvolver, realizar e analisar experimentos com indivíduos resolvendo problemas em redes conectadas.

B.4 Introdução à trabalhos realizados na área

A performance do trabalho cooperativo humano foi extensivamente estudado dentro de diveros contextos. Porém, estudos utilizando métodos comumemente encontrados na ciência da computação não eram frequentes antes da existência de diversas aplicações online, cuja importância é baseada no trabalho de milhões de usuários coordenados por uma rica infraestrutura eletrônica de comunicação. (BRAN et al., 2009).

Diversos papers têm sido publicado, nos últimos anos, mostrando novas e rápidas formas de processamento de imagem utilizando unidades de processamento gráfico (GPU) que, com o crescimento econômico da indústrica de jogos, se desenvolveu de simples processadores com um único núcleo até processadores com centenas de núcleos. Porém, a qualidade das imagens processadas ainda assim era inferior à qualidade de uma rede de humanos fazendo a classificação de imagens manualmente por inspeção visual, através do aplicativo reCAPTCHA ((AHN et al., 2008)).

A teoria que explica o porquê do sucesso do reCAPTCHA pode ser encontrada em um artigo publicado pelos autores do aplicativo (AHN MANUEL BLUM; LANGFORD, 2003). Porém, a teoria não demonstra como pode ser aplicada no desenvolvimento de novas aplicações, perdendo, desta forma, generalidade.

Outra aplicação de sucesso é o Wikipedia, uma enciclopedia cujos artigos podem receber contribuições de qualquer pessoa com acessoa a internet. O Wikipedia é o sétimo web site mais acessado do mundo (WikipediaRanking, 2016) e contém informações tão confiáveis quanto enciclopédias tradicionais (GILES, 2005).

Um tratamento matemático rigoroso de comportamento em grupo pode ser encontrado, originalmente, em Economia e, posteriormente, Teoria de Jogos Algorítmica (AGT). Em AGT, assume-se certas condições para a função de payoff de usuários e restringir suas ações para que possamos descrever o sistema composto pelo grupo como um modelo matemático que possua soluções analíticas ou numéricas conhecidas.

Há casos observados em nossos experimentos que não estão previstos em alguns trabalhos que tentam, através de AGT, modelar aplicações online (GHOSH; HUMMEL, 2014)(JAIN; PARKES, 2013), como demonstraremos ao descrever os resultados dos nossos experimentos.

Uma abordagem recente para estudar fatores que podem influenciar o comportamento online através de como os membros e uma rede se interconectam pode ser encontrado em (DAVID; JON, 2010) e ficou conhecido por identificar redes de escala livre naturais. (BARABáSI, 2002)

Outro trabalho que merece destaque e foi utilizado extensivamente como referência na construção dos nossos experimentos é (KEARNS; SURI; MONTFORT, 2006). Neste trabalho, o autor propõe um experimento em que indivíduos são solicitados a resolver colaborativamente o problema de coloração de grafos. Cada invidíduo representa uma cor do grafo e é conectado através de uma rede social virtual implementada no laboratório da universidade. Todos indivíduos podem mudar a sua própria cor voluntariamente para que os vizinhos aos quais está conectado possam observar, mudar de cor e, em conjunto, resolver o problema da coloração de grafos. Este é um exemplo de trabalho em que incentivo financeiro é provido. Relatórios escritos por participantes indicam que, mesmo que a única ação que um indivíduo possa exercer na rede é mudar de cor, alguns indivíduos estabeleceram um protocolo de mudança de cor em função da tempo para transmitir outros símbolos além das cores possíveis. Entretanto, a metodologia para esta análise não foi descrita, assim como uma justificava para o método escolhido.

Por volta de 2011, Duncan Watts and Siddhartha Suri publicaram um artigo em que eles realizam experimentos em uma rede social e confirmam resultados experimentais que também encontramos e apresentaremos mais adiante no texto. Os resultados indicam que a topologia da rede não era relevante para análise do comportamento em questão (SURI; WATTS, 2011). O artigo também argumenta que os resultados experimentais se aproximam da previsão teórica apenas quando os jogos que os indívudos no experimentam jogavam se aproxima do fim, momento este conhecido como *end game*. Os experimentos realizados no trabalho citado foram em torno de 50 e, assim como alguns de nossos experimentos, realizados em laboratórios da universidade.

B.5 Redes Meméticas Artificiais

No nosso modelo, denotaremos cada agente como a_i e o m-ésimo grupo de agentes como $A_m = \{a_1, a_2, ...\}$. Quando um índice não for especificado, A representará o conjunto de todos os agentes no contexto em questão (Internet, rede social, etc.).

Cada agente a_i possui *características culturais* que são (BIRUKOU et al., 2013)

> characteristics of human societies that are potentially transmitted by non-genetic means and can be owned by an agent

ou, em tradução livre:

> características de sociedades de humanos que são potencialmente transmitidos de forma não genética e podem ser possuídas por um agente

As características culturais compartilham, dentre divesas definições existentes, uma característica em comum: podem ser transmitidas e aprendidas por indivíduos(BIRUKOU et al., 2013).

Definition B.5.1. *O agente a_i possui zero ou mais características culturais τ_j e o conjunto de características a_i é T_{a_i}, em si contido no conjunto $T_{a_i} \subseteq T$. Características culturais podem* *ser aprendidas e compartilhadas por agentes. Agentes podem compartilhar características que nunca foram aprendidas, ou seja, agentes podem criar novas características culturais e compartilhá-las:*

$$T_{a_i} : t \rightarrow T_{a_i,t} \tag{B.2}$$

A função acima definida é utilizada para mapear o conjunto de características culturais que um determinado agente possui em um determinado momento.

No nosso modelo, T é um conjunto de elementos totalmente ordenado que pode ser indexado unicamente com j e $j \in \mathbb{N}_+$. A definição de cultura de um determinado grupo $A_m = \{a_1, a_2, ...\}$ é dado por $C_{A_m} = \{T_{a_1} \cup T_{a_2} \cup ...\}$.

Definition B.5.2. (Definição de cultura) *A cultura C_{A_m} do m-ésimo grupo de agentes A_m é*

$$C_{A_m} = \bigcup_{i=1}^{|A_m|} T_{a_i}. \tag{B.3}$$

Se o grupo m possui apenas o agente a_1 então a cultura do grupo é a cultura do agente $C_{A_m} = C_{a_1} = T_1$.

Definition B.5.3. (Evolução da cultura) *A evolução da cultura (ou diferença) da cultura de um grupo A_m do tempo t' até t'' é*

$$\Delta_{t',t''} C_{A_m} = C_{A_m}(t') \cap C_{A_m}(t''). \tag{B.4}$$

Em geral, estaremos interessados na evolução da cultura entre o início e o fim da atividade de um agente resolvendo um problema p_k. Se t' e t'' são momentos quando a instância de problema p_k começam e terminam, respectivamente:

$$\Delta C_{A_m,p_k}(t) = \begin{cases} \emptyset & \text{if } t \leq t' \\ \Delta_{t',t'+t} C_{A_m} & \text{se} \\ \Delta_{t',t''} C_{A_m} & \text{se } t \geq t'' \end{cases} \tag{B.5}$$

No nosso trabalho, características culturais estarão limitadas a constuir solução de problemas, ou seja, cada característica cultural é uma solução para um SAT ou para um Sudoku, naqueles experimentos. Intercambiaremos o termo *característica cultural* e meme no resto do texto.

Apresenaremos, agora, uma descrição da evolução da cultura em termos de teoria de jogos.

Cada agente a_i escolhe uma ação $\alpha_i(t) \in \{E, I, N\}$ estrategicamente, em que $\alpha_i = E$ indica que o agente i escolheu copiar o meme de um agente. Neste caso, podemos especificar que a' copiou o meme na ação i $\alpha_i = E(a')$ notation. Adicionalmente, $\alpha_i = I$ indica que i agiu introvertidamente e modificou T_i em um novo meme(ARAUJO; LAMB, 2008). Por fim, $\alpha_i = N$ indica que o agente i não tomou ação.

A estratégia do agente i é denotada por $s_i = \{\alpha_i(t_0), \alpha_i(t_1), ..., \alpha_i(t_{p_k})\}$ ou s_{i,p_k} quando especificamente se refere à estratégia utilizada pelo agente i enquanto resolve a k-ésima instânica de problema p_k. As estratégias serão denotadas por $S_{p_k} = \{s_0, s_1, ...\}$ para a k-ésima instância de problema p_k. Já que o nosso jogo é finito (possui número discreto de passos e é limitado por tempo), nós definiremos $s_i(t'') = s_i(t') \cup \{\alpha_i(t')\}$ quando cada agente tomar uma decisão (ação) individualmente. Cada próximo passo é indexado sequencialmente, ou seja, $t_1, t_2, ..., t_{p_k}$ e t_{p_k} é a duração do problema p_k.

Através da ação $\alpha_i(t') = I$, o agente a_i gera um meme τ' baseado em outros memes em T_i. A este processo chamamos de *geração* de meme. Nós nomeamos a função generativa da seguinte forma:

Definition B.5.4. *O agente a_i pode gerar memes com G_{a_i}:*

$$G_{a_i,\alpha_i,t} : (T_{a_i}, T_{a_{-1}}) \rightarrow \tau_t \tag{B.6}$$

e

$$\tau_t : \begin{cases} \tau_t \in T_{a_i} & \textit{se } \alpha_t = N \\ \tau_t \in T_{a_{-i}} & \textit{se } \alpha_t = E \\ \tau_t \in T & \textit{se } \alpha_t = I \end{cases} \tag{B.7}$$

Portanto, o processo de geração de memes pode gerar memes que não estavam previamente em T_{a_i} e $T_{a_{-i}}$.

O processo de geração de um meme é importante para a função de ganho (e custo). É nesta função em que agentes tomam decisões e têm de decidir entre copiar, mutar ou não fazer nada.

Agora, conectamos o modelo até aqui apresentado com o modelo em (ARAUJO; LAMB, 2008). De (ARAUJO; LAMB, 2008):

Definition B.5.5. Agregação através da cópia de outros agentes*: A é o conjunto de vértices adjacentes; u = argmax$_{x \in A}$eval(x). Se mais de um vértice em A satisfaz esta condição, u é escolhido aleatoriamente entre estes. Então, $v \Leftarrow u$, do contrário, v é mantido inalterado.*

Para traduzi B.5.5 para o nosso modelo, temos que $u = T_{a_{-i}}$ para agentes conectados, os agentes que estão conectados a a_i.

A definição B.5.5 afirma que agentes atuam racionalmente através da cópia da *melhor* solução de acordo com a função *eval*, exatamente como o modelo teórico.

Há mais passos importantes para entender o funcionamento de Redes Meméticas Artificiais. Porém, considerando os expeirmentos que planjemaos, é suficiente introduzir a definição acima, já que outros passos, como o *passo de conexão*, *passo de apropriação* e o *passo de agregação* estão fora do escopo deste trabalho.

Se $\alpha_i(t') = I$ e o conjunto de memes $|T_i(t')| = |T_i(t'+1)|$, $\forall t' < t'', |s_i(t')| < |s_i(t'')|$ ainda é válido e $|s_i(t)|$ é monotonicamente crescente e $s_i(t') = \alpha_i(t')$.

B.6 Objetivos e métodos

B.6.1 Objetivos

1. Modelar o comportamento humano através de AMN (AMN).
2. Identificar fatores que influenciam no comportamento humano em ambientes online utilizando AMN e realizar experimentos que possam prover evidências empíricas para embasar o modelo.
3. Aplicar o modelo, agora com os parâmetros definidos através de experimentos, em um novo experimento, confirmando as previsões.

Portanto, o nosso primeiro objetivo é prover fundamentação matemática para AMN no contexto de redes sociais de solução de problemas, uma fundamentação que possa ser usada e comparada com outros modelos e experimentos possam ser desenvolvidos e testados. Acreditamos que o nosso model é capaz de realizar descrições comportamentais mais detalhadas e mostrar em que ponto as teorias divergem de observações e quais fenômeons podem ser atribuidos á interação social. A metodologia precisa atender a um requisito que mostramos anteriormente: experimentos de baixo custo que possam motivar indivíduos a interagir e que estes experimentos possam gerar grandes quantidades de dados.

B.6.2 Metodologia

Para verificar o nosso modelo, baseado em AMN, vamos realizar experimentos em ambientes online em três diferentes configurações:

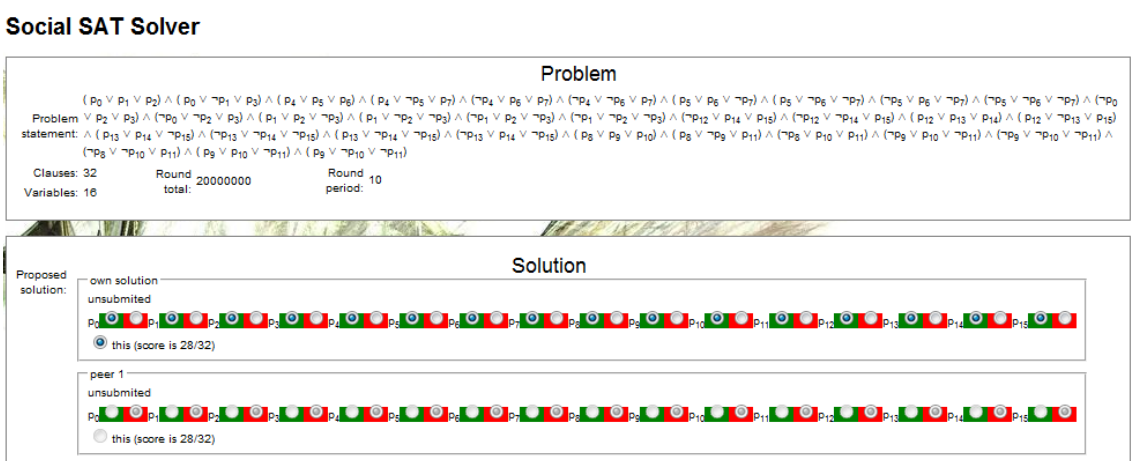
1. Fase de experimentos I: analisaremos fatores que são importantes em problemas de verdade objetiva.
2. Fase de experimentos II: nós estenderemos os resultados encontrados durante a Fase I para problemas de verdade subjetiva.
3. Fase de experimentos III: realizaremos experimentos que motivem financeiramente usuários e generalizaremos nossos resultados para o caso de incentivo financeiro.

B.6.3 Fase de experimentos I: experimentos em rede social

Para a Fase I, planejamos expeirmentos em que agentes humanos podem se comunicar em um ambiente controlado e são solicitados a resolver problemas matemáticos. Um primeiro grupo resolveu instâncias do problema SAT enquanto um segundo grupo resolveu instâncias do jogo Sudoku. Adicionamos o experimento com o Sudoku pois, ao contrário do SAT, já é sabido que existe uma motivação natural para jogar Sudoku, enquanto para SAT não há evidências.

Ambos os problemas vão ser estudados variando-se o número de agentes e topologia da rede e realizados nos laboratórios da universidade. Apenas a soluções de usuários conectados entre si, de acordo com a topologia, podem ser vistas pelos usuários vizinhos. Novas conexões não são permitidas depois que o experimento inicia. Em cada interface, um botão permitirá que uma solução seja copiada instantaneamente.

Figure B.16 – Interface de usuário no experimento envolvendo SAT



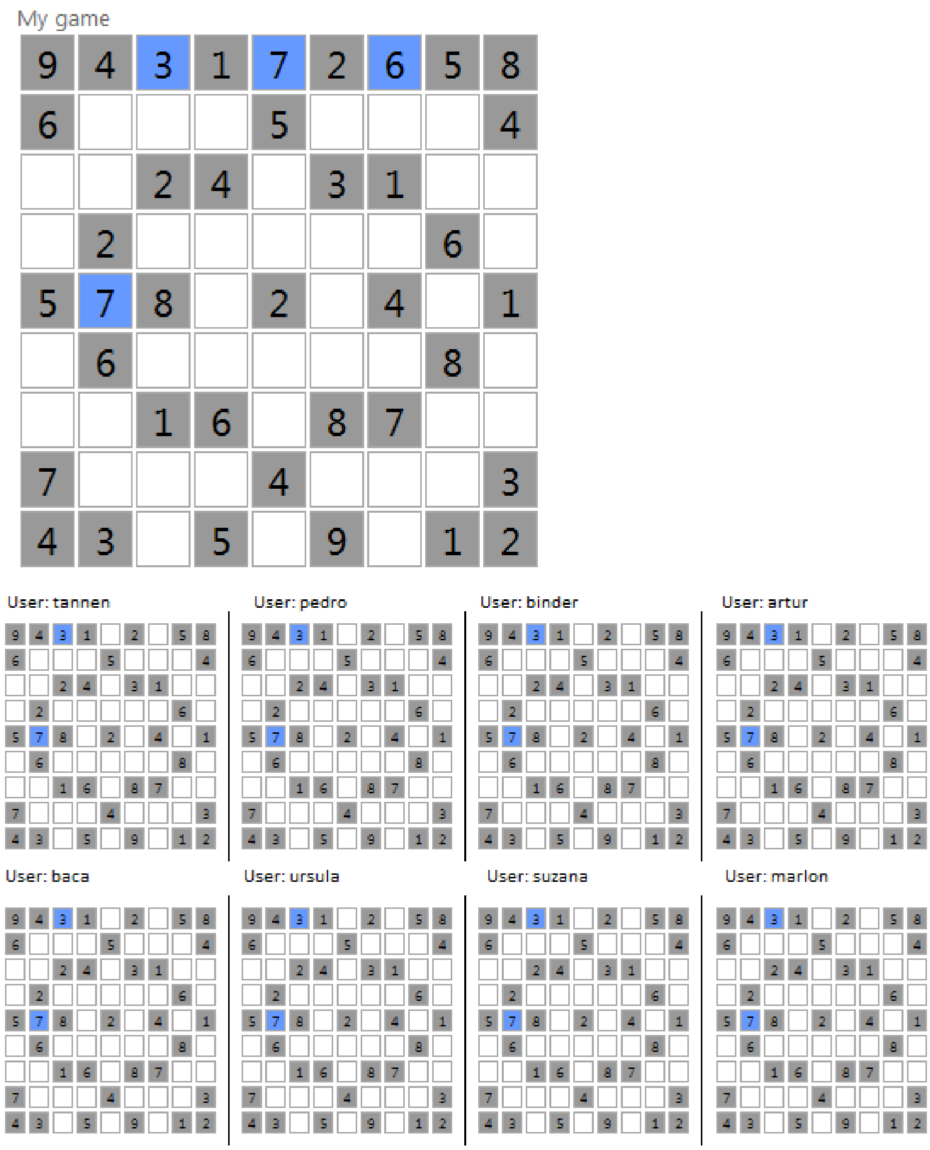


Figure B.17 – Interface de usuário no experimento envolvendo Sudoku

Como topologias, utilizamos redes em anel com tamanhos de vizinhança 4 e 6 (NEWMAN, 2010) e redes de escala livre(BARABáSI, 2002).

Durante os experimentos com SAT, nós explicamos o problema SAT, sua relevância para a ciência da computação e como os indivíduos devem utilizar a interface do aplicativo. Instruções sobre como visualizar a solução de seus vizinhos foram apresentadas também. Para motivar os indivíduos, anunciamos uma premiação para os melhores colocados.

B.6.4 Fase de experimentos II: votação de política pública

Nosso objetivo, nesta fase, é verificar se as conclusões da Fase I se mantém mesmo no contexto de problemas subjetivos. Especificamente, no caso de votações de política pública.

Nesta fase, compartilhamos uma página no Facebook em que pessoas poderiam votar e explorar opções de preços para passagens de ônibus em Porto Alegre (estado do Rio Grande do Sul, Brasil) B.18.

Figure B.18 – Interface de usuário do experimento de valores de passagem de ônibus

O valor total da passagem de ônibus é mostrado na parte inferior da página. A página, junto com o valor votado, podem ser compartilhados no Facebook.

Todos os clicks na página foram registrados em um banco de dados, assim como informação geográfica. Nossa hipótese é de que os usuários irão clicar nas opções mais próximas ao topo da página, evitando custos de analisar todas opções. As opções, para cada usuário, foi ordenadada de forma randomizada e individual para cada usuário.

B.6.5 Fase de experimentos III: mercado de trabalho online

Nosso objetivo com os próximos experimentos é verificar se nossos resultados anteriores podem ser aplicados para o caso de problemas com incentivo financeiro, através do mercado de trabalho online Amazon Mechanical Turk (AMT). Estes experimentos foram realizados em conjunto com a Cornell University.

AMT é um web site em que *requesters* oferecem uma recompensa financeira para a realização de tarefas, como um mercado de trabalho, porém de tarefas curtas como as microtasks. As tarefas são denominadas HIT (Human Intelligence Task) e qualquer pessoa, observando-se as políticas do site, podem aceitar e trabalhar na tarefa. Caso a tarefa não tenha sido realizada, o requester pode negar o pagamento. Antes de aceitar uma tarefa, um trabalhador pode ler a descrição da tarefa, como mostrado em B.19.

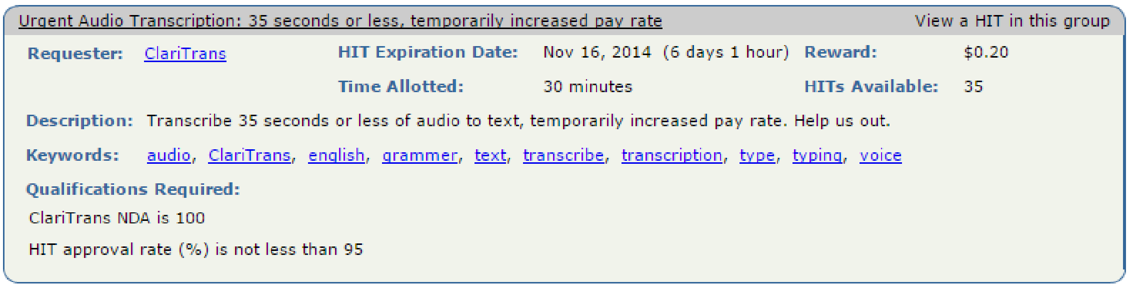


Figure B.19 – Detalhes de um HIT no AMT.

Os detalhes incluem, principalmente, uma descrição (*Description*), título (*Title*), recompensa (*Reward*) e duração estimada (*allotted time*).

Ao acessar o web site do AMT, o trabalhador (aquele que escolhe, aceita e realiza tarefas) é apresentado a uma lista de tarefas (Figura B.20).

Coletamos, periodicamente, o número de HITs de cada grupo de HIT de todas as tarefas disponíveis no site. B.20. O número de HITs consumidos por unidade de tempo seria um indicativo do interesse que determinados usuários tem por um determinado tipo de tarefa.

B.7 Resultados

Nossa análise mostra que há um comportamento padrão de agregação que pode ser identificado em todos os experimentos. Para nossa surpresa, agentes davam mais importância às primeiras soluções que pudessem ser copiadas, mais importância que o valor da solução para resolver o problema. Portanto, existia uma chance maior de uma solução ser copiada se estivesse

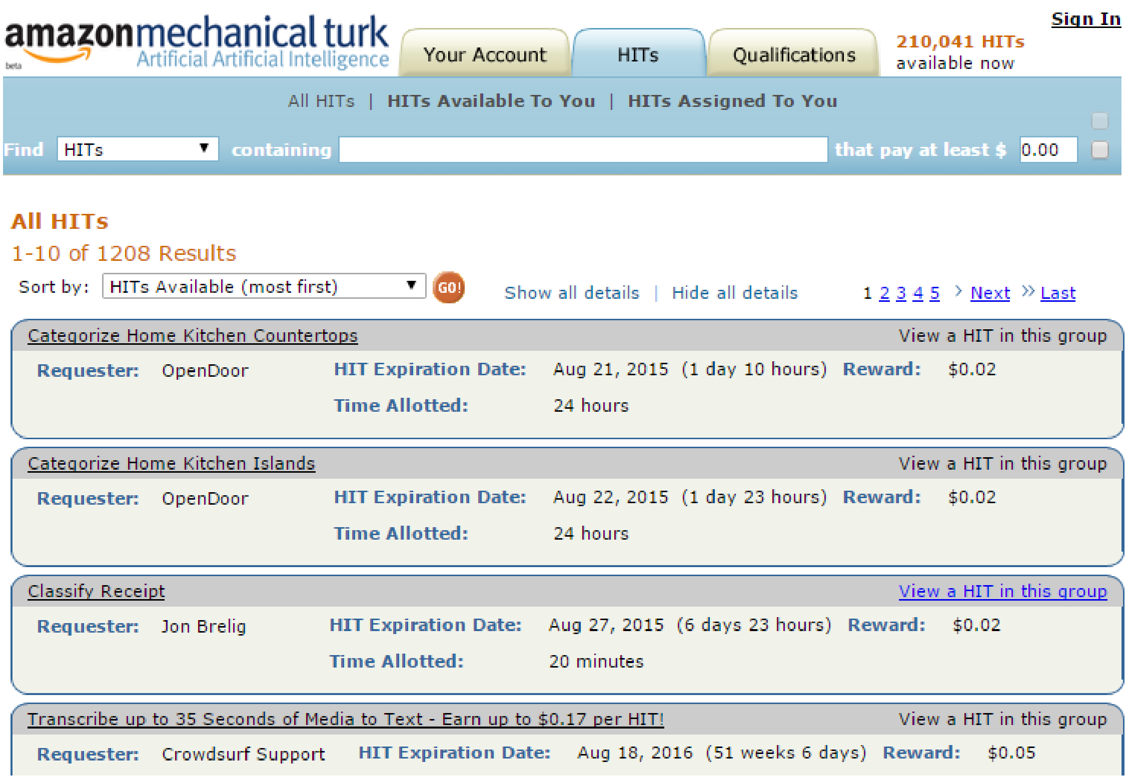


Figure B.20 – Interface de usuário do AMT

A tela inicial mostra as tarefas por ordem de número de HITs disponíveis para cada grupo de HITs.

em primeiro, mesmo que não fosse a melhor solução. A conclusão e análise detalhada foram publicados em (FARENZENA; LAMB; ARAUJO, 2010),(FARENZENA; ARAUJO; LAMB, 2011b) e (FARENZENA; ARAUJO; LAMB, 2011a).

B.7.1 Resultados da fase I

Durante a Fase I, publicado em (FARENZENA; ARAUJO; LAMB, 2011a), executamos 42 experimentos envolvendo 160 indivíduos. Para o experimento SAT, a interface é apresentada em linhas, enquanto para o Sudoku era apresentado em uma tabela de soluções (Fig. B.17).

Os experimentos SAT foram realizados em 30 sessões de experimentos, em que cada sessão os agentes resolviam diversas instâncias de problemas, sessões estas sumarizadas na tabela B.2. Das 30 sessões planejadas, apenas 14 foram agendadas e 8 obtiveram sucesso na realização. Algumas sessões não foram executadas por sobreposição com atividades acadêmicas, como provas e feiras.

Os dados coletados estãos sumarizados nas tabelas B.3) e B.4).

Nossa hipótese é a de que a posição inicial dificilmente influenciaria na decisão dos agentes,

Table B.2 – Plano para os experimentos SAT e Sudoku

Date	*Time*	*Type*	*Result*	*Comment*
9/23/2009	10:30	Individual	Negative	Falha no servidor.
9/30/2009	10:30	Individual	Positive	Instâncias SAT muito pequenas.
10/6/2009	15:30	Individual	Positive	Questionário para indivíduos.
10/7/2009	10:30	Individual	Negative	
10/8/2009	15:30	Individual	Negative	Faltaram instâncias adicionais.
11/9/2009	8:30	Individual	Positive	Problemas com rede (hardware).
11/9/2009	10:30	Individual	Positive	
11/10/2009	15:30	Em grupo	Negativo	Software bug.
11/11/2009	8:30	Em grupo	Negativo	Software bug.
11/11/2009	10:30	Em grupo	Negativo	Software bug
11/12/2009	15:30	Em grupo	Positivo	Anel N=26; $K_{ring} = 4$ and $K_{ring} = 6$.
11/18/2009	10:30	Em grupo	Positivo	
12/8/2009	10:30	Em grupo	Positivo	Software bug, mas funcionou.
12/14/2009	10:30	Em grupo	Positivo	$K_{ring} = 4$.

Resultados positivos expressam que os dados foram coletados com sucesso e que eles possuem qualidade. Resultados negativos indicam que os dados tiveram de ser descartados. Entre falhas comuns, tivemos bugs no software, falhas de hardware, dados corrompidos ou incosistentes ou parâmetros de testes incorretos.

considerando que agentes prefeririam copiar as melhores soluções com o objetivo de resolver o problema com mais eficácia. Para verificar esta hipótese, plotamos o número de vezes que uma solução vizinha foi copiada em função da sua posição na tela para o problema SAT (Fig. B.21). Foi possível identificar que a probabilidade da k-ésima solução ser copiada era maior. Ajustando-se a curva como uma funçõa massa de probabilidade da distribuição geométrica em que $\langle X(k)\rangle$ denota a probabilidade de um agente copiar a k-ésima solução, com parâmetro $p = 0.5479$, obtemos:

$$\langle X(k)\rangle = (1-p)^{k-1}p \tag{B.8}$$

Como demonstrado, as primeiras soluções tem maior chance de serem copiadas. Este resultado sugere que *indivíduos copiam as soluções mais acessíveis visualmente, mesmo que essas soluções não sejam as mais úteis para a solução do problema*. Na tabela B.3, mostramos que o número de vizinhos podem chegar a 20 para redes de escala livre, o que implica que alguns agentes poderiam ter até 20 soluções disponíveis. Mesmo assim, as soluções com ordenamento inferior são praticamente ignoradas.

Realizamos a mesma analise com o experimento com Sudoku a encontramos um padrão similar.

Table B.3 – Sumário de experimentos com SAT

Topologia	Experimentos (instâncias)	Variáveis	Cláusulas	Agentes	Soluções (copiadas)
Desconectado	16 (16)	3 to 26	5 to 63	125	-
Anel ($K_{ring} = 4$)	4 (3)	4 to 8	16 to 16	16 to 19	335 (13%)
Anel ($K_{ring} = 6$)	3 (2)	8 to 10	12 to 16	22 to 31	268 (14%)
Escala livre ($\gamma = 1.65$)	11 (9)	4 to 26	8 to 63	1 to 32	1179 (10%)
Total	34 (24)	3 to 26	5 to 63	1 to 125	7630 (11%)

* A razão de total de soluções trocadas entre agentes exclui soluções da topologia desconectada.

Table B.4 – Sumário de experimentos com Sudoku

Topologia	Experimentos	Agentes	Soluções (% copiadas)
Anel ($K_{ring} = 4$)	2 (2)	2 to 5	533 (25%)
Anel ($K_{ring} = 6$)	2 (2)	17 to 20	11114 (18%)
Escala livre ($\gamma = 1.65$)	4 (3)	14 to 35	6012 (87%)
Total	8 (6)	2 to 35	17659 (42%)

Todas conclusões foram estatisticamente verificadas com o teste de *Anderson-Darling* e níves de 5% de significância estatística. O modelo de distribução geométrico foi verificado com o teste de*duas amostras de Kolmogorov-Smirnov*.

Uma conclusão interessante que se pode chegar com os nossos experimentos é que há um limite quantitativo para o número de vizinhos que um indivíduo cooperará. Em nossos experimentos, este número é próximo de 6.

Em torno de $5.31\% \pm 6.93$ de todas soluções utilizadas por agentes, em algum momento, já tinham sido utilizadas pelos mesmos agentes. Portanto, ao modelar o comportamento humano online, é necessário utilizar um modelo que permita que humanos esqueçam de e repitam certas soluções.

B.7.2 Resultados da Fase II

Através do nosso modelo com a equação B.8, utilizamos o mesmo tipo de gráfico para visualizar como pessoas tem o comportamento afetado pela posição das opções na tela. Cada k-ésimo elemento na tela recebeu consecutivamente menos interação dos usuários, como mostrado na figura B.22.

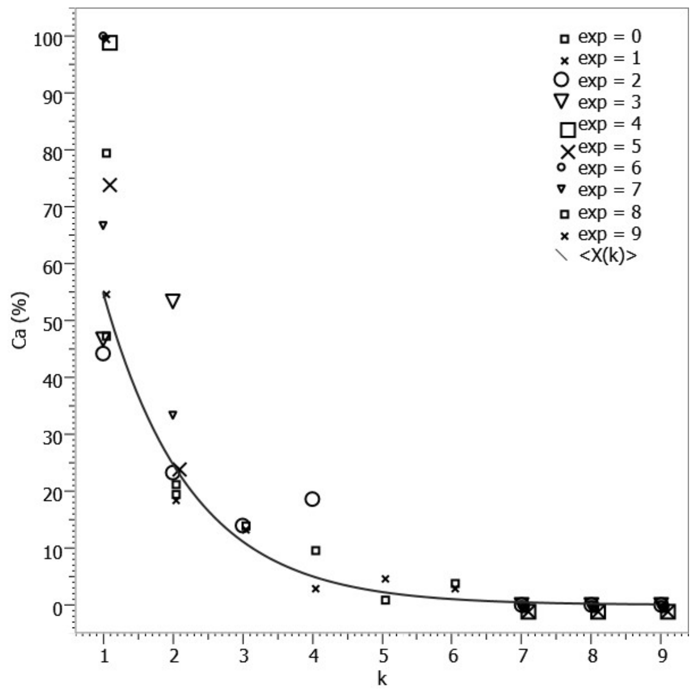


Figure B.21 – Influência da posição na estratégia do agente

Cada ponto mostra a percentagem Ca de uma solução ordenada na posição k na interface. Quando a solução foi copiada, o agente que copiou uma determinada solução resolveu exp instâncias antes daquela instância específica, indicando um certo nível de experiência.

$$\langle X(0)\rangle < \langle X(1)\rangle < ... < \langle X(N)\rangle \tag{B.9}$$

Como consequência, o usuário escolherá a com maior chance a primeira opção para votação, não importa qual seja essa opção. Portanto, escolhendo-se a ordem das opções, é possível determinar o resultado da votação, assumindo que $r(k)$ é suficientemente pequeno. Podemos ver o resultado na figura B.22. Para eliminar a influência do layout e, portanto, remover o efeito indesejável que polui a opinião coletiva da pesquisa sobre preços de política pública, para cada usuário o layout é randomizado e permanece assim até que o usuário elimine os cookies do navegador. Através da combinação da distribuição para cada layout, encontramos os seguintes resultados para a pesquisa de opinião, visto na figura B.23.

B.7.3 Resultados da Fase III

Realizamos a amostragem de dados de cada HIT no AMT a cada 5 minutos em médio durante 2 meses ininterruptos [1]. A tabela B.5 sumariza os dados coletados e compara com um

[1] Exceto em alguns curtos instantes quando a infraestrutura de coleta de dados não estava disponível.

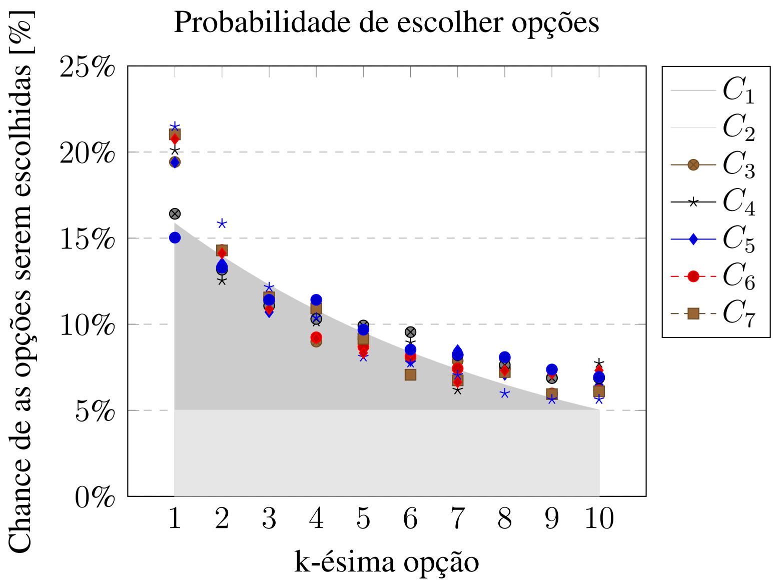


Figure B.22 – Influência posicional de opções em votações de política pública

Cada configuração de opções apresentou distribuições de escolha de opções similares. Os elementos, neste experimento, foram ordenados verticalmente.

Table B.5 – Dados coletados do AMT

	Nosso estudo	Chilton†	Proporção
Setup	Todas páginas	3 paginas	
Total de páginas	240	36	15.00%
Tempo total (horas)	1,440	32	2.22%
Grupos de HITs	67,181	2,040	3.04%
HITs (Total)	3,696,730	2,207,548	59.72%
Soma da recompensa ($)	$531,313.18	$232,051.15	43.68%

†Para referência, ver (CHILTON et al., 2010).

trabalho similar da área (CHILTON et al., 2010).

Para comparar o experimento realizado no AMT com os nossos experimentos anteriores, devemos comparar a taxa de decaimento do número total de tarefas por grupo de HITs com o número de soluções trocadas em nossos experimentos. Nosso objetivo, com a comparação, é provar que nenhum outro fator é mais importante que a posição da tarefa na interface, mesmo considerando a quantidade de dinheiro que um determinado trabalhador pode receber por determinada tarefa.

Na nossa análize, utilizamos uma regressão linear com regularização Ridge sobre os dados com 5 variáveis dependentes:

- **Número de HITs**: é o número total de HITs por grupo de HITs.

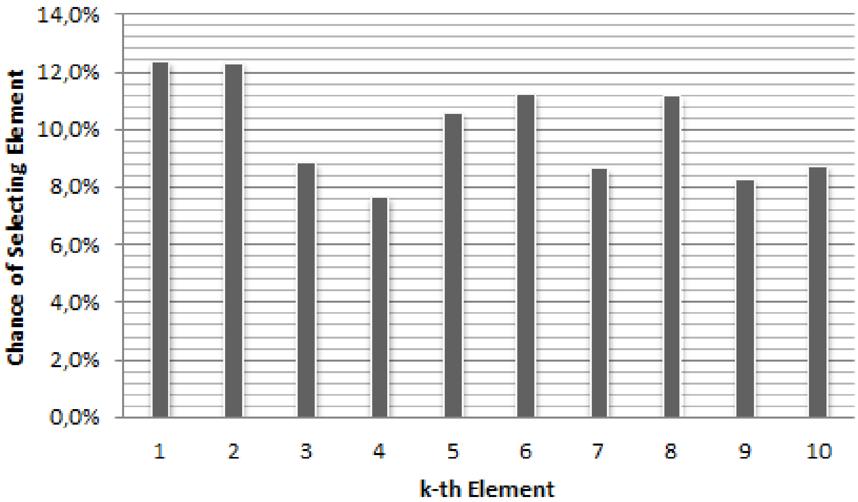


Figure B.23 – Combinando a distribuição dos layouts para eliminar a influência das primeiras opções

Através da randomização, podemos influenciar a escolha de opções que não foram baseadas na escolha de melhor opção segundo os critérios de preferência do agente votante.

- **Ordem alfabética**: para entener a influência deste tipo de ordenamento de tarefas, que é uma opção disponível no web site do AMT.
- **Tempo de expiração**: o quão próximo um determinado grupo de HITs está próximo de expirar e ser removido da lista.
- **Recompensa**: a quantidade de dinheiro recebida ao completar uma tarefa.
- **Tempo alocado**: a quantidade de tempo alocada estimada para realizar uma determinada tarefa.

Nossos resultados, mostrados na tabela B.6 confirmam a nossa hipótese de que a influência da recompensa financiera, inclusive, é inferior ao ordenamento das tarefas na tela. Entretanto, a estatística R^2 ficou muito baixa, mostrando que nosso modelo está incompleto.

Assumindo que outros *requesters* também podem ter percebiddo que aparecer entre as primeiras tarefas é um fator importante, podemos supor que alguns requesters estão manipulando o mercado e não representam o interesse real de trabalhadores do AMT. A maioria das tarefas seriam falsas apenas para inflar o número total de HITs. Removendo os dados dos primeiros colocados, realizamos uma nova regressão linear. Os resultados são mostrados na tabela B.7:

Todos níveis de significância melhoraram e o tempo de expiração e tempo alocado tem níveis de confiância $p < 0.05$.

Table B.6 – Regressão linear no AMT

	Estimativa	Erro padrão	t-Student	p
(Intercepto)	-345.46	29.462	-11.726	1.1553e−31
Ordenação por número de HITs	1.2442	0.11187	11.122	1.1555e−28
Ordem alfabética	0.061772	0.011668	5.2943	1.2055e−07
Ordem de tempo de expiração	-0.016155	0.0097449	-1.6578	0.097375
Ordem por recompensa	-0.0059581	0.0089948	-0.66239	0.50773
Ordem por tempo alocado	0.046183	0.0083148	5.5543	2.8178e−08

Número de observações: 23460, Graus de liberdade de erro: 23454
Erro RMS: 1.25e03
R^2: 0.0103, R^2 ajustado 0.0101
Estatística F vs. modelo constante: 48.7, p = 2.61e−50

Análise estatística dos dados coletados.

Table B.7 – Regressão linear no AMT, com correções

	Estimativa	Erro padrão	t-Student	p
(Intercepto)	-183.87	13.825	-13.299	3.2764e−40
Ordenação por número de HITs	0.78528	0.052312	15.012	1.069e−50
Ordem alfabética	0.030653	0.005438	5.6368	1.7521e−08
Ordem de tempo de expiração	-0.016715	0.0045583	-3.6669	0.0002461
Ordem por recompensa	-0.0081355	0.0042198	-1.9279	0.53878
Ordem por tempo alocado	0.022	0.0038769	5.6745	1.4078e−08

Número de observações: 23142, Graus de liberdade de erro: 23136
Erro RMS: 579
R^2: 0.0167, R^2 ajustado 0.0165
Estatística F vs. modelo constante: 78.8, p = 2.95e−82

Removendo o primeiro resultado melhorou o R^2 e os níveis de significância para as variáveis recompensa e tempo alocado.

Demonstramos, assim, que trabalhadores no AMT se comportam de modo semelhante a experimentos que realizamos anteriormente. Demonstramos, novamente, que as variáveis que normalmente são consideradas mais importantes na avaliação de usuários, elas não são tão importantes quanto fatores de ordenamento.

Entretanto, nossa análise, embora com valores de p satisfatórios, ainda apresentam um R^2 baixo e portanto há mais fatores que deveríamos considerar na nossa análise.

B.8 Conclusão, discussão e trabalhos futuros

Nossos primeiros experimentos indicaram que, quando indivíduos estão resolvendo microtarefas, então um componente forte no custo de busca a avaliação afeta como pessoas tomam decisões em ambientes conectados online. Estes custos de busca e avaliação são comparáveis aos custos de cooperar e portanto precisam reduzir o custo através de gastar menos tempo avaliando opções.

Neste trabalho, a cooperação existe quando agentes copiam soluções uns dos outros. Entretanto, há outras definições de cooperação que podem ser exploradas em trabalhos futuros.

Para embasar as conclusões iniciais, estendemos o nosso estudo com uma segunda fase de experimentos, em que problemas com verdade subjetiva foram explorados. Os resultados obtidos confirmam as conclusões da primeira fase de experimentos.

Ainda com o intuito de generalizar as conclusões obtidas nas duas fases de experimentos anteriores, em que indivíduos participaram voluntariamente de experimentos, nós realizamos novos experimentos com incentivo financeiro para os participantes, denominada Fase de experimentos III. Nesta fase, utilizamos o mercado de trabalho online Amazon Mechanical Turk. Nossos resultados apontam que, novamente, o ordenamento é mais importante para os usuários que até mesmo fatores como recompensa financeira ou tempo de execução das tarefas. Este ambiente é não-controlado.

Embora possamos mostrar consistência nos resultados de todos os experimentos, há algumas observações necessárias. Primeiramente, aplicamos o conceito de *meme* para soluções pequenas, de SAT e Sudoku. Não há informações sobre do resultado dos expeirmentos caso as soluções fossem maiores, inclusive em termos de complexidade computacional, como por exemplo representações binárias de programas. Porrtanto, consideramos experimentos em que, para o caso do SAT, 64 variáveis podem ser representadas por 64 bits, enquanto Sudoku 299 bits. Se tentásemos o modelo memético para problemas envolvendo soluções com linguagem natural, como interpretar conversação humana ou livros, o tamanho da entrada poderia ter bytes ou megabytes, para cada meme.

Nossa pesquisa foi limitada a microtarefas and nós não estendemos os nossos resultados a atividades longas. Embora tenhamos analisado casos em que agentes particiaparam em vários experimentos em sequência, a quantidade de tempo nessas atividades não passou de 1 hora. Como trabalho futuro, podemos realizer experimentos com duração de dias ou até semanas.