Understanding and Evaluating Human Behavior: An Application of Psychology and Machine Learning

Thesis submitted by

Tanveer Ahmed
12120101

under the guidance of

Prof. Abhishek Srivastava

in partial fulfilment of the requirements
for the award of the degree of

Doctor of Philosophy

Department Of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY INDORE

August 2018

THESIS CERTIFICATE

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled "**Understanding and Evaluating Human Behavior: An Application of Psychology and Machine Learning**" in the partial fulfilment of the requirements for the award of the degree of DOCTOR OF PHILOSOPHY and submitted in the DISCIPLINE OF COMPUTER SCIENCE AND ENGINEERING, Indian Institute of Technology Indore, is an authentic record of my own work carried out during the time period from January 2013 to August 2018 under the supervision of Dr. Abhishek Srivastava, Associate Professor, Indian Institute of Technology Indore, India.

The matter presented in this thesis has not been submitted by me for the award of any other degree to this or any other institute.

Signature of the student with date

(Tanveer Ahmed)

– –

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Signature of Thesis Supervisor with date

(Dr. Abhishek Srivastava)

– –

Tanveer Ahmed has successfully given his Ph. D. Oral Examination held on

Signature of Thesis Supervisor
Date:

Convener, DPGC
Date:

Signature of PSPC Member:1
Date:

Signature of PSPC Member:2
Date:

Signature of External Examiner
Date:

ACKNOWLEDGEMENTS

I am very grateful to my supervisor Dr. Abhishek Srivastava for his invaluable guidance, encouragement, and direction throughout this work. He gave me the freedom to work on the topic of my choosing and was very supportive throughout the way. Working with him ultimately resulted into a great deal of enjoyment in my dissertation research. My motivation especially to solve the problem presented in Chapter 6 was, at times, very shaken, but a few words of encouragement helped in getting the work done in an efficient way.

I would like to express my heartfelt gratitude towards my PSPC committee members Dr. Kapil Ahuja and Dr. Trapti Jain for their interesting discussions and suggestions towards my research.

I would also like to thank my mother for her continuous support during the tough phases of my Ph. D. Her suggestions and her words motivated me to continue the hard work during the course of this thesis.

I want to thank everyone who have, in one way or another, helped me to conduct this research. I express my appreciation and indebtedness to my friends Dr. Vipul Kumar Mishra, Amit Jain, Uday Kumar Singh, Rohit Verma, Dheeraj Rane who helped me in many ways during my thesis work.

ABSTRACT

KEYWORDS: Man-Machine Systems; Human Computer Interactions; Human Behaviour; Crowd; Psychology; Machine Learning.

In this thesis, we aim to understand human behavior and the associated humanistic properties through the use of computational methods. The work presented in this thesis is motivated by the fact that we as human beings do not completely understand the internal mental properties, e.g. Motivation, Interest, Altruism, of other people. It is indeed a challenge to have a mechanical machine do this. It could therefore very well be said that to look at such properties through the eyes of an artificial computational agent is non-trivial. In this thesis, we take on this challenge and try to show that there is a way through which we can handle the issue computationally. In doing so, our goal is to take one more step towards understanding the psychological properties of human beings through artificial agents. This is done by combining the core principles of two different fields of research: Psychology and Machine Learning. Further, to conduct a study of the humanistic properties, we perform the analysis of the human psychological properties in crowd based systems. These systems are chosen as they have a natural affinity towards both man and machine. Therefore, they present an excellent opportunity to focus on the technology of the machine and the psychology of the human simultaneously. In doing so, we address two major issues in the thesis.

First, we aim to devise efficient techniques of addressing the challenge of enhancing user participation. The objective is to understand the psychological conditions that make people participate at the online platform and simulate them in a computational environments. The goal here is to make these systems more productive and labour & cost-effective. This is done via analyzing group interactions and collaborative processes at these online platforms. To do this, we divide the problem into two categories: 1) we borrow elements from Machine Learning and propose a recruitment strategy that selects an individual so that the probability of getting a response is maximized; 2) we dig deep into human psychology and try to find alternate means of promoting user participation. We look for new and otherwise overlooked patterns in the behavior of people to find a few interesting facts.

Once we accomplish the previous objective(s), i.e. we are able to motivate users and generate their interest, we then move to the next significant challenge addressed in this thesis. We propose a framework that tries to quantify the interest of an individual towards any entity (say

Facebook, StackOverflow, Amazon Mechanical Turk and so on). Through the proposed framework, we make an attempt to model the long-term evolution of a person's interest. Furthermore, we estimate try to interest at any given day, hour, minute, and so on.

The problems addressed in this thesis are validated by performing simulations on one of the most mature crowdsourcing data repositories on the Internet: StackOverflow. The results show promise especially considering the fact that we have attempted to answer questions and explore some of the previously unexplored patterns in human behavior. We will show that the work carried out in the thesis complements existing literature, and at times, open a few new and previously unexplored dimensions of research in man-machine systems.

Contents

ACKNOWLEDGEMENTS ii

ABSTRACT iii

LIST OF TABLES viii

LIST OF FIGURES ix

ABBREVIATIONS xii

1 Introduction 1

1.1 Thesis Focus and Contribution 6

1.1.1 Questions Attempted to Answer in this Thesis 6

1.1.2 Thesis Focus: Why Study Psychology and Machine Learning Simultaneously? 8

1.2 Thesis Outline 9

1.3 Publications 10

2 Related Work 12

2.1 Work on Enhancing User Participation 12

2.2 Work on Human Interest 16

3 A Probabilistic Approach to Recruit Candidates 20

3.1 Methods 20

3.2 Parameter Estimation 23

3.3 Results 25

3.3.1 Data Collection and Prototype Development 25

3.3.2 DataSet Description 26

3.3.3 Comparison with Unbiased Probability 26

3.3.4 Predictive Capability 28
3.3.5 Importance of History 29
3.4 Summary 31

4 How to Promote User Participation? Applying Human Psychology to Understand and Promote User Activity 32

4.1 Understanding Motivation and Voluntarism 33
4.2 Observing and Understanding Human Behavior 35
4.3 Dimensionality of Work and Selection of Workers 36
4.4 Spatial Characteristics: Role of Nationality and Neighborhood 41
4.5 Response Time and Quality of Response 43
4.6 The Objective and Behavior after Achieving the Objective 48
4.7 Importance of History and Association between Worker and Requester . . . 50
4.8 Importance of Online Profile 51
4.9 The Gender 53
4.10 Stimulating Altruism 54
4.11 Discussion and Threats to Validity 56
4.12 Summary 58

5 Quantifying Uncertainty in the Internal Mental States to Predict Human Interest: An Application of Psychology and Machine Learning 60

5.1 Challenges for Predicting Human Interest 60
5.2 Broad Overview of the Approach 61
5.3 Proposed Framework 63
5.3.1 Predicting Interest via Activity: An Application of Bayesian Statistics 63
5.3.2 Process Step I: A Mathematical Definition for Activity 64
5.3.3 Process Step II: A Stochastic and a Self Evolving Model for Interest 68
5.3.4 Process Step III: Transforming Interest into Activity. A Dynamic and a Self Configuring Measurement Function 74
5.3.5 Process Step IV: Finding Interest from Measurable Activity, An Application of Recursive Bayesian Filtering 76
5.3.6 The issue of Activity Gap 77
5.4 Results 78

5.4.1 Prototype Development 78
5.4.2 Dataset Description 79
5.4.3 Experimental Setup 79
5.4.4 Data analysis 85
5.4.5 Comparison with Random Walk and Geometric Brownian Motion . 88
5.4.6 Stochastic Volatility and Effect of Varying Convergence speed . . . 88
5.4.7 Additional Investigation in Parameters 89
5.5 Discussion 91
5.6 Summary 93

6 Conclusion and Future Work **94**
6.1 Objectives Addressed 94
6.2 Future Work 96

List of Tables

4.1 Statistics of StackOverflow 37

4.2 Year Wise Score of Questions 38

4.3 Reputation Wise Quantity and Quality. Vertical Axis - Question Score, Horizontal Axis - Reputation 40

4.4 Nation Wise Responses. 42

4.5 Number of Subjects in Different Countries 43

4.6 City Wise Responses, Germany 44

4.7 City Wise Responses, France 45

4.8 Performance under Usual and Unusual Behavior 47

4.9 Effect of Profile Photographs. 51

4.10 Effect of Crowd's Profile. 52

4.11 Role of Gender 54

4.12 Statistics Regarding Politeness 55

5.1 An Example of the Attribute Matrix. 80

5.2 Subjective Objective Weights. 81

5.3 Activity Calculation. Bias Parameter $\beta = 0.4$. 82

5.4 Comparison with Random Walk and Geometric Brownian Motion. RW: Random Walk. GBM: Geometric Brownian Motion. Execution Time is in Milliseconds. 87

5.5 Accuracy for Different Variations in Equation (5.13). The variation is modelled via Mean Reverting Stochastic Procedure. 88

5.6 Mode of Variation I. The parameters follow Random Walk. 89

5.7 Mode of Variation II. The parameters follow Geometric Brownian Motion. . 90

List of Figures

3.1 A Snapshot of the Developed Application Deployed over MuleESB. 25
3.2 Proposed Method vs Unbiased Probability. 26
3.3 Probability of Getting a Response for Some Users. 28
3.4 Average Accuracy on a Monthly Basis. 28
3.5 Accuracy Averaged Over an Entire Year. 30
3.6 Importance of History. 30

4.1 Number of Responses per Month 36
4.2 Average Time to Answer in Minutes. 46
4.3 Usual and Unusual Behavior (User 893) 48

5.1 Bayesian Inference for Predicting Interest. 63
5.2 Evolution of Interest for One Random User. 85
5.3 Activity for One Random User. 86
5.4 Predicted Activity vs Actual Activity. 87

NOTATION

rr	A request
k	Total number of responses
N	Total number of requests
$p(i)$	A person from the crowd
$R(i)$	Random variable that denotes if a person responded to a request
π	Probability that a person will respond
$f(\pi)$	Distribution of π
$f(\pi; \alpha, \beta)$	Beta distribution with parameters α and β
$B(\alpha, \beta)$	Beta function
$\Gamma(.)$	Gamma function
$\phi_J(\pi)$	Jeffrey's Prior
$I(\pi)$	Fisher's information
N_{resp}	Number of candidates who responded
N_{recom}	Number of candidates recommended by the system
γ	Parameter Vector
I_t	Interest at time t
A_t	Activity at time t
$T_n(.)$	Transition function
$M_n(.)$	Measurement function
a_ϕ	ϕ^{th} perspective of activity
w_i	weight of the i^{th} perspective of activity
SM	Subjective Matrix
OM	Objective Matrix
e	Identity Matrix
$\mathcal{F}(.)$	Function of various attributes of activity
dW_t	Weiner Process
$\mathcal{N}()$	Gaussian Distribution
λ	Convergence speed for Interest
μ	Long term mean for Interest
σ	Volatility for Interest
λ'	Convergence speed for σ
μ'	Long term mean for σ

σ'	Volatility for σ
λ''	Convergence speed for λ
μ''	Long term mean for λ
σ''	Volatility for λ
$\mathcal{L}()$	Likelihood
X	Set of particles for the particle filter
e_t	Prediction error
l_{ER}	Empirical risk
ρ	Regularization parameter
Ω	Weight vector of RLS
$P(i)$	Inverted and regularized autocorrelation matrix
ρ	Regularization parameter
$K(i)$	Gain matrix
p	Dimension of RLS

ABBREVIATIONS

CPSs	Cyber Physical Systems
OU Process	Ornstein-Uhlenbeck Process
SM	Subjective Matrix
OM	Objective Matrix
LMS	Least Mean Square
RLS	Recursive Least Squares
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
VM	Virtual Machine
RW	Random Walk
GBM	Geometric Brownian Motion

Chapter 1

Introduction

"*To be motivated means to be moved to do something*" [1]. These are the exact words used by some of the most famous psychologists who have investigated interest and motivation. In the Aristotelian view of human development, people are often looked upon as possessing such capabilities. Human beings therefore endeavour tirelessly towards their psychological growth and development. Armed with such arsenal, people tend to seek challenges and aim to redefine their existing limits [2]. In general terms, a person interested or motivated in an object has a natural propensity to take on new challenges and put in extra efforts, thereby indicating the presence of some very promising features of the human mind. The object, towards which a person is motivated or interested to act upon, could be physical (e.g. an athelete interested in a sport of his/her choice), virtual (e.g. Social Networking Websites), intellectual (e.g. pursuit of knowledge), social (e.g. to volunteer for a societal cause) and so on. What motivates versus demotivates a person is often disputed and is attributed to a variety of factors that vary from one individual to another and are generally unknown. Further, it is also clear that sometimes the human spirit can dampen and an individual can reject the motivating factors that once drove him/her in a certain direction [3]. A person wanting to do something today, might feel reluctant to do it tomorrow. The circumstances and the personal attributes regulating one's interest (and motivation) are often subjective, and owing to the limitations imposed by the current state-of-the-art, are hard to simulate in an artificial environment. Consequently, if we study & understand these human properties, the societal impact would certainly be significant. This, however, is easier said than done. Studying internal mental properties in computational environments is non-trivial. In this thesis, we make an attempt to study and analyze some of the very basic human attributes. More precisely, the objective of this thesis is to study the internal properties of human beings through the use of computational methods.

We specified that this thesis aims to understand and study human mental properties. In this regard, a closely related field is: *Cyber-Physical Systems* (CPSs) [4]. The central idea of CPSs revolves around the notion of a machine that, integrated with the state-of-the-art computational and technological capabilities, can interact seamlessly with humans through the use of multiple modalities [4]. Such systems naturally have the tendency to interact and cover a wide array of functions capable of simulating human like behaviour in an artificial environment. The idea that one can go beyond the physical and can combine the virtual with the actual is one of the most fascinating examples that has led to the proliferation of this field. Following the

precedent set by this field, we not only aim to combine the physical and the virtual, but we go one step ahead and aim to combine the physical, the virtual, and the mental. That is, we aim to complement this area by analyzing the mental properties of human beings. To do that, we have to work at the intersection between man and machine. Furthermore, we have to focus on capturing and modelling the psychological properties of humans beings through computational algorithms. As specified in the previous paragraph, this is easier said than done. The rationale here is backed by work presented in [5], where the authors raise an important question: "*Why are people smarter than machines?*" Indeed, it is an important question asked in Artificial Intelligence. To answer this question, literature witnessed a plethora of work dedicated to the study of human cognition. However, and despite the attempts, several authors debated that the effort spent on studying these factors is not significant. This is evident by the work presented in [6] where the authors specify that: "*The effort to understand and simulate human cognitive abilities had been underway for over three decades, and despite initial promise, seemed not to have gotten very far*". From our point of view, the statement of the authors of [6] could be justified considering the fact that the question was asked well ahead of its time (The question was asked in 1986). Back then, the computational capabilities were not advanced enough to capture and simulate human-like intelligence in artificial environments. Since then, however, work has been trying to take the idea forward and has made several attempts to answer this allegedly simple question, e.g. see [7], [8]. In this context, although, work continues to investigate the technological part of the mixture, the psychological point of view is often ill-understood. This, by no means implies that researchers have ignored the human attributes altogether, but the point is that more emphasis is required on the other side of the fence. We believe that emphasis has to be given to three parts simultaneously: 1) technology; 2) human behavior; 3) the psychological properties regulating the person's behavior. Although, the first two points presented here have been considered in the past [8], [9], the internal properties, especially in computational environments, have only started to receive attention more recently. The rationale here is backed by the work presented in [10], where the authors specify: "*...even though computing systems are missioned to satisfy human needs, there has been little attempt to bring understandings of human need/psychology into core system design....*". These lines make a compelling argument for us to focus our efforts on the psychological properties of humans. In doing this, however, there are a few '***core***' challenges:

i. *How to computationally simulate something that is invisible?* Psychological properties are not meant to be directly noticed by a computational agent. Internal mental properties are invisible to any system in deployment, therefore, how to study them in computational environments.

ii. *How to focus on properties that are unique to humans in artificial environments?* Humanistic attributes are hard to simulate. How then should we proceed? What should be the *modus operandi*?

iii. *Is it possible to cover the entire spectrum of studies on human psychology?*

The questions highlighted in the previous points are challenging. The most important one being: The field of human psychology is broad and has witnessed several centuries of research. Work has discovered that there are many internal properties of human beings, e.g. Motivation, Interest, Recognition, Perception, and so on. Hence, it is not possible to study each and every psychological property in this thesis. Moreover, we have to find computational ways to understand these properties in artificial environments. We can understand that owing to limitations imposed by the prevailing technology, this is not possible. Having said that, we will nevertheless focus on understanding a few of the psychological properties. To do this, however, we have to limit the scope of the thesis. This has to be done because it allows us to get an better understanding of a few psychological properties. Therefore, we direct attention on a sub-set of the wide array of studies in human psychology, and focus our efforts on a single application area. In doing so, we will sacrifice on generality, but we can open-up a possibility of having a framework to better understand some of the internal properties of humans. This fractionation, however, is along the existing terms in literature where work has studied and has directed its efforts on studying the properties in specific fields, for instance in crowdsourcing [11], [12]. Therefore, following this precedent, we aim our attention on *crowd based systems*. These systems have been chosen for two reasons: 1) They naturally work at the boundary between man and machine. Recall that the overall aim of the thesis is to work at the intersection between the human and the system. Therefore, crowd oriented systems present an excellent opportunity to focus on the technology and the psychology. 2) There is enough precedent in literature that has tried to study some core human properties in these systems, e.g. [11], [13], [14], [15]. That being said, we must point out that the ideas discussed in this thesis, by no means, are limited to a particular field. The way, the work is formulated, and the mechanism through which the ideas are be presented, can be generalized and applied to a variety of research endeavours across disciplines. To do this, the thesis builds upon existing terms in literature, and tries to complement current work by showing that there is a different side to the same coin. Hence, with the scope of the thesis outlined, let's focus on crowd based systems.

The idea of crowd oriented systems, e.g. crowdsourcing [16], crowdsensing [17], mobile crowdsourcing [18], spatial crowdsourcing [19] etc., is not new. The last decade witnessed a plethora of work dedicated to the study of these platforms. In these platforms, work is divided into several sub parts, participants, often from an unknown audience (called the crowd) are requested to perform the job, and deliver the cumulative outcome. Indeed, the workflow, and therefore, the process-steps seem trivial. However, in reality it is hardly so. From the example presented in this paragraph, one can generalize that such systems rely heavily on their "*human*" workforce. Naturally, with humans in the loop, there are issues. To exemplify one problem, let's consider the case of crowd oriented paradigms that deal with mobile devices, for example

mobile crowdsourcing, crowdsensing, participatory sensing, and so on. In these different fields of research, a human volunteers to provide the requested information through his/her mobile device. To accomplish the functionality, there are several proposals in literature, e.g. [20], [21], [22], [23] (The list is not exhaustive). In these studies, the proposed middleware (that is deployed on a cloud server) selects a person from the crowd and outsources the task to his/her mobile device. Although acceptable, the methods do not give importance to a simple fact: *Will the selected individual respond or even comply to the requested task?* To understand the problem, we present an example commonly followed in literature [21], [22], [23].

Consider a person standing at XYZ square of ABC city at 1900 Hrs. Since the person matches the spatio-temporal requirement of an application (or say context requirement), a middleware deployed on a cloud selects the person. Thus, the task is outsourced to his/her mobile device. If the person complies with the request, he/she will get a suitable reward. In the context of this recurring example, we ask a question - Can we say that the person is definitely going to perform the task and will respond to the request no matter what? Are incentives sufficient to make the person comply to the request? Although, work has specified that "*incentives are probably the easiest way to motivate user participation in almost all types of Mobile Crowd Sensing and Computing applications*" [24], following 100 years of research in the psychology of *crowd dynamics* [25], we can safely say that the answer to both the questions is No. This is further backed by studies in literature that point towards the inference that incentives as the sole form of motivation often creates a negative impact on user participation. To clarify this, we quote from [26]: "*There is no doubt that the benefits of* ***piece-rate systems or pay-for-performance*** *incentive devices can be considerably compromised when the systems undermine workers' intrinsic motivation.*" These lines challenge the status-quo, and compel us to focus more on the human side in the so far mechanical paradigm. The problem that we found with existing literature is that though work tries to mix man and machine, it often treats both the entities on the same scale. But, we know that with machines the chances of getting (or not getting) an answer is high. In other words, even the most complex machines are straightforward, it is much easier to get a yes or a no response from a machine. However, with people there is always the uncertain human element. Treating both the entities on the same scale is therefore a slippery slope. In simple terms, we cannot handle the problem of the human factors without having a complete understanding of the *human psyche*. Moreover, and in the context of the use case presented here, if the motive is to get one's request accepted, we also have to understand what could urge people to respond. To be specific, the question that we raise here is: what could make an individual respond and participate more? Besides incentives or rewards, are there other ways to motivate people? Can we look behind the curtain of technology and inside the psychological properties to find a way of devising efficient techniques of generating interest in users, thereby maximizing and enhancing user participation? The point that we have raised here not only urges us to think differently, but an investigation on the questions raised here

could also open up a new front to explore a fresh set of ideas that, at least in theory, is worthy of a detailed exploration. This basic fact is an important factor with the potential to govern the future enhancements in the field.

The discussion in the previous two paragraphs was focused on exploring human psychology to find alternate, unconventional, and efficient means of enhancing user participation. For the time being, let's assume that we can somehow target the humanistic properties to motivate & generate interest in people. This raises the most significant question this thesis has tried to answer: If an individual is interested in an exercise (crowdsensing, crowdsourcing, swimming etc.), *can we then quantify the person's interest?* Can we find a number representing the person's interest? To exemplify this, consider a person interested in the online platform of StackOverflow (it is an example of crowdsourcing). On this platform, people ask and answer questions. In the context of this example, if a person is eagerly answering questions, he/she has a degree of interest in the platform (for varied reasons). The question that we raise here is: Can we numerically quantify the person's interest in StackOverflow? Can we model the long term evolution of interest? Interest is a mental or perhaps a psychological property that defines an individual's alignment or an innate characteristic towards an object, a subject, a topic or a thing [27]. According to [28], interest is accountable and is indeed a representative of one's desire that indicates the presence of a cognitive phenomenon that quantifies the tendency to engage with one's object of interest (in our case, the online platform). We must point out here that the idea of human interest is not new and has gone through a rigorous string of investigations in the last two centuries [29], [30], [31], [32], [33], [34] (The first paper on interest was published in 1806/1965 [35]). Yet to date, and despite such tremendous efforts, literature has been unable to answer the simple question that we have raised in this paragraph. To add additional complexity to the problem, we want to know how interest evolves in the long run, and what are changes that it goes through every minute, hour, day, week, and so on? We understand that it has not been possible for even a reasonable human being to quantify the interest of another individual. Therefore, it is expectedly a challenge to have a mechanical machine estimate this unique human property.

From the discussion in this section, we can understand that the challenges outlined so far have elements of both practical significance and theoretical import. In this thesis, therefore, we aim to address these issues. By articulating a set of statistical procedures concerning how each of these challenges is handled, the thesis aims to present a potential roadmap that could facilitate the understanding of properties unique to a "*non-mechanical*" entity (the human) and its corresponding interpretation by a "*mechanical*" agent. Though the challenge of modelling, understanding, and quantifying attributes unique to human beings is non-trivial, the goal is merely to discuss a few computational guidelines and theoretical tenets that could then be further explored and applied to a variety of future research endeavours. To do so, the ideas

discussed in this thesis revolves around two distinct fields: 1) Psychology and 2) Machine Learning. The motivation to choose these two distinct fields came from the work presented in [36], where it has been specified that "*To predict the behavior of such systems, it is necessary to start with the mathematical description of patterns found in real-world data*". We lay the foundation of the work in these lines and aim to find patterns in data to analyze, understand, and study a few human mental properties.

1.1 Thesis Focus and Contribution

1.1.1 Questions Attempted to Answer in this Thesis

With the scope and focus of the thesis defined, we now summarize the contribution of the thesis. In particular, the thesis has attempted to answer two interlinked questions:

Q1.What are the different ways to motivate and generate interest in an individual to participate in an exercise?

Frankly speaking, many tasks requested by requesters are not inherently interesting and enjoyable, therefore, understanding users' motivation to promote an active volition becomes a challenge. Though, motivating users is not a new issue, in fact, there is a plethora of work in psychology dedicated to the study of motivation, e.g. [37]. [38], our aim here is to present alternate methods to promote and enhance user participation. In doing so, we aim to complement the existing notion by providing a different way to look at people. In particular, we attempt to address the following questions: How to find the most reliable set of candidates? What motivates the crowd and how to motivate the crowd further? How to ensure persistence in the crowd's participation? Moreover, are *psycho-techno* methods acceptable and computationally operable? In sum, the motive here is to detail the circumstances that foster conditions to make a crowd oriented system more labour and cost effective. To do this, we break the problem into two parts.

- In the first part, we present a framework to select an individual that maximizes the probability of getting a response. We device a general framework and present a probabilistic method to recruit the most suitable set of candidates. We utilize various statistical methods and concepts of Data Science to do this. We propose a recruitment procedure that selects the best candidate based on the history of his/her participation habits. We test the method on real datasets. Through numerical investigation, we have found that the method shows good performance (The details of the framework are discussed in Chapter 3).

- For the second part, we draw inspiration from psychology to dig into the mental properties of people. Recall that for the first part (in the previous point), we focused our

attention on machine learning algorithms. In the next part of the problem, we complement machine learning by presenting the necessary psychological details. We perform a study of human psychology to find methods and techniques to generate interest, thereby maximizing and enhancing user participation. To discover such mechanisms, we investigate the effect of Nationality, Quality of Work, Altruism, Gender, History, and so on. This investigation acted as an addendum revealing a few interesting patterns in the crowd's behavior that so far eludes literature. We have tried to ground the observations in well-tested psychological theories. This is done to understand the core reasoning behind the behavior of people and to present data that supports versus contradicts the formulated hypotheses. Through the analysis conducted in this thesis, we also discuss new ways that show that there are other (and unconventional) ways of enhancing user participation. This is especially promising as the work tries to approach the problem from a fresh perspective. Based on the discussion, we propose a set of recommendations that present an argument to think differently in crowd based systems. Through the discussion presented in Chapter 4, we would also emphasize on the fact that we have to be realistic in our expectations and cannot enforce the Utopian assumption at online systems.

Q2. How to quantify a person's interest and How to model the long term evolution of a interest?

For the second objective, we aim to numerically quantify interest. The authors of [39] specify "*activity plays a significant role in the patterns of human behavior, which is a consequence of interest oriented human activity*". Hence, following this precedent and existing work in psychological literature [40], [41], we assume interest to be a mental state that makes a person engage with the entity of his/her interest. Thus, with this definition, the goal is to *quantify interest*. We present a framework that could model the long-term evolution of interest. To do this, we use basic principles of Bayesian Inference and infer interest indirectly from activity. We make an attempt to assess and evaluate interest using model-based approaches. The proposed framework is generic and can estimate interest towards any entity in the real world (e.g. Facebook, StackOverflow, WhatsApp, Amazon Mechanical Turk, Odesk etc.). We formulate the problem as a latent state estimation problem, and deduce an answer via Bayesian statistics. We specified that we estimate interest via activity. To do this, we first present a subjective-objective weighted approach to find a computationally feasible definition of activity. Subsequently, we discuss a method that models the long-term evolution of interest. We model interest as the Ornstein-Uhlenbeck (OU) process in Physics. Further, by correlating the mental property of interest with the OU process, we discover a few shortcoming. To fix them, we use concepts from stochastic volatility models in Economics and vary the instantaneous volatility of the OU process with time. Furthermore, the convergence speed of the OU process is also made stochastic. We utilize concepts from Adaptive filtering, and use the Recursive Least Mean Squares algorithm to capture the transformation of interest into activity. We use a black box approach that can alter it's internal mechanics on the fly to computationally capture the conversion of interest into activity. Lastly, we employ particle filter and provide a solution

via Monte Carlo Simulations. The proposed framework is validated by conducting simulations on datasets provided by StackOverflow. The experiments reveal a few interesting insights on modeling the mental property of interest via computational procedures (They are discussed in detail in Chapter 5).

It should be noted here that the work presented in this thesis has tried to address some tough challenges in studying human mental properties in computational environments. Therefore, it is not claimed here that the ideas presented in this thesis are cent-percent accurate and are applicable in each and every context. Human psychological attributes, as they are, are hard to study even in controlled laboratory environments. Moreover, our goal here is to study them using data-driven automated procedures. One can understand that this in itself is a challenge. We therefore sacrifice on generality and try to show a potential roadmap to analyze humanistic attributes via machine driven algorithms. In doing so, the objective merely is to present a few sequence of steps that although, imperfect, present a deeper understanding of human behavior through the eyes of a mechanical machine.

1.1.2 Thesis Focus: Why Study Psychology and Machine Learning Simultaneously?

It was outlined in the previous section that this thesis combines the fields of psychology and machine learning. The two disciplines are different and are indeed separate areas of research. A natural question in this context is: Why study them simultaneously? The answer to this question is summarized through the following points:

1. The first and the foremost reason to study machine learning and psychology is: This thesis has the objective of understanding human behavior through the use of computational methods. Psychology is a vital and a functionally notable branch of study that tries to understand typical peculiarities of the human mind. Machine learning, is a discipline that tries to induce intelligence in lifeless machines. Therefore, if the aim is to understand human behavior through statistical methods, it is logical to target the computational capabilities, at much the same time, target properties unique to human beings. The latter has to be done via an extensive study of the human psyche, whereas, for the former, an in-depth analysis of artificial intelligence is of utmost importance.

2. Computer scientists often think about a system in terms of its algorithmic and computational capabilities. Researchers, even working in man-machine systems, often form a bias towards such artificial perspectives (called as expert bias [42]). However, when dealing with a paradigm where humans are one of the major contributors of information, ignoring the crucial human factor is not the right way forward. We cannot take the human factor for granted. Therefore, the motive of this thesis is also to provide adequate representation to the "*human in the loop*".

3. Lastly, we aim to work with the notion of psychological computing. The term was introduced in [10]. In this paper, the authors specify: "*...paper barely scratches the surface of what psychological computing could be and puts forth assorted ideas to motivate the case*". In these lines, we can see that the notion of psychological computing is somewhat abstract and is at its initial stage. Moreover, there is no clear definition and no formal structure of what psychological computing is or could be. Nevertheless, we aim to work with this notion, thereby taking it one step ahead. We aim to complement psychological computing by combining it with machine learning to study the basic internal mental building blocks of human beings.

1.2 Thesis Outline

The work presented in the thesis begins by summarizing the state-of-the-art in Chapter 2. From Chapter 3 onwards, we begin the discussion on the proposed work. The rest of this thesis is organized as follows:

- Chapter 2 reviews existing approaches in literature on studying the mental properties of human beings. 1) We summarize work on maximizing user participation is discussed. The focus is on studies that try to engineer effective recruitment strategies and papers that have tried to understand various psychological properties of human beings; 2) An exhaustive review of computational techniques focusing on modelling interest is performed. We start by looking at the construct from the psychological point of view. Subsequently, we highlight the use of computational methods to study interest.
- Chapter 3 presents the proposed candidate recruitment algorithm. We utilize concepts from statistics and present a method to select the most suitable and reliable set of candidates from the crowd. We present the step-by-step derivation and present the necessary details that automates the procedure of candidate recruitment.
- Chapter 4 addresses the issue of finding new and efficient ways of enhancing user participation. We do this by going deep into Psychology. We look into the often ignored human aspects and study the problem from a psychological perspective. We discuss the importance of a few human factors that could teach us how to encourage user participation. We discuss the psycho-technological approach to observe, understand, and find a few details regarding behavior of humans in online systems. We further discuss the problems and issues with the observations.
- Chapter 5 addresses the most significant issue of the thesis wherein a data driven method that can estimate a person's interest is proposed. As specified previously, one of the objectives of this thesis is to model the allegedly unquantifiable property of interest. This chapter gives a detailed account on the challenges and provides a step-by-step solution to the issue. The chapter also highlights a few shortcoming with the approach.
- Chapter 6 concludes the thesis. This chapter provides a summary of the results and the limitations of the methods proposed. Further, we also identify several key areas to improve upon in future. The chapter also summarizes the list of issues that we intend to pursue.

1.3 Publications

At the time of writing this thesis, a considerable part of the work has been published and is under review at various venues. They are as follows:

Journals:

1. T. Ahmed, A. Srivastava. Understanding and evaluating the behavior of technical users. A study of developer interaction at StackOverflow. *Human-centric Computing and Information Sciences. Springer*, Vol. 7, pp. 8, (2017).

2. T. Ahmed, A. Srivastava. Predicting Human Interest: An Application of Artificial Intelligence and Uncertainty Quantification. *Journal of Uncertainty Analysis and Applications. Springer*, Vol. 4, pp. 9, (2016).

3. T. Ahmed, A. Srivastava. A Prototype Model to Predict Human Interest: Data Based Design to Combine Humans and Machines. *IEEE Transactions on Emerging Topics on Computing*, doi:10.1109/TETC.2017.2686487.

4. T. Ahmed, A. Srivastava. An Automated Approach to Estimate Human Interest. *Applied Intelligence, Springer*, doi:10.1007/s10489-017-0947-7.

5. T. Ahmed, A. Srivastava. Combining Humans and Machines for the Future: A Novel Procedure to Predict Human Interest. *Future Generation Computer Systems, Elsevier*, doi.org/10.1016/j.future.2018.01.043.

6. T. Ahmed, A. Srivastava. Analyzing crowdsourcing to teach mobile crowdsensing a few lessons. *Cognition, Technology, and Work, Springer*, doi.org/10.1007/s10111-018-0474-2.

7. T. Ahmed, A. Srivastava. Will You Accept My Job? A Recruitment Procedure for Mobile Crowdsensing. *Journal of Social and Humanistic Computing*. (**Accepted**).

8. T. Ahmed, A. Srivastava. A Novel Approach to Estimate Human Interest: An Application of Machine Learning and Artificial Intelligence. (**Under Revision**).

9. T. Ahmed, A. Srivastava. Combining Psychology and Technology: A Theory of Reasoned Action Inspired Framework for CrowdSensing. (**Under Review**).

10. T. Ahmed, A. Srivastava. How Much Are You Interested? A Computational Method for Estimating Human Interest. *IEEE Transactions on Knowledge and Data Engineering.* (**Under Review**)

Other papers published during Ph. D

Journals:

1. T. Ahmed, A. Srivastava. A Novel Physics Inspired Approach for Web Service Composition. *International Journal of Web Services Research*, Vol. 11.2, pp. 67-84, (2014).

2. T. Ahmed, A. Srivastava. OMC2: Opportunistic Mobile Computing and Crowdsensing. *International Journal of Mobile Network Design and Innovation* (**Accepted**).

Conferences:

1. T. Ahmed, A. Srivastava. Minimizing Waiting Time for Service Composition: A Frictional Approach. *In Proceedings of IEEE-ICWS*, pp. 268-275, (2013).

2. T. Ahmed, A. Srivastava. A Data-Centric and Machine Based Approach Towards Fixing the Cold Start Problem in Web Service Recommendation. *In Proceedings of IEEE-SCEECS*, pp. 1-6, (2014).

3. T. Ahmed, M. Mrissa, and A. Srivastava. MagEl: Magneto-Electric Effect Inspired Approach for Web Service Composition. *In Proceedings of IEEE-ICWS*, pp. 455-462, (2014).

4. T. Ahmed, A. Tripathi, and A. Srivastava. Rain4Service: An Approach Towards Decentralized Web Service Composition. *In Proceedings of IEEE-SCC*, pp. 267-274, (2014).

5. T. Ahmed, A. Srivastava. Service Choreography: Present and Future. *In Proceedings of IEEE-SCC*, pp. 863-864, (2014).

6. T. Ahmed, R. Verma, M. Bakshi, and A. Srivastava. Membrane Computing Approach for Decentralizing Scientific Workflow Execution in the Cloud. *In Proceedings of Conference on Membrane Computing, Springer*, pp. 51-65, (2014).

7. R. Verma, T. Ahmed, and A. Srivastava. A Generic Workflow Enactment Framework: A Membrane Inspired Approach. *In Proceedings of Conference on Membrane Computing, Springer*, pp. 345-355, (2014).

8. T. Ahmed, A. Srivastava. Choreographing Services Over Mobile Devices. *In Proceedings of International Conference on Service Oriented Computing (ICSOC), Springer*, pp. 429-436, (2014).

Chapter 2

Related Work

We specified that this thesis has attempted to answer two connected questions. In this chapter, we summarize the necessary motivation for the two issues. We start with the first question.

2.1 Work on Enhancing User Participation

The motivation behind the first problem came from [43], where the authors found that the number of responses to their posted requests is only 18.7%. In a similar experiment [44], under little "*favorable*" conditions, the authors found the number to be 42%. Our objective however was cemented through the study conducted in [45], where it was observed that within a social network (Facebook Friends) of average size 260.33, a person receives an average 1.42 no. of responses in 30 minutes, and 5.5 responses in three days. These numbers, for a social network, where in contrast to the unknown crowd, people are actually friends also gave us the initial purpose of this work. Further, we wanted to find out hidden information, if any, in the behavioral characteristics of individuals that could lead us to encourage user participation, and design a better and an effective platform. To do this, recall that the problem was divided into two parts. The first part proposes a statistical framework to recruit candidates that maximizes the chances of getting a response. Second, we look into psychological factors of human beings to find techniques that could urge "*strangers to help strangers*". For the first part, and much like our work, Reddy et *al.* [46] emphasizes on participation history of a volunteer and uses a mathematical system for reputation based volunteer recruitment. However, it fails to understand the importance of human participation in the practical domain of a human dependent computational system. In our attempt, we present observations that can help us understand several human aspects, thus allow us to learn and remedy the situation in early stages. Note, for the first part of the problem, there are several different schools of thought in literature. We specified that this issue is not new and is a current challenge of literature. In this regard, work could be broadly categorized into the following parts:

I) Work focusing on candidate recruitment under the area coverage constraint. That is, work tries to recruit candidates while they are available at a particular location, e.g. [21], [47], [46].

II) Candidate recruitment under cost (or incentive) constraint. In lay terms, the methods in this category tries to find the optimal number of candidates that meets the required budget constraint of the recruiter, e.g. [15], [48].

III) Candidate recruitment under the timing constraint. Work in this category attempts to find potential candidates that will finish the job within the specified deadline, for instance [23], [49].

IV) Work using a combination of the strategies. In this class work has tried to mix several different approaches (for instance a combination of the previous three categories) to select a potential candidate [50], [51].

V) **Miscellaneous**. Apart from these studies, there are also methods focusing on *trust based* candidate selection [52], [53], [54].

With respect to the work highlighted here, we must reiterate here that the idea of recruiting a candidate for crowd based exercises is not new and there are indeed many papers on the topic. However, the essence of the work lies in its capability to present a set of statistical guidelines that is a function of a person's socio-cultural environment in which he/she finds himself/herself. Moreover, the work presented in this chapter is flexible enough to complement existing studies in literature.

For the *second part* of the issue, we attempt to analyze the psychological properties of human beings. We perform a study of human psychology. We mine mature data repositories to look for new and otherwise overlooked patters. Specifically, we mine the datasets provided by StackOverflow to better understand and analyze human behavior. As we explore StackOverflow, it is necessary to highlight work on this platform. In this regard, it has been hypothesized in literature that forums like StackOverflow have the potential of turning into a huge software repository [55]. Consequently, there are several studies that focus on StackOverflow from a software engineering point of view [56], [57], [58]. Further, there are different papers that tries to analyze the technical aspect of StackOverflow. For instance, [59] tries to find patterns that makes a good code example. [60] provide details about the discussion between developers and the latest trends at StackOverflow. [61] studies activity at StackOverflow when the Android API changes. Literature has also analyzed StackOverflow in context of mobile development [62]. [63] uses the criterion of text based matching to find a potential programming question for a user to answer. Moreover, work has gone deeper and has found that high reputation users are efficient in providing good answers [64], [65]. However, in contrast to these work, the study conducted in this thesis makes an attempt to present a different side of StackOverflow. We will focus our attention on the human and the psychological factors related with the human only. Having said that, there are a few studies focusing on technical factors with a light touch on the human factors. This is done in the context of getting an answer accepted at StackOverflow. There is work targeting the human affective state [57] and the presentation quality [66] of an answer to get it accepted. We not only draw inspiration from these work, but go deep into the human aspects and explore the behavior of people in detail. Similar to one of our objectives, the authors of [67] tried to understand how StackOverflow works statistically. However, the

analysis was conducted till March 2010, with stackoverflow started in August 2008. In this respect, the one and half years of StackOverflow saw several upsides, but after almost many years now, StackOverflow has matured, and there are downsides. We not only discuss the downsides affecting people's participation, but, also present several facts that teach us how to understand the core reason behind such an issue and fix it in its early stages. Further, our objective is to try and find methods that could encourage user participation. In this line of thought, Yana et *al.* [68] specifies that experts as well non-experts users exhibit logarithmic growth in participation. However, we will present contradictory evidence in Chapter 4 that shows that this is not the case anymore. We analyze several human traits, one of which is gender based participation. The proposal in [69] also focuses on women in StackOverflow. However, as will be discussed in Chapter 4, our objectives differ, and our experiments are done on a more statistical scale. Also, our findings revisit the centuries old construct of sexism. One of the subtopics of our analysis concerns spatial dynamics of an individual. To this end, Schnek et *al.* [70] explore the behavior of people continentally (Asia, Europe, North America etc.). However, in contrast to continent based analysis, we went to the maximum possible details. We look at countries and the cities within those countries. Through our analysis, the pattern we observed raise several concerns, but at the same time, teach us several new lessons. The work presented in [71] focuses on performance of people after winning a badge in StackOverflow. In this context, the authors [13] specify - "*...upon reaching their achievement, (users) see no immediate need to continue the labour-intensive task*". Further, [71] also reconfirms this result from gamification point of view. We provide additional details to this line of work, and discuss a few directions that could help fix this problem. To determine the quality of answers, Posnet et *al.* [72] specify that experts show expert behavior from the beginning. Along a similar direction, [68], [73] use machine learning to identify experts in their early stages. We also present a similar picture that focuses on the habitual characteristics of a person to determine the quality of answer we can expect. In our attempt, we use unsupervised machine learning algorithms to uncover a hidden pattern in the crowd's behavior. In [74], and similar to our work, the authors focus on Free and Open source softwares. They try to understand what drove people to work on a specific software project and what types of projects are more attractive to people. In contrast to this work, we aim to understand the psychological properties of humans and their what could drive them to participate more in any scenario. Further, work presented in [75] performs a review on motivation and incentives that could help organizations plan better strategies to tap into people's talents. The authors of [76], [77], [78] argue on the fact that relying only on incentives can dampen people's creativity and can affect their productivity. Therefore, following this precedent, we do not focus on incentives mechanisms, rather, we try to find alternate techniques of enhancing and devising efficient labour and cost-effective systems. This does not mean that we negate the notion of incentives, rather, we aim to complement work on several fronts by showing that there are other ways of enhancing user's participation by adopting a *psycho-techno*

approach.

In crowdsourcing literature, there exists a plethora of works focusing on the question “What motivates the crowd?". Consequently, literature has extensively investigated the topics concerning financial incentives [15], skill development [79], task design [14], Intrinsic & Extrinsic motivation [11] etc. Further, comprehensive reviews on crowdsourcing can also be found in [80], [81]. As we trying to understand and encourage participation of users, therefore, we not only take inspiration from this line of work, but, also complement the ideas by presenting a few additional details. The purpose of studying these details is to find methods to get better results from people. In this respect, and to promote user participation, there exist work that shows instantly how a user's contribution makes a difference [82], make users play games [83], use reputation of others [84]. However, we are dealing with crowd based systems, where the result of contribution might not be instantly available. Further, expecting enhancement via games, in our view, is debatable in practical situations (not lab environments), especially considering the business environment, economic feasibility, and in the long run. However, showing reputation of others is important as it is in human nature to learn and look to individuals with high status. Our work also complements this fact by adding extra details. Further, we will make an attempt to find hidden and useful information from the crowd's behavior.

With respect to the state-of-the-art on maximizing the chances of getting a response and enhancing user participation, we highlight a few key differences between our work and existing literature in the following points:

I. We do not focus on recruitment under coverage constraint, cost constraint, or deadline constraint. We focus on the *willingness of the participant to accept a request and provide a response*. We discuss a probabilistic approach that chooses a potential recruit using the rules of Bayesian inference.

II. We do not limit the scope of the work, the proposed method is generic and can be applied to a variety of fields dealing with the crowd selection problem.

III. We focus on the core psychological properties of human beings and try to find new techniques and novel methods to target mental properties to make people participate more.

IV. We present several unnoticed and otherwise overlooked patterns in the behavior of people. We then look deep into human psychology to find reasons for such actions.

V. We take an interdisciplinary approach comprising of psychology and machine learning to understand the typical habits of the crowd.

VI. We target the typical mental attributes of human beings, e.g. Altruism, Role of Nationality, Gender, Online presence, Challenges, etc., to understand and engineer an effective crowd oriented platform.

2.2 Work on Human Interest

In this subsection, we summarize work on interest and try to give teh motivation behind the second question asked in section 1.1.1. The first paper on interest is more than 200 years old (published in German in 1806/1965 [35]). Following this work, the last two centuries have witnessed a huge body of work dedicated to the study of human interest [27], [41], [40], [85], [86], [87], [88]. Moreover, the concept has attracted attention across disciplines. In an endeavour to find an solution to the issue we found that interest is broadly classified into the following dimensions: Individual interest and Situational interest [27], [40], [88]. *Individual Interest* is a relatively long and persistent enduring preference towards an entity, subject, topic or anything, for example an academic interested in pursuing his/her research goals. On the other hand, *situational interest* is momentary, short, and is aroused by the temporary affect of the contextual stimuli, e.g. exotic natural scenery arousing momentary interest in the viewer. The simultaneous application of these two broad categories stimulates the feeling of interest. Both categories have received substantial amount of attention and have been investigated thoroughly in literature [29], [30], [31], [33]. Work has further gone deep to understand the concept and has found that interest has a deep biological foundation [89], [90]. It has been specified that the *seeking system of the brain* is a genetic and an evolutionary procedure that lays the foundation for the psychological state of mind. This causes a mammal to engage cognitively, physically, and symbolically with an object of interest [28]. Although these studies present a broad spectrum of analysis, it must be pointed out that significant work on interest started from *learning's point of view* i.e how does interest stimulates the learning capabilities of a person especially in educational context. The earliest work on *interest and learning* dates back to 1913 [91]. Following the seminal work, the idea went through a rigorous string of investigation. From a modern point of view, it still attracts attention in literature, e.g. see [92]. It should be noted here that the work discussed here provided the necessary theoretical foundation, but, the inability to provide an answer created several research gaps. They are: 1) We found that there is a lack of a method that can find a *number* for someone's interest towards any entity. 2) A mathematical model that can dynamically transform interest into activity eludes literature. 3) A statistical framework that can capture the continuous and 'long term' evolutionary dynamics of interest is unexplored. 4) There is no practically feasible solution to estimate interest towards '*any entity*'. Thus, to quantify interest, we have to address each these four issues.

It was specified in the previous paragraph that interest has received much attention in literature. In this regard, there are studies in the past that have tried to analyze interest using Artificial Intelligence. For instance, [34] tries to detect 'spontaneous' interest in natural conversations. This is done by analyzing the acoustic properties. However, in contrast to the work proposed in this thesis, where the aim is to quantify and model the long term evolution

of interest, the authors focus their efforts on determining the level of interest, e.g. disinterest, indifference, light interest, strong interest etc. Moreover, the idea of Individual interest is not investigated. In much the same way, work presented in [93] tries to the use concept of bidirectional long-short term memory and bottleneck networks to recognize the level of interest in natural conversations. [94] use the same idea, and employ the principle of lexical analysis and acoustic cues. Further, they combine the idea of contextual information in spoken languages. [95], [96], [97] use classification algorithms and combine the notion of multiple modalities to classify a student's interest into different categories, e.g. high, medium, low. Work presented in [98] focus on analyzing interest via content presentation and eye movement. The study in [99] tries to detect basic facial expression, Surprise, Sad, Happy etc., from videos and tries to study interest. Similarly, [100] tries to detect the *level of curiosity* from eye movements using data mining algorithms. Work presented in [101] use Bayesian Inference to predict the level of frustration. There is also substantial literature combining multiple modes to detect the degree of interest [102], [103], [104]. There is also a dedicated body of work to *analyze and deduce one's "topic" of interest*. The line of research here is available in the context of online searches and content recommendation. In this regard, the work of White et. al [105] focuses on detecting a user's topic of interest from the search history for website recommendation via contextual information. In [106], the authors propose a two-level learning procedure to track a user's non-stationary interest. In much the same way, the authors of [107] build upon the work presented in [106] and use Bayesian inference to track changes in a user's interest with built-in mechanisms for profile learning and tracking. The authors of [108] use the concept of three descriptor model to track a user's long term and short term topic of interest. The study discussed in [109] focus the same idea, but uses the criterion of long term and short term search history. Similarly, [110] uses click behavior to infer interest. Moreover, work has also investigated the importance of context to predict one's interest [111]. Similar to our work on time series based modeling of interest, work has also dedicated effort towards analyzing variables via time-based data analysis [112], [113]. [114] proposes a method for time-series prediction using Neuro-Fuzzy Inference based system. [115] extends the idea to multivariate prediction. Similarly, there is yet another body of work trying to analyze a user's engagement patterns. The application area of this line of work varies across disciplines. For instance, [116], [117] analyzes the time to update Wikipedia articles. Here the authors found a double power law distribution between simultaneous updates. Work presented in [118] analyzes huge data traces to find patterns in the activity of researchers at Sciencenet. The authors found a power law relationship between visiting frequency and the subsequent visitors. Further, they found the existence of Heaps' law and the memory effect. Along the same direction, work presented in [119] uses a time decoupling approach to analyze temporal patterns in online forums. There are bodies of work that employ numerical algorithms to study online forums in detail, with certain bits of work that contextualize the scope [120], [121], [122]. [123] analyzes short communication

and finds a bimodal existence of inter-event time distribution. Similarly, [39] finds evidence of power law distribution for rating movies. In sum, and on work investigating interest, we must specify here that the idea of analyzing interest via machine based algorithms is not new. Interest has attracted substantial attention in the literature. However, most efforts are unable to address the four points raised in the previous paragraph. Moreover, work has mostly focused on *spontaneous* interest. That is interest at the present moment of time. In this thesis, we not only concentrate on this criterion but, also focus on *long-term* interest. As will be discussed in detail in Chapter 5, we estimate interest via activity. An impedence to estimating interest via activity is that data about activity is often not available. Therefore, interest estimation in such scenarios is not possible. This issue is known as the problem of *activity gap* (See Chapter 5 for more details). In this thesis, we propose a method to overcome this issue and present a reasonable guideline to build additional constructive work in the field.

In literature there is a growing body of work dedicated to the study of artificial intelligence and robotics where several authors have tried to study and analyze basic human properties. In this regard, and perhaps the closest to our work on quantifying interest, from a theoretical point of view, is [124]. The authors of this paper use information theory framework to present several possible models for Intrinsic Motivation. For instance, the authors discuss potential mathematical models for Uncertainty Motivation, Information Gain Motivation, Novelty Motivation, and so on. This paper presents a broad theoretical perspective on different types of Motivation. We not only draw inspiration from this study, but build upon this work to present a new method to analyze one of the basic internal human properties. We go deep into interest and try to understand it using data driven computational methods. Moreover, by conducting numerical simulations on StackOverflow datasets, we offer a possible guideline to explore other internal mental states. Although, [124] is closest to this work on a theoretical level, from a practical point of view, the closest is [125]. This work focuses on three critical issues: 1) Mathematical distribution of the time for which interest lasts. 2) Statistical distribution to model the return of a user to a previous topic of interest. 3) Ranking of interest and its transition. Though, as will be discussed in Chapter 5, we do not focus on any of these criteria, nevertheless, we follow the initiative of the authors, and take one more step towards bridging the gap between data analytics and its capability to quantify internal human states. The authors of [125] have specified "*As a branch of the science of "Big Data", the field of human-interest dynamics is at its infancy*". The motivation of this work is laid down in these lines to model interest using data driven algorithms.

In Chapter 5 of this thesis, interest is modeled as an Ornstein-Uhlenbeck Process, [126]. This process was proposed in 1930 to describe the motion of a physical particle in space. Following the success of this process, it has been employed in many fields, for example interest rate [127], electricity prices [128], membrane depolarization [129], neurological spikes in the

brain [130], phylogeny (genetic evolution of continuous traits) [131] and so on. Though, the model served perfectly as the base criterion, the aim was to further improve the work. To this end, it was found that literature in Economics deal with complex dynamic systems. Therefore, the core mathematical concepts of this discipline were explored and a Stochastic Volatility model was employed. Stochastic Volatility models were proposed to overcome the shortcomings of the famous Black-Scholes formula. The formula was unable to capture the volatility dynamics in a financial asset. Stochastic Volatility models overcame this problem by making the volatility in the underlying asset stochastic [132], [133]. Taking inspiration from this line of work, the volatility of the OU process is made stochastic. In addition, experimentation is also performed by varying the convergence speed of the process. Moreover, three different types of variations are explored. Although, these variations added additional complexity and increased the execution time, but, it also improved accuracy.

To highlight the novel contribution towards addressing the second question of the thesis, the following points summarize the key differences between the proposed framework and existing work:

I) We neither study the level of interest nor do we infer the topic of interest. The objective is to quantify interest with emphasis on its long term evolution.

II) The aim of the proposed method is to present a general framework to quantify interest towards any entity. Though, the model cannot extensively cover every aspect, we have attempted to present a general framework.

III) The objective of the method is to find a number for interest at any given time interval. For instance, we discuss a procedure to quantify an individual's interest (towards Amazon Mechanical Turk, ODesk, StackOverflow etc.) at any point in a day, hour, minute, and so on.

IV) We propose a computationally feasible definition to measure activity.

V) We further propose a novel method to model the dynamics of interest.

VI) Lastly, we discuss a method to dynamically convert interest into activity.

Chapter 3

A Probabilistic Approach to Recruit Candidates

In this Chapter, we present a method to recruit potential candidates so as to maximize the probability of getting a response from the crowd. Recall that in Chapter 1 we specified that it is not compulsory for any human being to provide a response to a request. Therefore, to maximize the chances of getting a response, we should target individuals who have a high probability of sending a response. In this Chapter, we present a statistical perspective to do this. The goal here is to make the procedure of candidate recruitment automatic and more reliable. We utilize concepts of probability theory and draw inspiration from statistics to do this.

3.1 Methods

To accomplish the objective of selecting a candidate, we employ prior information available to the system, and predict the most likely estimate of the posterior. To achieve this, we derive the probability of getting a response from the person. To explain the method, lets assume that the person has responded k times out of a total of N requests in the past.

We define a random variable $R(i)$ for the person $p(i)$ as

$$R(i) = 1 \text{ \{if the person responds to the i}^{th} \text{ request.\}} \tag{3.1}$$

and,

$$T(N) = \sum_{j=1}^{N} R(i) \text{ \{The total number of responses\}} \tag{3.2}$$

Lets assume that the probability that a person $p_{(i)}$ is willing to comply to a request r is π.

$$\Rightarrow P(R(i) = 1) = \pi; f(\pi) \tag{3.3}$$

where, $f(\pi)$ is the distribution function of the parameter π. Similarly, the probability that the person $p(i)$ is not willing to provide a response is expressed as

$$P(R(i) = 0) = 1 - \pi \tag{3.4}$$

As a request can be initiated by any number of requesters at any time. Therefore, we assume that each request is initiated independently. As a result, the distribution of $T(N) = k$ follows binomial distribution.

$$\Rightarrow P(T(N) = k) = \binom{N}{k} \pi^k (1-\pi)^{N-k} \tag{3.5}$$

In this formulation, our objective is to calculate the probability of getting a response at the (N+1)th request, given a history of N requests (out of which k responses are obtained). In simple words, given a history of participation, we have to predict whether we can expect future participation from the person or not. To do this, we use Bayes theorem. According to Bayes theorem, posterior is proportional to prior times likelihood. Therefore, using this property, we express the issue as:

$$P(R(N+1) = 1 | T(N) = k) =$$

$$\frac{P(T(N) = k | R(N+1) = 1) \times P(R(N+1) = 1)}{P(T(N) = k)} \tag{3.6}$$

We know,

$$\Rightarrow P(T(N) = k | R(N+1) = 1) = P(T(N) = k) \tag{3.7}$$

In this chapter, we use the assumption that $T(N) = k$ and $R(N+1)$ are independent given π [134]. Therefore,

$$P(T(N) = k) = \int_0^1 P(T(N) = k | \pi) f(\pi) d\pi \tag{3.8}$$

We know that the probability that a person is going to comply to the next incoming request is π. Therefore, we have

$$P(R(N+1) = 1) = \pi \tag{3.9}$$

Substituting (3.7), (3.8), (3.9) in (3.6), and simplifying the equations, we get

$$P(R(N+1)=1|T(N)=k)=\frac{\int_0^1 \pi P(T(X)=k|\pi)f(\pi)d\pi}{\int_0^1 P(T(X)=k|\pi)f(\pi)d\pi} \tag{3.10}$$

Substituting the expression from (3.5) in (3.10) and simplifying, we get

$$P(R(N+1)=1|T(N)=k)=\frac{\int_0^1 \pi^{k+1}(1-\pi)^{N-k}f(\pi)d\pi}{\int_0^1 \pi^{k}(1-\pi)^{N-k}f(\pi)d\pi} \tag{3.11}$$

To solve the above equation, we require a particular distribution for the probability of a person responding to a request. In other words, we need a distribution function ($f(\pi)$) for the parameter π. In crowd oriented applications, the probability of success at each trial is not fixed but random. As a result, an ideal candidate for this particular situation is conjugate beta prior density function. Thus, we use the Conjugate beta prior to define the probability density function of π. This function is defined as:

$$f(\pi;\alpha,\beta)=\frac{\pi^{\alpha-1}(1-\pi)^{\beta-1}}{B(\alpha,\beta)} \tag{3.12}$$

where, α and β are the two parameters, $B(\alpha,\beta)$ is the beta function defined as

$$B(\alpha,\beta)=\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt \tag{3.13}$$

The mean and variance of the beta distribution are well known and are as follows:

$$E[\pi]=\frac{\alpha}{\alpha+\beta} \tag{3.14}$$

$$var(\pi)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{3.15}$$

Representing equation (3.12) in terms of Gamma function, we get

$$f(\pi;\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)(\Gamma(\beta))}\pi^{\alpha-1}(1-\pi)^{\beta-1} \tag{3.16}$$

Substituting the above expression from equation (3.16) in equation (3.11), we get

$$P(R(N+1)=1|T(N)=k)$$

$$= \frac{\int_0^1 \pi^{k+1}(1-\pi)^{N-k}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)(\Gamma(\beta))}\pi^{\alpha-1}(1-\pi)^{\beta-1}d\pi}{\int_0^1 \pi^{k}(1-\pi)^{N-k}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)(\Gamma(\beta))}\pi^{\alpha-1}(1-\pi)^{\beta-1}d\pi} \tag{3.17}$$

Simplifying the above equation and manipulating the numerator and denominator, we get

$$P(R(N+1)=1|T(N)=k) = \frac{\int_0^1 \pi^{k+\alpha}(1-\pi)^{N-k+\beta-1}d\pi}{\int_0^1 \pi^{k+\alpha-1}(1-\pi)^{N-k+\beta-1}d\pi} \tag{3.18}$$

This equation has similarity to the Beta function described in equation (3.13). Therefore, rewriting the equation in terms of Beta function, we get

$$P(R(N+1)=1|T(N)=k) = \frac{B(k+\alpha+1, N-k+\beta)}{B(k+\alpha, N-k+\beta)} \tag{3.19}$$

After observing the partial result, we go back to the assumption of equation (3.12), where we assumed the prior as a Beta distribution. Using the property of conjugate priors, the posterior distribution of π is also beta distribution. As a result, we substitute the values of α and β as $\alpha + k$ and $\beta + N - k$ respectively [135]. Using these values and by simplifying the fraction, we get

$$P(R(N+1)=1|T(N)=k) = \frac{B(2k+\alpha+1, 2N-2k+\beta)}{B(2k+\alpha, 2N-2k+\beta)} \tag{3.20}$$

3.2 Parameter Estimation

The derivation presented in the previous subsection provides a method to estimate the probability of getting a response from the person. However, from equation (3.20), we can see that the probability is dependent on two unknown parameters: 1) α and 2) β. Therefore, we need a data driven method to estimate their numerical values. In this regard, we estimate the parameters via Jeffry's prior. This method is chosen because it is invariant to the effect of reparametrization [136], thereby following the principle of "let the data do the talking". Jeffery's Prior is defined as :

$$\phi_J(\pi) \propto \sqrt{I(\pi)} \tag{3.21}$$

where, $I(\pi)$ is the fisher's information, defined as

$$I(\pi) = -E[\frac{d^2 \log p(K|\pi)}{d^2\pi}]$$

In our case, $k \sim Binomial(N, \pi)$ and

$$p(k|\pi) = \binom{N}{k}\pi^k(1-\pi)^{(n-k)}$$

Taking the log of above expression and differentiating twice, we get

$$\begin{aligned}
\log(p(k|\pi)) &= \log\binom{N}{k} + k\log(\pi) + (N-k)\log(1-\pi) \\
\frac{d\log(p(k|\pi))}{d\pi} &= \frac{k}{\pi} - \frac{N-k}{(1-\pi)} \\
\frac{d^2\log(p(k|\pi))}{d\pi^2} &= -\frac{k}{\pi^2} - \frac{N-k}{(1-\pi)^2}
\end{aligned}$$

We know that the expected value ($E[K]$) of Binomial Distribution is $N\pi$, therefore, substituting k as $N\pi$ in above equation, and using Fisher's informtion, we get

$$\begin{aligned}
I(\pi) &= -E[\frac{d^2 \log p(K|\pi)}{d^2\pi}] \\
&= \frac{N\pi}{\pi^2} + \frac{N - N\pi}{(1-\pi)^2} \\
&= \frac{N}{\pi} + \frac{N}{1-\pi} \\
&= \frac{N}{\pi(1-\pi)}
\end{aligned}$$

where, N is a constant. Therefore, from equation (3.21) we have

$$\phi_J(\pi) \propto \pi^{-1/2}(1-\pi)^{-1/2} \tag{3.22}$$

This expression follows a Beta distribution $B(\alpha, \beta)$ with parameters 0.5 and 0.5 (for details see [135]). Thus, for the proposed framework we choose these values. An advantage of using these particular values is that they represent *non informative priors*, thereby following the principle of - "*Let the data do the talking*". This is important as we have an element of objectivity in the system.

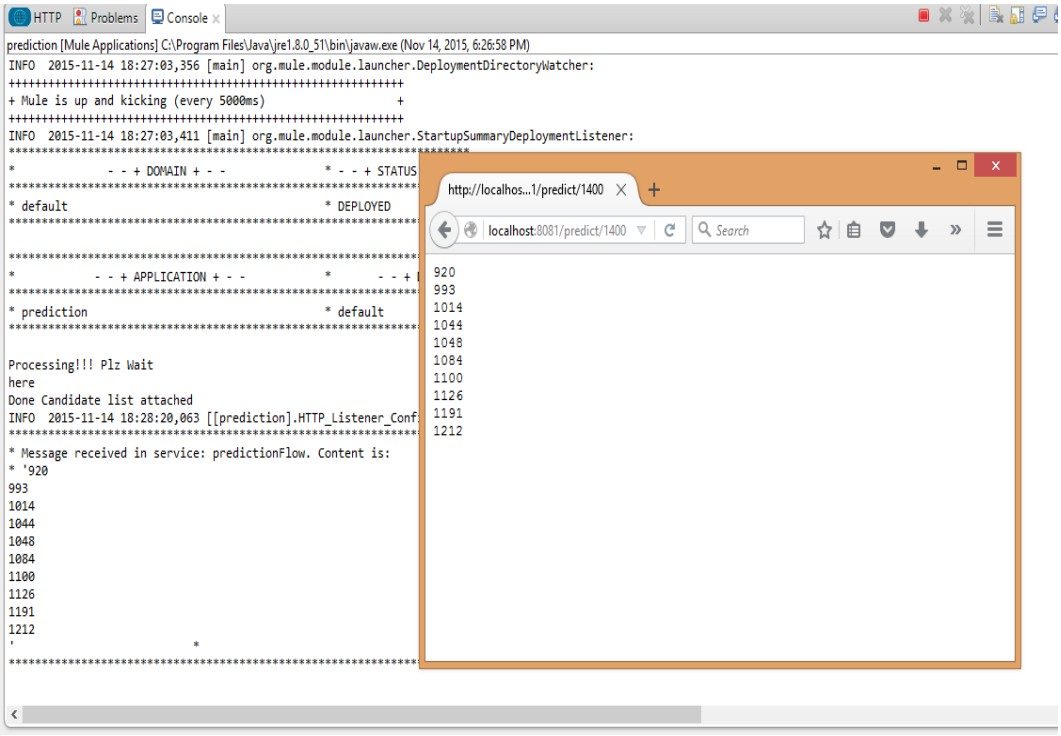

Figure 3.1: A Snapshot of the Developed Application Deployed over MuleESB.

3.3 Results

3.3.1 Data Collection and Prototype Development

To validate the viability of the proposed method in actual deployment, we have developed a prototype. The prototype was developed as a standalone application. The prototype deployed as a Web based application was implemented using Java, and is deployed over an Enterprise Service Bus, MuleESB[1]. We chose MuleESB for two reasons. 1) It is open source and freely available. 2) By deploying the proposed framework on an ESB, we show the feasibility of the method in current cloud based computational environments. The application was developed via Anypoint Studio v5.3.0. The inbuilt server package was deployed on an Machine with i7 Processor, 8GB Ram, and 2.4 Ghz processing speed with Windows 8 as the Operating System.

[1]https://www.mulesoft.org/

The application developed using RESTful principles provided a uniform method of accessing the information stored at the middleware. Thus, using this type of a methodology, we provided a universal strategy to invoke the application from any entity in the real world. A snapshot of the application deployed using this settings is shown in Fig. 3.1.

3.3.2 DataSet Description

To numerically test the performance of the proposed recruitment algorithm, we experiment with datasets provided by StackOverflow. It is a crowd oriented discussion forum. It has one of the largest data repositories and has been documented extensively in literature. The dataset describe the details of all the posts, comments, questions, answers, votes and so on. From this dataset, we assumed that a question posted by a user is analogous to a request by a requester, and the answer is analogous to a response provided by the worker. From this dataset, we have collected the details of 10,000 users. We did this for a period of one year.

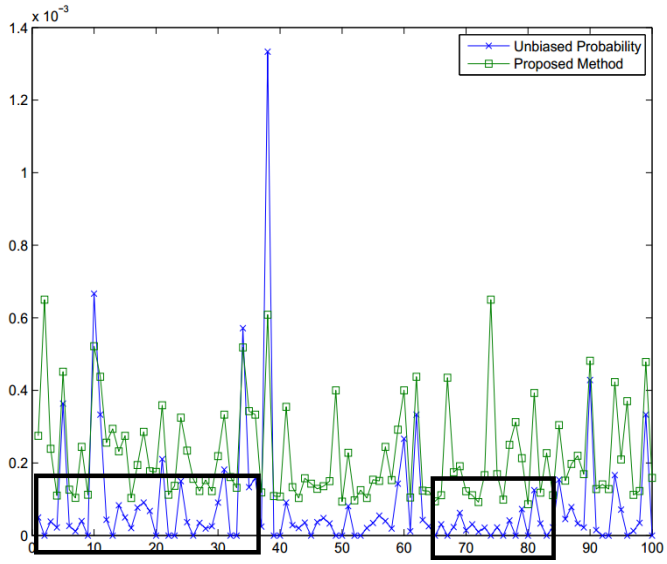

Figure 3.2: Proposed Method vs Unbiased Probability.

3.3.3 Comparison with Unbiased Probability

$$\text{Unbiased Probability}, P(R(N+1)) = \frac{k}{N} \tag{3.23}$$

To compare the performance of the method with Unbiased probability (equation 3.23), we have shown the probability of getting a response for 100 days in Fig. 3.2. The data to calculate the probability on the current day was taken from the previous day. In this figure, we have highlighted a few cases where the number of responses from a person is zero. In this scenario,

since the person has not responded at all, therefore, the unbiased method is producing a zero numerical value ($k = 0$, $N \neq 0$). In other words, the system is certain that the person is never going to respond to any of the future requests. This is infeasible in practical situations, especially considering the fact that we are dealing with a human crowd. With humans, the uncertainty factor is high, consequently, the participation at an exercise can change any time. We know that human behavior is erratic and can go through several changes. Therefore, if we want to work with human beings, we need information that incorporates typical human factors. In this regard, and in contrast to unbiased probability, the proposed method is producing a lower numerical value i.e. the probability of getting a response is less. This is acceptable because if the users did not responded to any of the request, we can say that the probability of such users participating in future exercises is also less, but note, it is not zero.

To further show the importance of the proposed method, consider the case of the cold start problem. By cold start problem, we imply that the user is new to the system, and has neither received nor responded to any of the requests. In that case (k=0, N=0), the unbiased probability (equation (3.23)), will produce $\frac{0}{0}$. In other words, the system is stuck in unstable state. This is problematic in real situations. In contrast, using the derivation shown in Section 3.1, the proposed model is not stuck at all. To highlight this, we substitute the values of k, N as zero in equation (3.20) and get:

$$P(R(N+1) = 1|T(N) = k) = \frac{B(\alpha + 1, \beta)}{B(\alpha, \beta)} \tag{3.24}$$

From the derivation shown in Section 3.2, we know that the the values of α and β is 0.5. Therefore,

$$\begin{aligned} P(R(N+1) = 1|T(N) = 0) &= \frac{B(1.5, 0.5)}{B(0.5, 0.5)} \\ &= \frac{1.57}{3.14} \\ &= 0.5 \end{aligned}$$

Thus, the probability of getting a response is 0.5. This is intuitively as well as practically feasible. In other words, when a user is new to the system, there is a 50% chance that he/she will comply to a request.

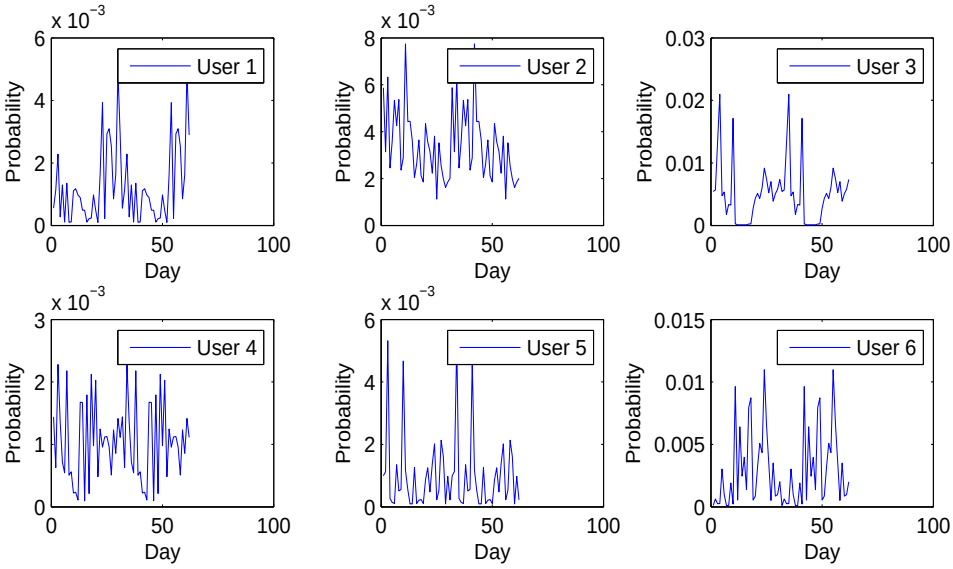

Figure 3.3: Probability of Getting a Response for Some Users.

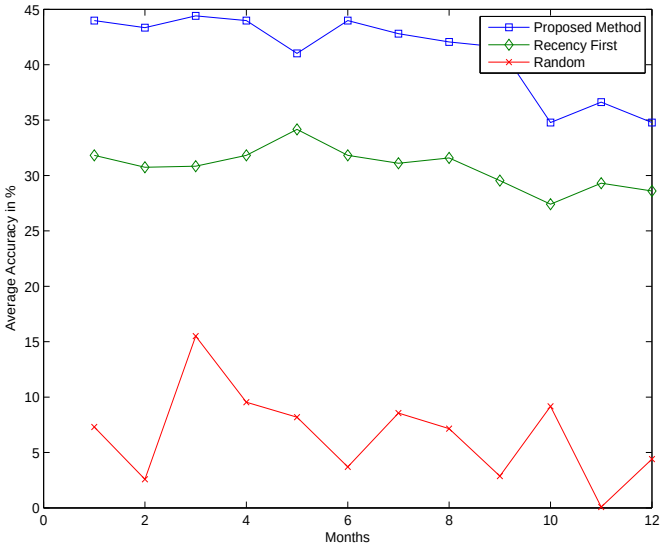

Figure 3.4: Average Accuracy on a Monthly Basis.

3.3.4 Predictive Capability

To begin with the analysis on the predictive capability of the method, we have shown the daily evolution of probability for a few users monitored for a continuous period of 60 days in Fig. 3.3. It is visible from the figure that the probability for these users follow several ups and downs. This type of a pattern is expected as no person from the crowd is going to participate with the same rigor everyday. Owing to certain circumstances in a person's daily routine, these type of situations are expected. However, a pertinent question in this context is: With this erratic and constantly changing behavior, what is the accuracy of the system? In the next series of experiments, we test the accuracy of the method. In the experiments, accuracy is defined as

follows:

$$Accuracy = \frac{N_{resp}}{N_{recom}}$$

where, N_{resp} is the number of candidates who actually responded, and N_{recom} is the number of candidates who were recommended by the system.

To test the method in real scenarios, we have compared the performance with Random user selection method and recency first method [21]. By recency first method, we imply selecting a person who has recently provided a response. To begin with the test, we chose each day in the dataset, and calculated the probability of selection for the next day. Therefore, the method automatically selected a few candidates and recommended them to the requester. With this type of a testing methodology, the result for *each month* is shown Fig. 3.4. Further, the accuracy values averaged over the year is also presented in Fig. 3.5. It is clear from the figures that recruiting candidates randomly is certainly not the best way forward. In this context, and according to the recommendation of literature [21], selecting a person based on recency first method might seem a good option. However, from the results, the accuracy of the proposed method is much better than the accuracy of recency first method. To be precise, the accuracy of the proposed method is $\sim$42%, whereas the accuracy of recency first method is $\sim$31%. Thus, the method showed good performance.

3.3.5 Importance of History

The next series of tests were conducted to test the behavior of historical values in predicting the future behavior of the crowd. Specifically, we wanted to find out the answers to the following questions: If a person has responded to a request today, then what is the probability that he/she will respond tomorrow? Moreover, what is accuracy? Further, if a person has been active for one week, then what is the probability that he/she will be active tomorrow?

To find the answers to these questions, we conducted a few tests. The tests were designed as follows: We wanted to check the predictive capability of the framework by taking in the entire data for the previous one day, previous seven days, last fifteen days, and the last one month. The results are presented in Fig. 3.6. As shown in the figure, we get a high accuracy value when when we look into the last one day and the past fifteen days. Though, accuracy is high for the test concerning the last one day, but the difference is not significant. To be precise, the values for one day is 41.76%, and the number for fifteen days is 41.54%. The exact

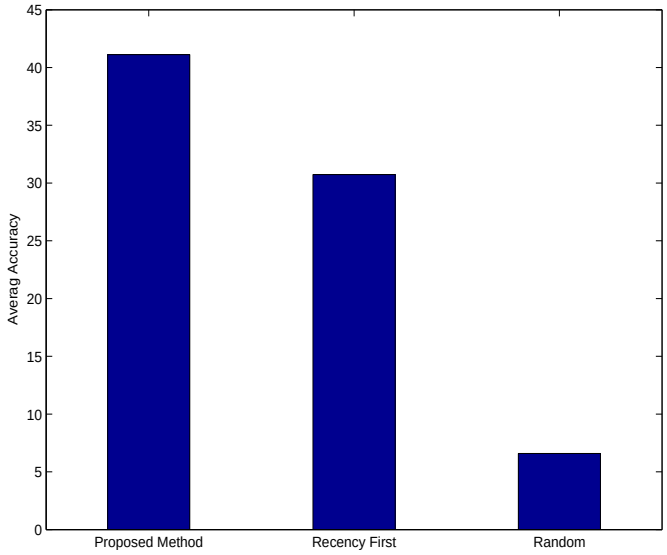

Figure 3.5: Accuracy Averaged Over an Entire Year.

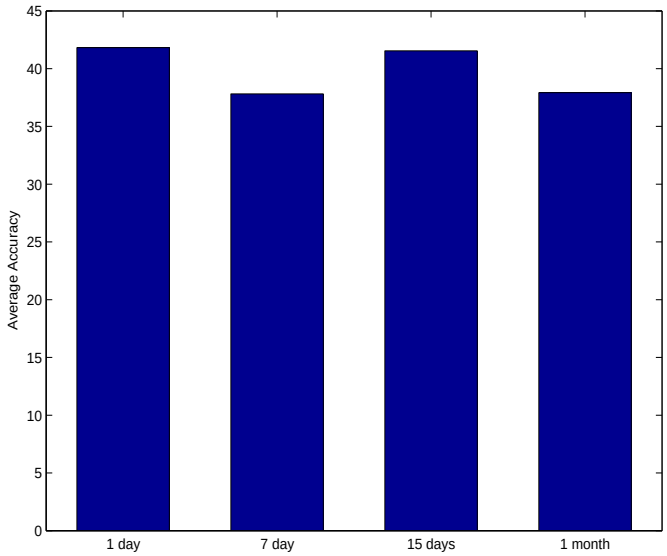

Figure 3.6: Importance of History.

reason why the accuracy for these two numbers (one day and fifteen days) is high is, however, unknown. But, this result gave a few insights. First, to predict the future behavior of a person, it is more plausible to look into the recent activity rather than taking into account the entire historical data. This is because the interest to participate in an activity can change over time. Thus, it is more practical to look into the recent participation habits. Second, this process also has computational advantages. Mining the data to look deep into historical values takes lot of computation time, for example mining the last five years of data. Moreover, as the process is not expected to yield good results, therefore, it not logical to proceed this way. Thus, we recommend using more recent activity for predicting the future participation habits.

3.4 Summary

In this Chapter, we revisited the topic of mathematical models capturing human behavior and their capability to predict future participation habits. We proposed a probabilistic method to recruit a candidate to participate in crowd oriented exercises. We focused on a human centric recruitment algorithm and employed data engineering. We dug into statistics and proposed a framework to select a potential candidate so as to maximize the probability of getting a response. We used concepts from Probability theory to engineer the model. Further, the underlying parameters of the model were estimated via Jeffry's prior. The efficacy of the proposed method was validated by experimenting on real world datasets. Through numerical simulations, we found that the method showed good performance.

Chapter 4

How to Promote User Participation? Applying Human Psychology to Understand and Promote User Activity

In the previous chapter, we looked at the participation habits of an individual to propose a worker recruitment model. The drawback of the work in that chapter was its inability of look at the other side of the coin. That is, the framework was skewed towards machine learning. In this chapter, we look into human psychology and try to find a few reasons that could show us a potential direction to answer the following questions: Why would people want to help other people? How to devise unconventional psycho-techno techniques to enhance user participation? In doing so, we present a few set of recommendations that could help us understand different ways of generating interest and maximizing the chances of getting a response. The study carried out in this chapter reveals a few interesting and otherwise neglected observations about typical human mentality. In this regard, and before going into the details, we quickly summarize the contribution of this chapter in the following points.

1. Mature crowdsourcing repositories (provided by StackOverflow) are mined in order to find new and otherwise unnoticed patterns in the crowd's behavior. This is done to present details that could further improve the participation of users.
2. The foundation of the chapter is laid down in well tested psychological theories. We try to explore the core reason behind the presented observation is explored by looking deep into psychology.
3. Based on the findings discussed in the chapter, a few recommendations are presented to the requester as well as to a potential online platform.
4. The chapter tries to find explanations and alternate methods to motivate, encourage, and enhance user participation.

The systematic workflow with respect to the content discussed in the chapter is presented in the following points.

1. As crowd oriented systems are dependent upon the efforts and motivation of the volunteer, therefore, the chapter starts by exploring the psychology behind Motivation and Volunteerism.
2. After explaining the psychological principles, data repositories provided by StackOverflow are mined.

3. The chapter then explores, studies, and presents eight different types of dimensions. For each of the discussed dimensions, data is collected and the reason behind the observed behavior is analyzed by looking into human psychology.

4. Based on the different dimensions explored in this chapter, a set of recommendations is proposed. The aim is to advise a potential requester as well as the platform to try and look at humans differently.

Before beginning the discussion on the proposed work, we must point it out here that the work presented in this chapter focuses mostly on open-call based crowd sourcing and on Stack-Oveflow. The study of the human attributes and the idea presented here, however, can be generalized to different platforms as well.

4.1 Understanding Motivation and Voluntarism

Before beginning the discussion, several fundamental reasons that could urge people to participate in crowd based exercises are presented. The discussion presented here starts by looking at the psychological principles dictating the behavior of an individual under the factor of motivation.

In psychological literature, it is a well accepted notion that motivation of a person is categorized into: Intrinsic and Extrinsic Motivation [137]. Intrinsic motivation allows individuals to engage in activities for the purpose of pleasure and satisfaction. That is, motivation without seeking any reward or incentive [38]. Though stimulating thee pleasure centre of the brain could be categorized as Intrinsic Motivation, we emphasize more on the extrinsic rewards. Further, the notion of Intrinsic Motivation is sub-categorised into: to know, to accomplish, and to experience stimulation [138]. To know, gives the pleasure in learning and exploring something new. To accomplish gives the satisfaction of excelling in some activity. To experience stimulation involves the pleasure obtained in experiencing fun, excitement, and positive stimulation of senses. On the other hand, extrinsic motivation deals with the motivation of a person to get involved in an activity for the mere purpose of a reward. Similar to Intrinsic Motivation, Extrinsic Motivation is also a multidimensional construct characterized by external regulation, introjection, and identification [139]. External regulation is one of the most common types of extrinsic motivation found in people. It includes engaging in an activity to gain reward or to avoid punishment. The next category, introjection, implies involving in an activity with a somewhat 'Internalized' feeling that causes the self to become more involved. Though the category is extrinsic, it presents an illusion that the actions are caused owing to internal intentions. For example Introjection of the form: I write as I would like people to think that I am a good writer. In identification, the actions are performed because the behavior is more valued and is

considered important. Although, the actions are performed by extrinsic factors, the individual is convinced that the intention is a part of the self and is caused by a goal-directed behavior.

After describing the motivational categories of the human psyche, let's discuss the factors that govern the participation of people as volunteers. As online crowd based systems involve recruiting volunteers, it is necessary to study the factors causing such type of a behavior. To this end, one of the most influential attempts to identify factors governing the participation of people in volunteering, Clary et *al.* [140] came up with six different categories.

1. *Values*. Participating in crowd oriented exercises give people ample opportunities to express their altruistic nature and show their concern for society. Further, people who give preference to values have often been found to show concern for others. For instance, considering the case of the lost child [141], people favoring values will certainly participate in helping others in finding their lost child.

2. *Understanding*. Through participating in crowd activities, some may expect to see a different perspective on a few things. It provides them with a new experience, and a chance to practice their skills that may otherwise go unused. For instance, the Secret London[1] project saw 150,000 members within two weeks with Londoners sharing suggestions and photos of London. The project gave people an opportunity to gain new experience and practice their urban skills.

3. *Social*. The next category deals with the social behavior of a person. This category is closely related to the helping nature of a person and the capability of a person to comply with social norms to achieve societal objectives. For instance, people favoring *social* will participate in projects that report pollution in their surroundings.

4. *Career*. This dimension deals with the long term career goals of a participant.

5. *Protective*. This category includes protecting one's ego from the negative features of the self. Further, it may serve as an alternate to reduce one's guilt or one's personal problems. This dimension may or may not be a significant factor contributing towards user participation in online crowd based events.

6. *Enhancement*. This dimension, similar to the protective category, deals with the ego's relation with the affect. The dimension deals with the effect of postive or negative mood in helpfulness. For instance, when in good mood, one would like to maintain the current state of mind by helping others. On the other hand, when in a negative state of mind, the participation of a person is uncertain.

[1] https://www.facebook.com/groups/259068995911/

4.2 Observing and Understanding Human Behavior

In this section, we draw inspiration from one of the most mature platforms in crowdsourcing to discuss some of the factors that could guide a potential requester and an organization towards engineering a better and an affective crowd based system. Accurately, this section analyzes StackOverflow in detail. In StackOverflow, there are more than 9 Million questions, 16 Million answers, 4.5 Million registered users, and has seen 8 successful years. Even though the numbers are convincing, one can still question the choice of this testbed. To answer this, we would stress upon the importance of the fact that we are trying to understand human participation level in the real world. Second, the motive here is to learn and understand human behavior, therefore, it is reasonable to study one of the most widely used commercial platforms employed by human beings. Therefore, the purpose of studying the behavior of people in responding to requests is perfectly aligned with this mature system. To justify the choice of this platform, a few lines from the users of StackOverflow are presented here:

---

1. My participation in Stack Exchange is because it's fun, enjoyable, and contributes to me professionally. **User - 1048539**. This shows experiencing stimulation (Intrinsic Motivation) and demonstrate the *career* dimension of Volunteerism

2. At the end of the day, StackOverflow is about helping people find answers to their questions. **User - 1394393**. This example shows *Society* and *Value* dimension of Volunteerism.

To begin with the goal of objectively understanding the behavior of people in crowd based systems, several informal emails were sent to StackOverflow users. Further, the StackOverflow authorities were requested to provide with some user information. However, we obtained a very few replies, and owing to privacy concerns the request to the authorities was denied. A few replies however directed us to the StackOverflow community discussion forum where similar to StackOverflow's main website, the community is very active. There are 132K users, 73K questions, 114K answers, 572K comments, and similar to stackoverflow it has been running for more than eight years. This forum, run by StackOverflow users themselves, gave us ample opportunity to analyze people's mood, their anger, their disappointment, their motivation, and much more. Therefore, the discussion forums was manually crawled to understand the problems and find answers to these problems. It should be noted here that the discussion forums can sometimes be subjective, therefore, to find evidence on the problems, data was collected online from StackOverflow's databases[2]. Owing to the lack of a system in literature to quantify

[2]http://data.stackexchange.com/stackoverflow/query/new

the quality of an answer, the *score* (of the answer) was assumed as a quality cue. Possibly inaccurate, the rationale here is backed by the work of Gantayat et *al.* [142] where the authors found that 81% of answers with the highest score (Score=Upvotes-Downvotes) were accepted as the best answers.

Considering the case of StackOverflow, a few statistics on user participation are shown in Table 4.1. In this Table, users have been divided into six categories as per their reputation (StackOverflow works on a reputation basis). It is visible from the Table that people with high reputation respond more in terms of both quality and quantity, whereas, people with low reputation flood the platform with low quality responses. This observation is evident by the fact that an average high end user answers 1185 questions with an average score of 3.62/response, on the other hand, a low reputation user has a score of 1 per response and on an average has responded only once so far. These numbers clearly indicate the superior participation level and the quality of the answers posted online by high end users. Therefore, ideally and looking at the numbers in Table 4.1, one may form a conjecture that recruiting high end users is the best way forward. However, in the subsequent sections, we present a few observations that negates this proposition.

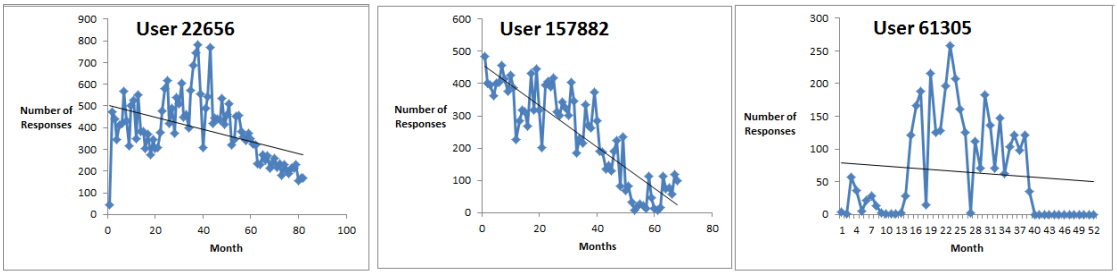

Figure 4.1: Number of Responses per Month

4.3 Dimensionality of Work and Selection of Workers

Before beginning the discussion in this subsection, we present how high end users responded to questions posted on StackOverflow. The figures are shown in Fig. 4.1. In the Figure, the number of responses per month beginning from 2008-to-date is shown for a few users. It is visible that in the initial phase the number of responses increased and the curve experienced a positive slope, consequently, it could be said that user participation increased. However, the behavior did not stood the test of time, and after attaining a maxima, the curve experienced a negative slope. In fact, in the later stages, and for a few users, the number of responses came down to *zero*. This fact raises a critical question, if the users were initially highly motivated, if they were participating heavily, what caused this type of a degrading behavior? To find the answer, let's look at a few dimensions of the work requested by requesters.

Table 4.1: Statistics of StackOverflow

Reputation	No. of Responses	Score of Responses	Score/Response	No. of Users	Response/Users	Score/User	Avg Age
1-to-1000	**4791531**	5373836	1.121	**4313102**	1.110	1.245	29
1001-5000	4045707	8184301	2.022	63644	63.567	128.595	32
5001-10000	1690988	4369711	2.584	9106	185.700	479.871	34
10001-15000	933864	2612674	2.797	2943	317.312	887.758	35
15001-20000	588887	1785161	3.031	1386	424.889	1287.994	35
$>$ 20000	3919865	**14393964**	**3.672**	3306	**1185.682**	**4353.891**	35

Table 4.2: Year Wise Score of Questions

Year	Avg Score of Questions
2008	19.946090368
2009	6.7616608685
2010	3.8913057356
2011	2.6143824707
2012	1.734857727
2013	1.0870153938
2014	0.5559946298
2015	0.3807140762

The first step in this direction is the attribute of '*Quality of Work*'. Quality of Work has always been a construct that has attracted decades of research in psychology [59]. It is studied many times that bigger the challenge, higher the effort and greater the reward. But, in situations where there is no challenge, there is little effort. In this context, if a person of high intellect is asked to perform mediocre tasks, then the dimensions of intrinsic motivation (to know and to accomplish) are undermined and no amount of incentives can help maintain an increased level of participation. During analysis, it was observed that the average quality of questions posted on StackOverflow has reduced. This is evidenced in Table 4.2. High end users therefore feel they are not learning anything new. As a result, they feel bored and uninterested. To exemplify this, few words of such users are quoted here.

---

1) ...I have almost 114K on StackOverflow, but my last answer was on February 6th. And it will be my last answer.- ***User 61305****.*

*2)...a huge mess of low quality questions with a doubled up unanswered rate. "You deserve what you give" -****User 157882****.*

The line "*you deserve what you give"* clearly shows the disappointment of people when faced with this situation. The users gave their best, spent their personal and valuable time on StackOverflow answering questions (for instance user 157882 answered 24.04 % of ALL question in JSF, highest being 43.45% in 2010 and lowest 4.79% in 2014). But, their efforts started going in vain when the work was below par and didn't match up to their expectations. Furthermore, the statement "*...but my last answer was on February 6th...*" signals the lack of interest in such exercises. Therefore, there is no doubt that the participation of such users is in jeopardy.

Though, Quality of Work is a powerful construct that can regulate the participation of users,

another dimension that is equally important is duplicate and redundant. In an endeavour to find new patterns regarding the neglected human factors, it was found that most people participated in crowd based activities because of intrinsic motivation. It was observed that the attribute of Intrinsic Motivation, *to Know*, gave most people the pleasure to explore and engage in something new. Further, the attribute - *to experience stimulation*, gave people an opportunity to demonstrate their *smartness* and *problem solving* skills at StackOverflow. However, an anomaly occurred when the whole purpose of these dimensions was undermined. This happened in situations when people were asked to do the same task over and over again. We observed that this process compromised people's intrinsic motivation, and thus, according to the psychological principle discussed in [26], no amount of extrinsic motivation can fix this situation. To justify this statement, we present the comments of a few users.

---

1) I'd just noticed how much my participation has dropped of late. For me, like many, I've grown a little bored of the same questions over and over again. ***User 97337***

2) ...because people ask the same questions over and over, and people get tired of answering the same questions over and over.. ***User 16587***

Therefore, combining these two dimensions of work, the lesson to learn here is when dealing with humans we have to take care of their intellectual capability and have to respect the threshold of duplicate requests. However, the analysis also revealed that people with low reputation kept answering questions, albeit the quality of answers was not good. Evidence is presented in Table 4.3. In this Table, the questions are divided according to their score. Subsequently, the quantity & quality of answers is cross-referenced by comparing people of different reputations. As visible, low reputation users answer in bulk in all categories, but lack quality. Therefore, in real situations, we cannot favor a particular class of people. A requester has to maintain a trade-off between the number of users recruited, from each class, to get the job done.

Lesson 1. If the aim of system designers and engineers is to engineer a better crowd platform, the recruitment procedure must respect the expectations of volunteers. [143] specifies that crowd forums leverage human intelligence. Therefore, considering human intelligence and learning from the different dimensions of work, it is clear that a requester should try and outsource tasks to people according to their expectations, and at much the same time, try not to send out duplicate requests over and over again to the same person. Moreover, a requester must mix up people with different reputations to get responses both in quantity and quality. Therefore, considering the feasibility and contextual requirement of the requested task, it is recommended that a requester mix different people of various class to get better results and

Table 4.3: Reputation Wise Quantity and Quality. Vertical Axis - Question Score, Horizontal Axis - Reputation

			Quantity							Quality		
	1-1000	1001-5000	5001-10000	10001-15000	15001-20000	>20000	1-1000	1001-5000	5001-10000	10001-15000	15001-20000	>20000
1 – 50	**2390316**	2158175	973673	544663	354264	**2390203**	1	2	2	2	3	**3**
51-10	**44457**	29854	12017	6553	4224	25433	5	23	26	27	29	**37**
101-150	**15677**	9923	4146	2284	1431	7887	5	31	40	42	45	**59**
151-200	**7761**	4926	2074	1881	678	3601	5	33	52	58	58	**76**
201-250	**4646**	2907	1141	600	361	1947	5	33	77	80	81	**98**
251-300	**3346**	2030	857	432	286	1486	6	30	87	77	71	**110**
301-350	**2349**	1510	624	360	226	1045	8	29	82	98	82	**118**
351-400	**1578**	1035	425	189	146	688	5	23	26	27	29	**37**
>400	**7840**	5381	2064	1133	691	3363	11	38	98	142	161	**283**

good participation.

4.4 Spatial Characteristics: Role of Nationality and Neighborhood

In this section, the spatial characteristics of a human being is explored to analyze people' behavior according to their respective locations. To do that, the nationality of people participating at StackOverflow was considered. In the approach, the top ten countries whose nationals are ranked the highest in terms of answering questions, quantitatively, were considered. In our dataset, we took USA, UK, France, Germany, India, Russia, Canada, China, Brazil, Netherlands, Italy, Sweden (They rank high in terms of quantity of posts).

To begin the experiment, questions asked in one country were extracted, then the answers provided by people from a different country were cross-referenced. The results for this series of observations are shown in Table 4.4. During analysis, we found that people with the same nationality helped each other more than people with a different nationality. This observation was in contrast to the initial expectation, where it was hypothesized that in online crowd based systems there is no boundary of nationalities. Furthermore, as none of the posted questions had anything to do with neither nationality nor geography, therefore, there was a strong belief in the hypothesis. Consequently, after getting this result, the test was repeated 10 more times by taking different number of subjects. For instance, we tested with 500, 1000, 1500, 2000 and so on users from each country. However, the result was the same. The numbers were always in the favour of the parent country. Nonetheless, after the first round of analysis, the hypothesis failed. However, taking inspiration from this failed hypothesis, additional details were added to the study. In the next series of experiments, people having the same nationality were taken and their behavior towards a person from the same city was analyzed. The dataset consisted of ten cities, ranked quantitatively, for each country. For instance, questions asked in Paris and responded to in Paris were checked. Further, the number of responses given by the rest of the country was analyzed (in this case, France excluding Paris). The result for this series of experiments is shown in Table 4.6 and 4.7 (For reasons of brevity, results for Germany and France are presented). One can see that for every test the number of responses by people from the same city is high. Even though people have the same nationality, but the statistics clearly indicate the importance of the neighborhood and emphasize more on locality [144]. This pattern and the importance of neighborhood has been overlooked by researchers working in crowd based systems, but like a two sided coin, we cannot ignore the importance of these human traits. It must also be pointed out that during our literature study, it was realized that the "*theoretical*" definition of the crowd signifies "*a faceless entity*" [25], but, one has to

Ans by/ Ques by	USA	UK	India	Germany	France	China	Brazil	Netherlands	Russia	Sweden	Italy	Canada
USA	**32526**	16666	16328	6562	2992	1782	2738	2971	2151	2679	2434	9576
UK	18054	**82993**	44603	16562	7680	3471	4594	8058	5266	7247	6044	14224
India	8559	16439	**111842**	7795	4143	3116	2462	3861	2900	3091	4042	6189
Germany	8849	19606	27007	**31698**	4850	2663	2703	4370	3404	3837	3831	7572
France	4074	8825	12818	5050	**13951**	1378	1371	1971	1711	1781	1841	3396
China	865	1322	3239	754	397	**4182**	258	294	346	325	285	732
Brazil	1849	2991	3382	1376	706	310	**7279**	638	434	546	542	1423
Netherlands	4178	9776	12496	4484	2072	939	1445	**11525**	1389	1823	1701	3841
Russia	2420	4619	6797	2524	1234	792	723	1028	**7640**	890	881	1978
Sweden	3006	6704	8294	2789	1330	787	844	1218	947	**8353**	994	2439
Italy	1624	3576	5383	1701	883	469	553	798	552	705	**7209**	1292
Canada	8941	12590	11391	4893	2161	1175	2011	2084	1422	1986	1727	**25930**

Table 4.4: Nation Wise Responses.

US	UK	India	Germany	France	China	Brazil	Netherland	Russia	Sweden	Italy	Canada
73488	41216	74434	23259	14156	15721	13219	11216	11507	8381	8426	18383

Table 4.5: Number of Subjects in Different Countries

understand that to the crowd the requester is not faceless. Therefore, it would be wise if we could objectively use this property to enhance and engineer a better strategy to promote user participation.

Lesson 2. The observed spatial characteristic, concerning human ideology, teaches us that if requesters were to ask people in their vicinity (not limited to city), then the probability of getting a response is much higher than randomly sending out requests to total strangers. Though, it is true that in most cases we cannot expect people from the local vicinity to be present everywhere all the time, to maximize the chances of getting a responses, it is advisable to look into these factors first. Otherwise, the selection of the unknown crowd workers is already defined. Though the statement negates the ideal laboratory experimental mentality, but to handle the real world's real problems, it would be wise to use unbiased and objective approaches.

4.5 Response Time and Quality of Response

The next series of experiments are conducted to find how do human beings respond to requests? For this purpose, the test was conducted in two different ways. First, users were categorized based on their reputation. Then their average time to respond was analyzed. Second, for each user irrespective of reputation, the objective was to find if there is any useful hidden information at all. We start by presenting the results for the first test.

To begin with the setup, users were categorized based on their reputation. Subsequently, the average time to answer was analyzed for each category. The resulting observation for this test is shown in Fig. 4.2. The mean and standard deviation for the observation presented in Fig. 4.2 is 21,837 and 17,631 respectively. As visible, people with low reputation have a low response time, but, it has been shown that the average answers lack quality. Therefore, answering at a whim is not always the best choice, and should not be the way forward. Taking inspiration from this observation, and after learning the response time of people, we attempted to find the answer to the next few questions:

1. What is the best way to respond?
2. What type of a behavior always results in the best quality of work?

To answer these questions, human psychology was explored to find an explanation. Psychology teaches us that the human behavior can be categorized as Passive, Aggressive, and

Table 4.6: City Wise Responses, Germany

	Munich	Berlin	Hamburg	Cologne	Stuttgart	Frankfurt	Karlsruhe	Dresden	Leipzig
Response from Same City	**1485**	**3915**	**857**	**403**	**350**	**357**	**365**	**316**	**152**
Rest of the Country (Germany)	686	1670	448	231	128	242	156	*189*	*77*

Table 4.7: City Wise Responses, France

	Paris	Lyon	Toulouse	Nantes	Grenoble	Bordeaux	Lille	Rennes	Montpellier	Nice
Response from Same City	**3966**	**416**	**100**	**185**	**248**	**144**	**148**	**177**	**124**	**141**
Rest of the Country (France)	697	113	32	47	73	29	55	36	40	52

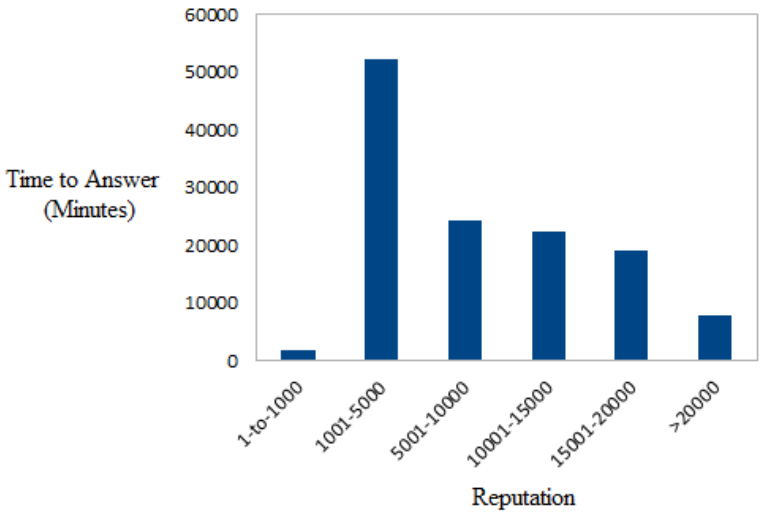

Figure 4.2: Average Time to Answer in Minutes.

Assertive, Submissive. Therefore, a human is classified as any one of them. With respect to this statement, we have to understand that for a particular person, any one of these categories becomes the usual behavior, and others becomes unusual. Considering human nature, it is expected that the usual behavior last for a long time, and unusual behavior doesn't. Therefore, in our case and based on this reasoning, the number of responses under the usual category is high. Taking inspiration from this principle, the response time of a person was divided into two categories. For this purpose, unsupervised machine learning algorithms were used to differentiate the time to answer questions into separate categories for each user individually. For this purpose, the open source machine learning toolkit WEKA[3] was used. It must be pointed out here that the process was done manually, consequently, only 100 users were analyzed. The result and the corresponding classes obtained after this procedure is shown in Fig. 4.3. Result for only one user is presented.

After dividing the answers based on response time, the quality of answers for each of these two classes was calculated. The result for this experiment is shown in Table 4.8. In our venture, we found a pattern. We observed that the score of answers provided under usual category is high and most users followed this behavior (71/100 for this dataset). However, after getting the results, we thought that there might be a bias in the test, therefore, the experiment was repeated by taking more users with an objective to dismiss this analysis (users from different reputation categories were also mixed). But, the result was against the objective, and supported the first observation (86/151 for different dataset, making 157/251). Therefore, this observation led us to realize that people when respond in a usual manner, produce good quality results.

Lesson 3. [24] raises the issue of data quality caused by participation of humans in online

[3]www.cs.waikato.ac.nz/ml/weka

Table 4.8: Performance under Usual and Unusual Behavior

User	893	1043	1053	1097	1190	**1219**	1228	1242	1310	1432	1709	1527	1559
Score in Usual Behavior	**10**	6	**11**	**6.33**	**12**	**12**	6	**5**	**33**	**9.07**	23	**14**	**5**
Score in Unusual Behavior	3	**9**	0	2.4	8	6	**36**	3	9	7.91	**39**	4	2

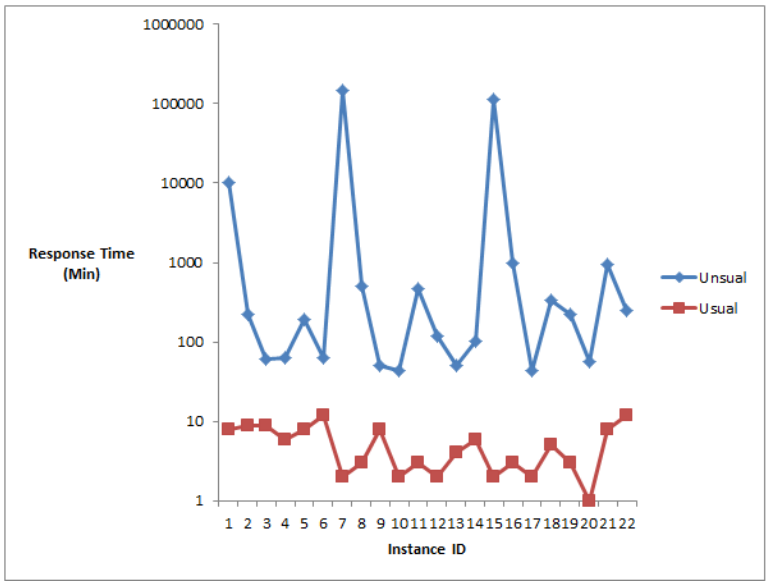

Figure 4.3: Usual and Unusual Behavior (User 893)

forums. Through the observations presented in this subsection, we discovered a previously unexplored pattern. Therefore, to take advantage of this pattern, if a requester wants get the best efforts from a volunteer, it is advisable to recruit those people who frequently provide responses in a usual manner. It should be noted here that a result produced under the usual category might not be the best globally, but, it is the best a person can do.

4.6 The Objective and Behavior after Achieving the Objective

StackOverflow works on a system of reputation and badges, where if a user gives a good answer, he/she is rewarded with good reputation, and after providing several good answers, the person is rewarded with a badge. It was found that when a person became interested in this platform, he/she tried to gain as much reputation as he/she could. Further, users also tried to win badges. To exemplify this, on the profile page of User 106224, the person has proudly written, 1*st user to earn the bronze, silver and gold css3 badges, 1st user to earn the bronze, silver and gold css-selectors badges*, and so on. Similarly, user 100297 and several others have done something similar. Such observation led us to realize that people at StackOverflow take pride in earning these badges. Therefore, it can be generalized that for some users earning badges and gaining high reputation becomes an objective. This statement is supported by the growing interest of literature in gamifying badges at StackOverflow [13], [71]. As a result, and according to goal setting theory [145], people will work hard and will participate more to achieve these objectives. However, work has also found that once people won a badge, they do not maintain

the same efforts, they once put in before winning the badge. The rationale here is backed by the authors of [13] who have specified that - "*...upon reaching their achievement, (users) see no immediate need to continue the labour-intensive task*". This statement, rather this fact in literature, is an anomaly. In other words, if people were participating heavily, if they were putting extra efforts, then what happened after they won a badge? To understand this, lets take a look at human psychology.

To proceed with the problem, one has to answer the following the question: Why was the person interested to participate more? To this end, it can be easily deduced that the person wanted to win a badge. Further, winning a badge is a reward, hence, the motivation was Extrinsic Motivation. Along similar lines, and for any worldly scenario, we have to understand that if people's motivation (either Intrinsic or Extrinsic) is reduced, then the self is not stimulated, and people do not feel the urge to perform. This is common knowledge and the phenomenon has been observed many times. In case of StackOverflow, once a person won the badge, the factor stimulating Extrinsic Motivation was dead, hence, there was no need to participate heavily, hence, a decreased level of engagement. In simple words, if there is no motivation, it is expected that a person will not participate enthusiastically anymore. Moreover, when extrinsically motivated, we have to face other entwined consequences. It has been found many times in literature that when extrinsically motivated, people's moral commitment to help is decreased [146],[147]. In other words, people who are extrinsically motivated, are less likely to help others out of internal desire and in the absence of rewarding mechanisms. This psychological finding is directly applicable to the case in question. Objectively speaking, users wanted to win a badge, therefore, they participated more. Once they got the prize, the factor stimulating Extrinsic Motivation was removed. However, this removal also effected the person's internal desire to participate [146],[147]. Combined with the these psychological facts, one has to face the situation of decreasing participation as discussed in [13], [71]. Consequently, to fix this problem, it is imperative that we find new ways of stimulating users' Extrinsic Motivation. This is especially important considering the fact that stimulating one's internal desire is tough. However, stimulating extrinsic motivational factors is comparatively easy. For example, it is comparatively easier to target the ambitions of an ambitious person, whereas, it is tough to create ambitions in an unambitious person. For crowd based systems, therefore, we should aim to keep offering users highly prestigious and awe inspiring trophies. Therefore, according to the psychological principle discussed in [145], people will put in more efforts. Having said that, it is also true that sooner or later people's desire to engage will decrease, but, if that inevitable situation can be delayed, then in the meantime, we can generate interest and get good participation from users.

Lesson 4. The discussion presented in this subsection points to the importance of providing people with an objective, and the corresponding increment in their performance to achieve it.

However, it was also observed that after achieving the objective, the participation rate saw a drop. The lesson to learn here is that a platform should take note and must offer several challenges to people. People have a natural tendency to take on challenges, therefore, they will try to participate more and will indeed put in extra efforts. However, it is of prime importance that the challenges should be smartly designed to encourage and stimulate the interest of people for a long time, both before and after winning. It should be noted here that badges, reputation etc. are not the only way to offer objectives. Moreover, it is crucial that the process must be 'evolutionary' and the winnings must be awe inspiring, otherwise, the enhanced participation will be dead in some time.

4.7 Importance of History and Association between Worker and Requester

The next category of observations deals with the relationship between the requester and the worker. Like many common human traits, relationship between two individuals is an important criterion that encourage users to put in extra efforts for people they care about (or are known to them). Taking inspiration from this social fact and basing the argument in commitment theory [148], we put this idea to test. In other words, the aim is to find out if there is any pattern from which the importance of this human attribute can be better understood. To this end, the test was conducted in two different ways. The first one is similar to Quid-Pro-Quo, where if a person answered somebody's question, the objective was to see if the requester returned and helped the responders answer his/her question. The second concerns how many people come back to answer questions posted by the same requester.

In the analysis, it was observed that the first test, Quid-Pro-Quo, did not produce fruitful results. In the experiments, participation habits of 10,000 users were analyzed, but the analysis revealed no significant result. This signifies that crowd systems do not respect the principle of Quid-Pro-Quo. Therefore, it is not viable to expect people to come back and reply if one has responded to their request in the past. The second test also did not yield any significant result. For this test, a total of 1,00,000 users were analyzed. Out of 1,00,000 only 6,457 returned to provide an answer for the same requester. Though, it was also observed that a returning user provided an average of 3 answers per requester. But, the numbers are not enough to make a difference. Therefore, it can be said that in crowd based systems, relationships (or history in our case) does not necessarily promote an increased level of participation. Therefore, it is infeasible to rely on this human factor.

Lesson 5. Commitment theory [148] and studies on online commitment towards communi-

Table 4.9: Effect of Profile Photographs.

Requester/Responder	With Photo	Without Photo
With photo	1203770	1042917
Without photo	339570	607323

ties [149] have emphasized on the importance of commitment and people building relationships with online communities. Though, as discussed in [149], this is true for the relationship between people and 'the platform', but through our analysis, we found that at a human level, i.e human-to-human, this is not the case. Any potential crowd based platform, therefore, cannot rely on this fact and expect people, either recruiters or volunteers, to show an obligation or commitment towards each other. Although, this fact is against the commonly held notion of recruiting people with a history, the facts and statistics presented in this subsection negates the notion.

4.8 Importance of Online Profile

In this subsection, the role of online profiles is discussed. It is frequently observed and is very common for people to create profiles at online platforms, e.g. there are many individuals who have created a lengthy and a detailed profile at LinkedIn for the purpose of advancing their careers. Therefore, taking inspiration from this fact, the idea was to test the importance of online profiling. To be specific, experimentation was performed with two hypotheses.

1. What if the effect of a requester's profile in getting a response to his/her posted request?
2. What is the result of a volunteer's profile on his/her response frequency?

To test the two hypotheses, a comparison was done for the response frequency by checking whether the profile has a picture or not. Subsequently, we examined the profile length (length is an indication of details). For this purpose, people who have a profile length of more than 100 were selected.

To start with the test, the result corresponding to the status of a profile photo is shown in Table 4.9. It is visible from the Table that when the profile of a requester has a photograph, the requester get more responses from the crowd. This observation, from a requester's point of view, shows the importance of a simple photograph. Uploading a photograph at one's profile is considered trivial, but, this triviality resulted in a huge difference in the numbers. Taking inspiration from this result, we dug deeper. It is clear that a requester with a photograph gets more response, therefore, the next objective is to see whether the length of the profile matters or not. In this context, if the length of the requester's profile is more than 100, then a total of

Table 4.10: Effect of Crowd's Profile.

With photo and long profile	Without photo and long profile	With photo and short profile	Without photo and short profile
2298897	313350	2017931	929545

8,68,481 responses were obtained, on the other hand, if the length of the profile is less than 100, then 1,772,134 responses were obtained. Therefore, from a requester's perspective, a photograph indeed makes a difference, but, the length doesn't matter. Next, we tested the responder's point of view. The results for this experiment are shown in Table 4.10. It is visible from the table that people who have a profile photograph and have build up their profiles respond more. Building a profile indicates that the person has spent some efforts, and is interested to let others know about his/her competency. Thus, there is a desire to actively participate in crowd based activities.

Lesson 6. From the observations presented in this subsection, two lessons are learned.

1. It is advisable for requesters to upload a photograph at their respective profiles. Further, although, the evidence is against the point, it is also recommended that a requester should spend some effort on writing a few lines on his/her profile page. Though, the details of the profile do not live up to the objective of enhancing participation, it is a wise option to let the opposite party familiarize themselves with one's portrait. Keeping a good online profile initiates an indirect contact between humans, and introduces a sense of familiarity. This is backed by uncertainty reduction theory [150]. Prior information (e.g. through one's profile) about an individual aids the opposite party in minimizing uncertainty in interaction.

2. To maximize the probability of getting more responses, it is recommended to approach individuals that have a healthy online presence. It should be noted here that this statement is not universal, and there are good workers who like to maintain their privacy. But, on a general scale and from the numbers shown in this section, it is recommended to look into these factors first, otherwise, the selection of unfamiliar individuals is already defined and is the last resort.

4.9 The Gender

In the next series of analysis, tests are conducted to see whether gender has a role in modifying the behavior of people? To be specific, the motive is to find out whether people of the same gender prefer responding to each other or whether people act independently of gender. To do that, the names of both males and females were gathered from Carnegie Mellon University's (CMU) name corpus[4]. Using this corpus, StackOverflow users were categorized as Males and Females. Subsequently, we cross-referenced the questions posted by one sex and answered to by the other. On the basis of classification against the CMU's name corpus, we got 2,05,935 males and only 58,052 females participants at StackOverflow. Therefore, several male names were randomly removed from the male corpus to tip the scales. The first dataset had 58,052 females and 57,800 males. The result conducted with this dataset is shown in Table 4.11. It

[4]http://www.cs.cmu.edu/afs/cs/Web/Groups/AI/areas/nlp/corpora/names/

Table 4.11: Role of Gender

Ques by/Answer by	Male	Female
Male	25624	8954
Female	8880	23546

is visible from the table that the number of responses between people of the same gender is more than that between those with the opposite gender. Prior to beginning the experiment, it was hypothesized that online crowd based systems act in a gender neutral manner. However, this result negates the initial hypothesis. Therefore, to thoroughly test this observation, the experiment was repeated ten more times by taking in different number of subjects in each category. We manually and carefully kept similar numbers is both categories (For instance, 16,091 Males and 17,014 females and so on). However, the result was the same. It was found that people preferred answering to questions posted by the same gender. Though, this result might have a skew owing to sampling bias or bias in the name corpus, but in the study, this behavior was observed. Further, the result is analogous to the spatial characteristics of humans, where they preferred answering to a person of the same locality. We, however, will not recommend recruiting people based on their genders as it is not only unethical, but, psychologically and scientifically it also compromises the definition of the faceless crowd.

Lesson 7. The observations in this subsection emphasizes on the importance of interaction between people of the same gender. However, it is not advisable to focus on this recruiting strategy. Further, the platform as well as the requester should follow the rules that respect a faceless crowd. It should be noted here that the observations does not imply that people are sexists in online communities, but sheds some new light on the topic.

4.10 Stimulating Altruism

In the next series of experiments, tests are conducted to check for the importance of request framing. Specifically, the test targets the *Value* dimension of Voluntarism. In this context, it is understood that to initiate a crowd oriented exercise, a requester has to request the volunteer to respond. Therefore, to maximize the probability of getting a response, we should try to frame requests in such a way that stimulate altruism. To be specific, when we write, the procedure, the choice of words, the method of request framing, and the link between adjacent sentences matters. To exemplify this, and to emphasize on the importance of altruism, two application dependent (garbage reporting) statements are presented below.

1. Please send me pictures of garbage inside the university campus.

Table 4.12: Statistics Regarding Politeness

	Polite	Neutral	Impolite
Avg Score of Questions	66.351	293.95	46.58
Avg no. of Answers	8.556	10.904	6.79
Avg. Score of Answers	13.52	13.72	13.51

2. We are trying to clean our *dirty* campus to make it a better place to study. Therefore, Nature's club request you and it would be much appreciated, if you could please send a few pictures of garbage in our, with your help a soon to be lovely, campus :)

Though, the two statements reflect the same idea and focus on the same problem, but the way the second statement is framed captures attention. It has elements of narration (reason behind the request was given), human connection (by using the word *Our Campus*), sentiment (using words like 'Dirty' and 'soon to be lovely'), length (signaling the efforts put in to form the statement), politeness (using words like request and appreciated), result (a clean campus), dependency (with your help), social status of requester (Nature's Club), humor (using :)) and several others. It has been investigated in social psychology that the way a request is framed stimulates the altruistic nature of humans. Therefore, to test this, and along the lines of the work of Althoff et *al.* [151] on Reddit, a test was conducted for politeness. In the test, 100 questions were taken and their politeness was checked via Stanford's Politeness API[5]. Out of 100 questions, 24 were classified as impolite, 20 were neutral, and 46 were polite.

During analysis, it was observed that politeness and neutrality of the questions resulted in an increase in the number of answers, and the question getting better score. But, it was also observed the politeness had no effect on the quality of the answers, evidence shown in Table 4.12. Therefore, the system is getting more responses but there is no major effect on quality. Nevertheless, this is acceptable as we are getting something from either quality or quantity. However, work in this area is often limited by the important question - "How to frame a request with what one is asking for?" Though, this is indeed a challenging issue, the observations presented in this subsection points to the importance of framing requests appropriately. Further, learning from the discussion in [152], it is advisable to avoid the trap of misleading words. For example, it is recommended to use statements like "Could you please send me pictures for the ABC project" rather than using "Please send me pictures for the ABC project". The first statement is polite, whereas, the second statement is not much polite.

Lesson 8. If a requester wants to increase the probability of getting a response from the crowd, then it is advisable to frame requests that stimulates the altruistic nature of volunteers. From a computational point of view, human beings in an online platform are treated as a group

[5]http://politeness.mpi-sws.org/

of discrete computational units, however, we have to understand that in crowd based systems we are dealing with *non-mechanical* entities. Thus, it is beneficial to target properties unique to these non-mechanical entities. Throughout the chapter, it has been pointed out that a request is sent to a person (in the crowd), and it is the human that is responsible for generating the data. Therefore, it is recommended to put in a few extra efforts to make the request more polite and altruistic.

4.11 Discussion and Threats to Validity

The discussion in this chapter merely scratches the surface on investigating the humanistic attributes at online crowd based systems. Through the analysis, a few lessons were learned. Some of the finding were good, a few, however, showed a different side of people. In theory, we assume that the ideas expressed on the platform are neutral and free of bias. The discussion in this chapter, however, brings forth the fact that the reality is far from this utopian assumption. There is much bias and one needs to be pragmatic and realistic in one's expectations from such online fora.

Before we proceed any further, it must be pointed out that it is not claimed here that the observations presented in this chapter are universal. That is, they need not be applicable to every individual. Although, the work attempted to bring into light a few neglected issues at crowd based forums so that additional constructive work can be built upon, we must emphasize on the point that the observations should not be taken as the final choices. More efforts and more analysis is needed to go into the details of the raised issues. In this regard, emphasis must be given on a few shortcomings with the study conducted in this chapter. They are as follows:

1. The discussion in this chapter is based on analysis performed on StackOverflow. The statistics presented here are *descriptive statistics*. Our motive here is not to discuss the How/When/Why of the observations discussed in this paper. The objective here is to highlight a few issues that could be further explored upon. For example, by drawing inspiration from any of the discussed observations, one can take a survey to confirm vs contradict the ideas by taking in the actual data "*directly*" from the users. We presented the patterns by analyzing the data obtained from the platform. However, a survey would definitely help understand the observations from users' point of view.

2. The next issue is the validity of the numbers presented in the paper. We extracted the data from StackOverflow datasets and presented the observations that support vs contradict the original hypotheses. In this regard, we must point out that the numbers presented in the paper are mere observations. We could not conduct statistical testing to prove the existence of such patterns. This is because the data that we worked upon is cleansed for privacy. Any trace that could leak a user's privacy is removed. This compromises on the availability of data (The public dataset released by StackOverflow is not comprehensive).

Hence, we could not conduct statistical tests on the "*public*" data provided by StackOverflow. Despite that, we worked with the constraints and found the necessary numbers by following the procedure discussed in Section 4. We must highlight here that the motive of the work presented here is to bring the neglected factors into the community so that additional constructive work can be built upon. Moreover, the guidelines presented here can help future work to draw inspiration and conduct an in-depth analysis on each (or some) of the topics discussed in the paper.

3. In this paper, a few observations were discussed in detail. However, an issue with the observations is: we cannot statistically prove causation. In other words, we cannot prove the relationship between correlation and causation. Though, the numbers presented here show the case for correlation, we cannot statistically prove causation. This therefore is the next shortcoming with the work presented in this paper.

4. The methodology followed in this article is similar to quasi experimental design. Hence, it is not claimed here that the observations discussed in the paper are universal and are applicable to each and every individual. Moreover, we conducted the tests on StackOverflow datasets *only*. Though, for the users of StackOverflow included in the datasets, the observations are correct, we do not claim that the observations are true for every platform and scenario.

5. The subsequent issue is with the observation discussed in Section 4.4. We discussed that users preferred responding to people who belong to the same locality. To uncover this, we needed to classify users according to their location. In this respect, we must point out that it is not mandatory for the users of StackOverflow to display their respective locations. Hence, the analysis presented in Section 4.4 does not include all users of StackOverflow. Despite that, we found precedent in literature that confirms this fact from a sociological point of view [144]. In this article, we found this pattern in online crowd platforms. In such platforms, people do not have to be present in person. This is an important result discussed in the chapter. However, we must point out that we have not included all the users of StackOverflow.

6. The subsequent threat is in Gender based participation. Similar to point 1, it must be pointed out that all users of StackOverflow could not be included in the analysis. This is because some users have display name like NPE (userid 367273). It is therefore difficult to classify such users as either males or females. Moreover, classifying users as Males or Females on the basis of their names is by definition biased (for example Simon can be used for both Males and Females). As pointed out in Section 4.9, the test was conducted 10 times by taking in different number of test subjects from the dataset. But, it is not claimed here that the observation is true for the entire user base of StackOverflow. Moreover, it is also not claimed here that there is Sexism at StackOverflow. To confirm the existence of such a pattern, one needs to include the entire users of StackOverflow. This, however, is easier said than done. This is because one has to classify all users into males and females accurately.

Despite the shortcomings highlighted in the previous points, the chapter attempted to present the linking step between the studies in crowdsourcing and the humanistic (and psychological) point of view of users at such forums. Moreover, this chapter offers a few potential research

questions that needs to be answered in crowd forums. Though, the analysis conducted in the chapter was limited to StackOverflow, the existence of such patterns and the observations presented in this chapter can teach us to have realistic expectations from other forums as well.

4.12 Summary

In this chapter, StackOverflow was observed to understand and learn a few valuable lessons. To achieve the primary objective of engineering a better crowd based platform, human behavior was analyzed and several important observations were looked at. Evidence was presented that the 'worker' cannot be treated as a mechanical entity and must be handled as an *individual* of the human crowd. To understand human behavior in online crowd based systems, actions of individuals for crowdsourcing were studied in detail. Further, analysis on the dataset provided by StackOverflow revealed a few interesting facts. Based on the observations, a few recommendations were proposed that could help engineer an effective human dependent computational platform. The recommendations presented in the chapter are briefly summarized in the following points.

1. We were able to explore various dimensions of work to present evidence that compels us to respect the intellect of people. Through the work presented in the chapter, we discussed that we should not send duplicate requests to the same volunteer over and over again. Also, to get a response, in both quality and quantity, we resented numerical facts that showed that we should mix different classes of volunteers to get responses in quality and quantity.

2. We studied the behavior of people according to their locations in detail. We found that people preferred responding to a person from the same locality more. Therefore, to maximize the chances of getting a response, we recommended that volunteers with similar spatial characteristics (for instance, same city) be approached more often. Otherwise, the selection of a "stranger" is already defined and is the last resort.

3. For the next series of findings, we analyzed the response patterns of people based on time to answer. We found that people responding in a manner that is natural and usual, in terms of time to answer, gave good quality responses. Thus, if one wants to get the best efforts from a recruit, it is beneficial to approach users who provide responses with a natural and usual time lag.

4. We studied the behavior of users both before and after "winning" a challenge. Based on the discussion in the chapter, we recommended making the "challenge" as well as the reward evolutionary in nature.

5. We further presented evidence behind the fact that having a history between a requester

and a worker does not guarantee a professional commitment in the future. It is beneficial not to rely on the shared history between two people to initiate the procedure of candidate recruitment.

6. We also recommended that a requester should build his/her profile at a potential crowd platform. The presence of an online profile initiates a sense of familiarity between the requester and the responder. Further, through the analysis presented in the thesis, we also advised that those recruits be approached who have a healthy online presence.

7. The analysis conducted on StackOverflow also revealed that gender plays an important role in crowd based systems. We found that people preferred a person of the same gender more. However, we recommend not to focus on this recruitment strategy as it not only unethical but it scientifically violates the definition of the “faceless" crowd.

8. Lastly, to maximize the number of responses from the crowd, we recommended that requests be framed appropriately. That is, there should be attempts to make requests more polite. We found that politeness results in a request getting more number of responses. But, we also found that the quality of responses stays unaffected irrespective of the degree of politeness.

Chapter 5

Quantifying Uncertainty in the Internal Mental States to Predict Human Interest: An Application of Psychology and Machine Learning

In the previous chapters, we studied human behavior from a statistical and psychological point of view. The aim of the previous two chapters was to understand the humanistic attributes and generate interest in a user to participate more. In this chapter, we shall go one step ahead. We focus our attention on quantifying the ostensibly unquantifiable property of human interest. Here, the aim us to model interest through the use of computational methods. That is, we try to answer the following question: How "*much*" are you interested? To do this, we propose the design of a system that can estimate the intangible property of human *interest*. This is one of the non-trivial problems of literature. In this chapter, we present a potential solution for this issue. To do that, we assume that interest in an any entity stimulates the person to takes actions. Furthermore, it is a well tested theory that interest and activity do not occur in a vacuum [40], [41]. It was specified in section 1.1.1 that the idea here is backed by the authors of [39] who have specified that - "*activity plays a significant role in the patterns of human behavior, which is a consequence of interest oriented human activity*" In lay terms, a person is compelled to take actions. The motive here is to use statistical models to quantify the hidden phenomenon that promotes activity. In this regard, we begin by highlighting the challenges we have to face while modeling this unique internal mental property. Subsequently, we provide a detailed account of the method proposed in this thesis. We start with the challenges.

5.1 Challenges for Predicting Human Interest

C1. How to measure interest? It is understood that interest is a property that can not even be computed accurately by a human. Therefore, expecting a mechanical object to measure it, is challenging. What would be the modus operandi for measuring interest? Should we mount complex head gears on normal people, read their brain waves, and expect that the contraption will give us a good estimate? This strategy, especially in the context of ordinary people outside a laboratory environment is not only infeasible, but practically debatable.

C2. It has been specified in literature that interest is representative of a person's actions and quantifies the phenomenon that provokes activity [28]. A brief reflection on the term activity,

however, clearly implies that activity is hardly a unitary construct. People have always been able to express their actions through different perspectives. To exemplify this, if a person is interested in Swimming, for example, the possible perspectives of activity are: The numbers of hours spent swimming, the number of hours spent training one's muscles, the number of hours spent learning new strategies, the number swimming sessions in a day, and so on. A straightforward implication of this dialectical viewpoint leads to fractionation of activity into several sub-features that converges towards the notion (activity) being characterized as an abstract concept. Moreover, looking from this vantage point, it could also be concluded that the sub-elements of activity are different for every application or object of interest. For example, the perspectives of activity in the case where a person is interested in working at Amazon Mechanical Turk is different to the case where the person was interested in swimming. As a result, the second challenge is to find a computationally feasible definition for activity.

C3. In one's daily routine, there are a variety of social factors that flourish versus forestall interest. This is one of the most frequent and naturally occurring phenomena of daily import. Though the circumstances that encourage versus subvert interest is a function of dynamic and every day erratic circumstances, this nevertheless induces a change in one's interest that so far cannot be captured computationally. For example, a person was highly interested in participating at StackOverflow, for instance, but with time interest decreased and the person felt more and more reluctant to participate (for any reasons). In this use case, though the object of interest did not change, the amount of interest certainly did. With respect to the idea behind this use case, the next challenge is to find a statistically feasible definition that can account for this change in interest. That is, how to model the long term evolution of interest?

C4. From the previous two points it is clear that interest evolves with time, moreover, interest stimulates the self to take actions. However, a pertinent question while considering the previous two points simultaneously is: *How did interest converted into activity*? Considering the example in the previous point, how to map the phenomenon that converted interest into activity and model it through computational procedures? Thus, the last challenge is to engineer a method that can statistically define the transformation of interest into activity.

5.2 Broad Overview of the Approach

The challenges outlined in the previous Section have elements of both practical significance and theoretical import, especially considering the fact that the overall aim of this chapter is to model an internal mental property through computational approaches. The motive of the work presented in this chapter, therefore, is to outline a general contour that can address each of these issues, and in doing so, help engineer a meaningful framework to understand the evolution of

interest. By articulating a set of statistical procedures concerning how each of the previous four issues are handled, the aim here is to present a potential roadmap that could facilitate the understanding of interest and its corresponding interpretation by an artificial computational agent. Though the challenge of modelling and quantifying interest is indeed non-trivial, the goal is merely to discuss a few statistical guidelines and theoretical tenets that could then be further explored and applied to a variety of future research endeavours. Therefore, with this motivation in mind, the contribution of this chapter is highlighted in the following points:

1. The problem of predicting interest is formulated as a hidden state estimation problem. Subsequently, fundamental principles of Bayesian Inference are used to deduce interest indirectly from activity.

2. To address the issue highlighted in C2, a subjective-objective technique is used to combine several viewpoints of activity into a computationally feasible construct. In doing so, the subjective and the objective nature of a person is used simultaneously to design the procedure.

3. To tackle the issue highlighted in C3, interest is modelled as the Ornstein-Uhlenbeck process in Physics. Subsequently, inspiration is drawn from Economics and Stochastic Volatility models are employed to improve the performance of the base model. Moreover, further exploration is performed by varying the convergence speed of the process.

4. To address C4, concepts from adaptive filtering are used and Recursive Least Mean Square algorithm is employed to represent the transformation dynamics of interest resulting into activity.

5. To provide a solution to the problem, Monte Carlo simulations through particle filters are employed. Furthermore, numerical experiments are conducted on StackOverflow datasets to present a practically feasible solution.

Before we begin the discussion on the proposed framework, we must reiterate here that interest is an intangible variable that we are trying to quantify via machine-driven algorithms. Therefore, the notion (interest), as presented in the paper, is expected to be imprecise. It is not claimed here that the idea of interest covers the entire theoretical array of literature. However, the phenomenon is of practical importance. Interest in any entity causes a person to takes action [40], [41]. Moreover, interest is inseparably entwined with activity. Drawing inspiration from this basic fact, we make an attempt to assess and evaluate interest using model-based approaches. We work with the presumption that a person is interested in an object and is compelled to take actions. The goal therefore is to employ data-driven procedures to estimate hidden phenomenon provoking activity. Moreover, by following guidelines in Uncertainty Quantification and Machine learning, we do this by presenting a framework that can model the long-term evolution of interest [127], [132], [126].

5.3 Proposed Framework

5.3.1 Predicting Interest via Activity: An Application of Bayesian Statistics

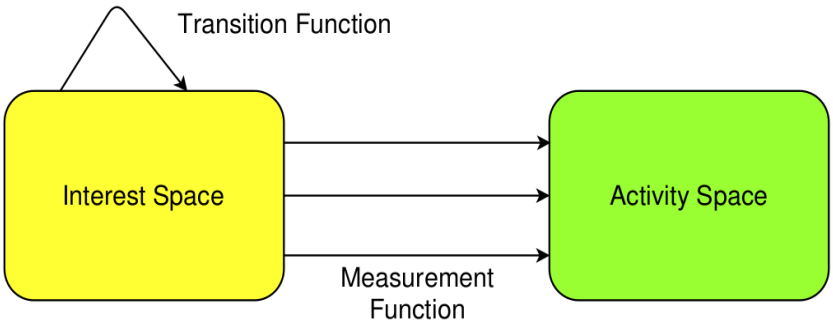

Figure 5.1: Bayesian Inference for Predicting Interest.

In the previous section it was pointed out that the method to deduce interest is formulated as a hidden state estimation problem. The method draws inspiration from Bayesian Inference. With this mode, the broad idea of the solution is presented in Fig. 5.1. It is visible from the figure that there are two state spaces: 1) The Interest Space and 2) The Activity Space. The interest space consists of all the possible numerical values of interest. Similarly, activity space is the collection of all possible activity values. The goal of Bayesian Inference is to use these two state spaces to define: 1) A function that can update interest values in the interest space. 2) A function that transforms interest into activity. For the former function, interest evolves itself with time. Therefore, this function is responsible for evolving the interest values in the interest space. This function is also called as the *Transition function*. The definition of which is as follows:

$$I_n = T_n(I_{n-1}, \theta_n). \tag{5.1}$$

where, I_n is the interest at the n^{th} unit of time, T_n is the Transition function, θ_n is i.i.d noise. Similar to the transition function, and to produce a functional map that can convert interest into activity, a *measurement function* is needed. In general words, interest stimulates a person to take actions, hence, there is activity. The measurement function is responsible for mapping this event in computational terms. This is represented as:

$$A_n = M_n(I_n, \hbar_n). \tag{5.2}$$

where, A_n is the Activity at n^{th} unit of time, M_n is the Measurement function, $\hbar_n$ is i.i.d process noise. Once the two functions are defined, the next step is use them to estimate interest from activity. To do that, Bayesian Inference relies upon the following two rules: 1) Predict and 2) Update. In order to make a prediction, the method relies upon prior information and uses the so-called Chapman Kolmogrov transition equations for the purpose. It is represented as:

$$P(I_t|A_{t-1},\gamma) = \int_I P(I_t|I_{t-1},A_{t-1},\gamma)P(I_{t-1}|A_{t-1},\gamma)I_{t-1}. \tag{5.3}$$

where, γ is the parameter vector. Once a potential value for interest is predicted, the system moves to the update stage. In this step, the predicted value is updated via the newly fed information about activity. In terms of Bayesian statistics, the aim is to calculate the posterior. This is done using the Bayes rule as:

$$P(I_t|A_t,\gamma) = \frac{P(A_t|I_t,\gamma)P(I_t|A_{t-1},\gamma)}{P(A_t|A_{t-1},\gamma)}. \tag{5.4}$$

where, the denominator is expressed as:

$$P(A_t|A_{t-1},\gamma) = \int_I P(A_t|I_t,A_{1:t-1},\gamma)P(I_t|A_{1:t-1},\gamma)dI_t. \tag{5.5}$$

Once the necessary processing of information is complete, the next step is to filter the numerical samples of interest. To that end, Bayesian Inference problems rely on procedures implementing the Markov Chain Monte Carlo methods. This step is comparatively simple. The issue, however, is in the actual implementation of this discussed *theoretical* framework. In this regard, the following workflow (process steps) is defined to find concrete and computationally feasible definitions of the discussed framework.

I. There is a requirement for a computationally feasible definition of activity.

II. The next requirement is the statistical definition of the transition function.

III. Subsequently, a computationally feasible measurement function is required.

IV. Lastly, a Bayesian filter has to be defined to find numerical estimates of interest from activity.

5.3.2 Process Step I: A Mathematical Definition for Activity

In the beginning of the chapter (In point C2), it was pointed out that activity in general terms, is often treated as a mere singular variable. In reality, however, people are moved to act towards

their object of interest via perspectives that are not limited to a set of congenial attributes. In this subsection, we present a method that can transform this abstract notion into a more concrete and a computationally viable option.

To understand the procedure, lets first focus on a general use case. Lets consider the case where a person (say John) is interested in Instagram (An online photo sharing platform). For this use case, the various modes of activity could be: The length of a login session, the number of photos uploaded, the number of login sessions in a day, the number of photos viewed and so on. A short insight into this use case, and these various facets, immediately reveals the diverse nature of activity. Moreover, the attributes presented in this particular example were limited to a specific platform, Instagram, but when we consider the case where John is interested in StackOverflow, we immediately realize that the attributes in this scenario are completely different to that of Instagram. Consequently, we have to consider the viewpoints on activity for every object of interest separately. To this end, lets consider that for a specific object of interest, E^i, there are a possible ϕ perspectives. This is mathematically expressed using the following equation:

$$A^n = \mathcal{F}(a_1, a_2, ..., a_\phi). \tag{5.6}$$

where A^n is the numerical value for activity at the n^{th} unit of time, a_ϕ is the ϕ^{th} attribute of activity (e.g. the number of photos viewed). With the notion of activity defined, the next goal is to find a procedure that can numerically simulate the function $\mathcal{F}$. In this regard, a reasonable choice is the thoroughly tested weighted approach. This method has a predominant tradition of being applied to work spanning across multiple disciplines [153]. This chapter works along these terms and therefore represent activity using the following equation:

$$\mathcal{F}(a_1, a_2, ..., a_\phi) = \sum_{i=1}^{\phi} w_i a_i. \tag{5.7}$$

where, $w_i \in \{0, 1\}$ and $\sum_{i=1}^{\phi} w_i = 1$. Here, w_i is the weight of the i^{th} perspective of activity.

Algorithm 1 Subjective Weight Calculation

Initialization. Input the pairwise comparison decision matrix. This matrix is also called as the Saaty's Matrix S. Each cell of the matrix should satisfy $S_{kk} = 1$, $S_{jk} > 0$, $S_{jk} = \frac{1}{S_{kj}}$

---

where S_{jk} denotes the preference of the person towards the attribute a_j w.r.t the attribute a_k.

Optimization Problem Formulation. Weights are calculated by minimizing the following optimization problem:

$$\min C = w_T F w = \sum_{i=1}^{n} \sum_{j=1}^{n} (s_{ij} w_j - w_i)^2 \tag{5.8}$$

subject to

$\sum w = 1$ where, $F = [f_{ij}]$ for $i, j = \{1, 2, ..., n\}$

$f_{ii} = n - 2 + \sum_{i=1}^{n} s_{ij}^2$, for $j = \{1, 2, ..., n\}$

and, $f_{ij} = -(s_{ij} + s_{ji})$, for $i, j = \{1, 2, ..., n\}$

Solving by non-linear programming. Solving the above optimization problem leads to the following expression

$$w' = F^{-1} e / e^T F^{-1} e \tag{5.9}$$

where, e is the identity matrix, and $w' > 0$, $\sum w' = 1$, w' is the subjective weight matrix.

With this particular definition of the function $\mathcal{F}$, the next logical issue is: How to calculate the value of the weights? From a statistical point of view, weights specify the numerical preference of a user towards a specific attribute of activity. This, in non-technical terms, implies that two people, though interested in the same object, need not show an equal amount of numerical preference towards a specific attribute of activity. For instance, John and Alice both are interested in Facebook (for any reasons). Alice, on the one hand, likes to upload pictures of herself and likes to provide a lot of comments. John, on the other hand, spends a lot of time browsing through his friends' profiles, moreover, the length of every login session is high. To generalize this idea, both John and Alice is interested in Facebook, but the way they expressed their interest is different. Therefore, it is imperative that the procedure consider the '*subjective*' and the personal nature of the person to calculate activity. However, going with the intuitive and sometimes judgmental nature of humans is not the most optimal strategy. Literature has repeatedly specified that clouded by incomplete information and owing to the lack of a judicious procedure, subjectivity alone is often compromised [154], [155]. The discussion here argues that any practical human dependent system must work with "*subjectivity*" and "*objectivity*" simultaneously. The former uses the personal preference of the user, whereas, the latter provides an element of impartiality and logic expected from a computational procedure. As a result, the two points of view are combined and weights are computed via the subjective-objective weighted

approach. The method to implement this strategy is taken from [156]. With respect to this choice [156], it should be noted here that there are many strategies in literature for calculating subjective objective weights. This strategy has been selected as it is computationally inexpensive and uses a programming model that is robust and can very easily be applied to a variety of problems. Moreover, the framework employs ideas from the thoroughly tested weighted least squares method and the objective programming model. These methods are commonly used in a variety of problems across disciplines. Owing to such advantages, the subjective objective weights are computed according to the method discussed in [156]. The methods are summarized in Algorithms 1 and 2. As per this approach, the subjective approach comprises applying the method of the least squares to solve a sequence of algebraic equation, whereas, the objective approach comprises of minimizing the distance between the most optimal and several alternate solutions. In the algorithms discussed below, only the final expressions of the methods are presented. The details and the analytical derivation of the method is not included here. A detailed mathematical analysis (and derivation) of subjective-objective weights falls within Decision Making literature and is therefore neither in the scope of the chapter nor this thesis. The interested reader is referred to [156], [157] for the details. The two methods are combined and activity at any time period t is calculated as:

$$A^t = \gamma \times SM \times AM^t + (1 - \gamma) \times OM \times AM^t. \tag{5.10}$$

where AM is the attribute matrix, OM is the objective weight matrix, SM is the subjective weight matrix, $\gamma \in (0,\ 1)$ is the bias parameter. This equation results in a computationally (numerically) feasible definition of activity.

Algorithm 2 Objective Weight Calculation

Initialization. Input the normalized decision matrix $D = (d_{ij})$, for $i = \{1, 2..., m\}$ and $j = \{1, 2, ..., n\}$.

Transform D into weighted normalized decision matrix $WN = (d_{ij})w_j$, for $i = \{1, 2..., m\}$ and $j = \{1, 2, ..., n\}$.

Define d^*, WN^*, and X as

WN_j^*= $\max\{WN_{1j}, WN_{2j}, ..., WN_{mj}\}$, and

d_j^*= $\max\{d_{1j}, d_{2j}, ..., d_{mj}\}$, and

$X = \{x_1, x_2, ..., x_m\}$, where

$x_i = \sum_{k=1}^{n}(WN_k^* - WN_{ik})^2$, for $i = \{1, 2, ..., n\}$

Optimization Problem Formulation. Weights are calculated by minimizing the following optimization problem:

$$\min w^T G w \tag{5.11}$$

subject to

$e^T w = 1$, and $\sum w = 1$

where, G, a diagonal matrix, is defined as

$g_{ii} = \sum_{j=1}^{m}(d_k^* - d_{jk})^2$, for $k = \{1, 2, ..., n\}$

Solving by non-linear programming. Solving the above optimization problem, leads to the following expression

$$w'' = G^{-1}e/e^T G^{-1}e \tag{5.12}$$

where, e is the identity matrix, and

$w'' > 0$, $\sum w'' = 1$. w'' is the objective weight matrix.

5.3.3 Process Step II: A Stochastic and a Self Evolving Model for Interest

In this subsection, a method to model interest is presented. More specifically, with respect to the systematic workflow of the chapter, this subsection addresses point II highlighted in Section 5.3.1. However, before beginning the discussion, it must be pointed out that owing to lack of literature on modelling the long-term evolution of interest, that is, owing to the lack of statistical guidelines on modelling interest, the development of the function begins with everyday observations. Subsequently, we make a few assumptions. We begin with the observations.

Observation I: **The first point of note about interest is that it is stochastic**. The rationale here is backed by work in analytical psychology where literature examines various internal mental states via stochastic methods [158], [159]. Moreover, stochasticity in interest is a direct implication of the uncertainty in the human routine. If, on the other hand, this is false, then interest is deterministic and we can predict human behaviour at any point in time. In fact, we can estimate every internal mental state with absolute certainty. It can be deduced that this is an anomaly. Therefore, we can conclude that interest is a stochastic process.

Observation II: **Interest is a variable that does not increase indefinitely with time**. Proceeding along similar lines, if this is not the case, then interest is an ever increasing function. However, this is an anomaly as we has experienced that the cycle goes through several ups and downs. An individual does not engage with the object of his/her interest with the same rigour every time. Therefore, interest does not always increase with time.

After discussing the starting ideas about the observations on interest, the following are the assumption made in the model.

Assumption I: **It is assumed that interest is a diffusion process**. That is, it follows Markov property with no jumps. This is a standard assumption in literature dealing with uncertainty quantification and machine learning [127], [132], [133]. The method proposed in this chapter is along this existing notion in literature. This assumption is made for computational advantages.

In light of the discussion and these statistical properties, interest is modelled via the Ornstein-Uhlenbeck (OU) process in Physics [126]. It is a mathematical realization of the Brownian Motion. The process is one of its kind and allows for a linear transformation in space and time. This process has been selected to model interest owing to the following reasons: 1) It matches with the assumptions and the observations of interest discussed in this section. 2) In contrast to similar methods in its category, it's analytical properties are extensively studied and significant research has been dedicated towards understanding its statistical characteristics. Hence, its computational properties are well documented. 3) It has a good discrete (needed to "computationally" simulate any process) representation (It is discussed in the following text). Owing to these properties, interest has been modelled as an OU process. The process in its differential form is represented by the following equation:

$$dI_T = \lambda(\mu - I(T))dt + \sigma dW_t. \tag{5.13}$$

where μ is the mean, σ is the volatility component, λ is the convergence speed, I_T is the interest, dW is the Weiner process. The following points summarize the physical description of the OU process in brief:

i. The equation is a representative of the movement of a particle, e.g. a molecule, in space time.

ii. The movement is random at each interval of time.

iii. The randomness in the motion is controlled by σ, also referred to as volatility.

iv. The speed of the physical particle is denoted by the term λ, also known as convergence speed.

v. The process though moves randomly in space, but it converges to a specific point in space. This point is represented by μ. It is also referred to as the mean. This property is known as *mean reversion*.

vi. At each time step, the instantaneous drift $\lambda(\mu - I(t))$ corresponds to the physical force that pulls the process toward the long term mean μ.

With this understanding, lets find a solution to equation (5.13). Replacing $f(I_t, t)$ as $e^{\lambda t} I_t$ in equation (5.13) and simplifying produces the following final equation:

$$I_t = e^{-\lambda t} I_0 + \mu(1 - e^{-\lambda t}) + \int_0^t \sigma e^{\lambda(h-t)} dW_h. \tag{5.14}$$

With this starting point, let's dig into the OU process. This is owing to the fact that we are making an attempt to deduce the internal property of a human through the use of computational approaches. Specifically, additional efforts are needed to find the discrete model (corresponding to equation (5.14)). A simple way to do this is to apply the Euler-Maruyama Method [160]. In this thesis, however, a more advanced version is explored. Literature in Mathematics and Physics have dedicated much efforts to deduce and understand several computationally feasible statistical properties of the OU process. Therefore, the work follows the discussion in [161], [162], and consequently, the discrete model (corresponding to the OU process) is shown below (This is one of the most extensively studied computational representations of the OU process).

$$I_t = e^{-\lambda t} I_{t-1} + \mu(1 - e^{-\lambda t}) + \sigma\sqrt{(1 - e^{-2\lambda t})/2\lambda}\epsilon_t. \tag{5.15}$$

where, $\epsilon_t \sim \mathcal{N}(0, 1)$, t is the time difference.

The discussion presented so far has highlighted the use of advanced data analytics to model interest. However, as the objective here is to model interest, the dynamics of which are unclear, let's take one more step. Analyzing and correlating interest & the OU process simultaneously reveals two shortcomings. They are as follows:

I. The first drawback with the OU process is the constant value of σ. Recall that σ controls the amount of uncertainty (or randomness) in the process. For human behaviour, uncertainty is a natural consequence that happens when people encounter everyday circumstances. Furthermore, and to a large extent, the apparent variations in a person's daily routine as a function of the dynamic environmental factors or variables leads to continuous uncertain and erratic situations that are sometimes beyond one's control. For instance, any person cannot accurately predict what will happen in the next one or two hours. As a result, to model the stochastic nature of interest via the OU process, *we cannot assume that the amount of randomness entering into a person's interest is always constant*. This is analogous to assuming a constant value of σ in equation (5.13). This is the first shortcoming with the process.

II. The second problem is with the convergence speed, denoted by λ. Similar to the previous point, where it was discussed that the OU process assumed a constant value of σ, *we cannot assume a fixed value of* λ. This is owing to the fact that it is not possible to determine how slow or how fast will interest converge to its long term value. Though, it fluctuates around the mean (a property of the OU process), but as interest is a stochastic process, it cannot be definitely claimed that the process will converge at a constant speed to its long term mean. This therefore is the second drawback of the process.

To fix the problem highlighted in the first point, inspiration from Economics is drawn to make the volatility component (σ) of the process stochastic, thereby introducing the notion of *stochastic volatility based OU process*. The motivation behind this method came from the seminal work presented in [163], where the authors found that the famous Black-Scholes formula was unable to approximate the dynamic nature of financial assets. Therefore, the authors proposed a fix by introducing the notion of stochastic volatility models. The proposed work follows the same procedure, but the fix is applied to the OU process. Moreover, the fix follows the discussion presented in [132] to model the stochasticity in σ via the mean reverting procedure (the OU process). As a result, the volatility component of equation (5.13) is now represented as:

$$d\sigma(t) = \lambda'(\mu' - \sigma(t))dt + \hat{\sigma}d\hat{W}_t. \tag{5.16}$$

At this point, the first shortcoming has been fixed. To solve the second issue with equation (5.13), the convergence speed (λ) is also allowed to follow the OU process. As a result, the equation for the parameter λ is now represented as:

$$d\lambda(t) = \hat{\lambda}''(\hat{\mu}'' - \lambda(t))dt + \hat{\sigma''}d\hat{W}''_t. \tag{5.17}$$

From equations (5.13), (5.16), (5.17), a statistical procedure has been engineered that can now model the evolution of interest. It should be noted here that for the rest of this chapter, equations (5.13), (5.16), (5.17) are called the ***stochastic parameters based OU process*** for interest (In the Results section, the feasibility of the method of variation in λ and σ is discussed. Further, the motive to vary the parameters is also validated).

Before proceeding to the next component of the system, it must be pointed out here that any OU process is dependent on the three crucial parameters: λ, μ, σ. Therefore, one has to find a procedure that can estimate their values from data. In this respect, there is a huge body of work dedicated to the study of parameter estimation in Finance, especially for stochastic volatility

models. This chapter follows one of the most appreciated methods of literature: Maximum Likelihood Estimation (MLE) [164]. MLE is one the achievements of statistical literature in the last century. The method to predict the parameters of the OU process is described in the following subsection.

Parameter Estimation of the OU process

An advantage with the proposed method is equations (5.13), (5.16), (5.17) are modelled via the OU process. Therefore, in this section the procedure for equation (5.13) is described. Similar method is applicable for equations (5.16) and (5.17).

For any OU process, described by λ, μ, σ, the conditional probability density is given by:

$$f(I_{i+1}|I_i; \lambda, \mu, \varphi) = \frac{1}{\sqrt{2\pi\varphi^2}} e\left(-\frac{Q^2}{2\varphi^2}\right) \tag{5.18}$$

where,

$$Q = (I_i - I_{i-1}e^{-\lambda\delta} - \mu(1 - e^{-\lambda\delta})); \varphi^2 = \frac{\sigma^2(1 - e^{-2\lambda\delta})}{2\lambda} \tag{5.19}$$

here, I_i represents the i^{th} Interest value, δ is the time step. Assume that a series of n interest values is available. Then, the log-likelihood is represented as:

$$\mathcal{L}(\lambda, \mu, \varphi) = \sum_{i=1}^{n} \ln f(I_{i+1}|I_i; \lambda, \mu, \varphi) \tag{5.20}$$

Expanding equation (5.20) by substituting the expressions from equations (5.18) and (5.19), and simplifying, we get

$$\begin{aligned} \mathcal{L}(\lambda, \mu, \varphi) &= -\frac{n}{2}\ln(2\pi) - n\ln(\varphi) \\ &\quad -\frac{1}{2\varphi^2}\sum_{i=1}^{n}(I_i - I_{i-1}e^{-\lambda\delta} - \lambda(1 - e^{-\lambda\delta}))^2 \end{aligned}$$

Maximum Likelihood is found at the position where the derivatives are zero. Therefore, differentiating the equation with respect to μ, equating the partial derivative to zero, and simplifying produces the following expression:

$$\frac{\partial \mathcal{L}(.)}{\partial \mu} = \frac{1}{\varphi^2} \sum (I_i - I_{i-1} e^{-\lambda\delta} - \mu(1 - e^{-\lambda\delta}) = 0$$

$$\Rightarrow \mu = \frac{\sum_{i=1}^{n} (I_i - I_{i-1} e^{-\lambda\delta})}{n(1 - e^{-\lambda\delta})} \tag{A}$$

Following the same procedure for λ and φ results in the following equations:

$$\lambda = -\frac{1}{\delta} \ln \frac{\sum_{i=1}^{n} (I_i - \mu)(I_{i-1} - \mu)}{\sum_{i=1}^{n} (I_i - \mu)^2} \tag{B}$$

and,

$$\varphi = \frac{1}{n} \sum_{i=1}^{n} (I_i - \mu - e^{-\lambda\delta}(I_{i-1} - \mu))^2 \tag{C}$$

From equations (A), (B), (C), we have three variables and three equations. Therefore, simplifying the three equations produces the final estimates as:

$$\mu = \frac{I_a X_{bb} - I_a X_{ab}}{n(I_{aa} - I_{ab}) - (I_a^2 - I_a X_b)} \tag{5.21}$$

$$\lambda = -\frac{1}{\delta} \ln \left(\frac{I_{ab} - \mu I_a - \mu I_b + n\mu^2}{I_{aa} - 2\mu I_a + n\mu^2} \right) \tag{5.22}$$

$$\sigma = \sqrt{\frac{2\lambda\vartheta^2}{1 - e^{-2\lambda\delta}}} \tag{5.23}$$

where,

$$\begin{aligned} \vartheta = \frac{1}{n} [& I_{bb} - 2e^{-\lambda\delta} I_{ab} + e^{2\lambda\delta} I_{aa} \\ & - 2\mu(1 - e^{-\lambda\delta})(I_b - e^{-\lambda\delta} I_a) + n\mu^2 (1 - e^{-\lambda\delta})^2] \end{aligned} \tag{5.24}$$

In the above equations, $I_a = \sum_{i=1}^{n} I_{i-1}$, $I_b = \sum_{i=1}^{n} I_i$, $I_{aa} = \sum_{i=1}^{n} I_{i-1}^2$, $I_{ab} = \sum_{i=1}^{n} I_{i-1} I_i$, $I_{bb} = \sum_{i=1}^{n} I_i^2$.

5.3.4 Process Step III: Transforming Interest into Activity. A Dynamic and a Self Configuring Measurement Function

This section addresses the next process-step in the workflow presented in Section 5.3.1 (Point III). More specifically, the procedure that transforms interest into activity is presented here. In doing so, inspiration is drawn from adaptive filtering and function approximation.

To find a statistical procedure that can simulate the transformation dynamics of interest changing into activity, a functional map $M : I \longrightarrow A$ is needed so that the input to the function is the set S of data points $\{(I_1, A_1), (I_2, A_2), (I_3, A_3), ..., (I_n, A_n)\}$, where $(I_i, A_i) \in X = I \times A$, I is the set of interest values and A is the set of activity values. Although, it is well known that to find such a function is non-trivial, the goal is to produce an approximate map that minimizes the following cost function:

$$\min l_{ER} = \sum_{j=1}^{N} (A_i - m(I_i))^2. \tag{5.25}$$

where, N is the sample size, *m(.)* is the function of interest that can predict activity, l_{ER} is the empirical risk. By exploring literature, it was found that the solution of equation (5.25) is ill-posed and is dependent upon the potential hypothesis space (A space containing potential definitions of the function). Therefore, work uses approximation techniques. Though this particular problem has been investigated in literature, the issue, however, is in finding the most suitable candidate. To that end, we follow the theoretical discussion presented in [165] where the authors reviewed existing literature and found that work considers a positive correlation between curiosity and actions. That is, if one variable increases, the other related variable also increases with it and vice-versa. Taking inspiration from this phenomenon, the formulation of this subsection puts this theoretical notion into practice. Speaking from a statistical point of view, one of the ways to do this is to use the Least Mean Square (LMS) algorithm. But, the disadvantages of the LMS algorithm are well documented in literature. Consequently, to solve the above equation, the Recursive Least Mean Square (RLS) Algorithm [166] is employed. Though, it is computationally expensive (requires matrix manipulation), but it achieves a high convergence speed. To apply the theory of RLS, one has to rewrite equation (5.25) as:

$$\min l_{ER} = \sum_{j=1}^{N} |A(j) - I(j)^T \Omega(N-1)| + \rho ||\Omega(N-1)||^2. \tag{5.26}$$

where, ρ is the regularization factor $\in$ (0, 1) and Ω is the weight vector of RLS. The weight vector Ω is especially important as it provides a measure of the importance of previous errors

made in prediction. Predicting activity from interest, especially in light of uncertainty, is a significant challenge. The weight vector Ω comes in handy as it helps to change the internal mechanisms of the system. In lay terms, this method gives a *reconfigurable black-box*. To do this, RLS updates the weight vector in equation (5.26) by recursively updating the error matrix and the data autocorrelation matrix. Along this principle, the procedure to predict activity from interest is summarized in the following algorithm:

Algorithm 3. RLS Algorithm to Convert Interest Into Activity.

Input: Data Points consisting of Interest and Activity (I_n, A_n), $n \in \{1, 2, ..p\}$

Output: Predicted Activity Vector $\hat{A}$, Weight Vector Ω.

I. Initialize the System. $\Omega(0) = 0$,$P(0) = \rho^{-1}I$

II. For $n > 1$, do

$r_e(n)$ = 1 + $I(n)^T P(n-1)I(n)$

$K(n)$ = $P(n-1)I(n)/r_e(n)$

$e(n)$ = $A(n) - I(n)^T \Omega(n-1)$

$\Omega(n)$ = $\Omega(n-1)$ + $K(n)e(n)$

$P(n)$ = $P(n-1) - P(n-1)I(n)I(n)^T P(n-1)/r_e(n)$

II. end for

In the above algorithm, $I(n)$ is the interest vector, $A(n)$ is the activity vector, $P(i)$ is the inverted and regularized autocorrelation matrix, $K(i)$ is the gain matrix, and $e(i)$ is the error. Also, the dimension of $P(i)$ is $p \times p$, where p is the dimension of the Interest Vector.

Applying Algorithm 3, the computational definition of the phenomenon that can convert interest into activity is now complete. Further, a statistical definition of an adaptable and a self-adjusting measurement function is obtained that can dynamically learn and alter itself based on changing circumstances. As already noted in Section 5.3.3, interest changes itself stochastically, therefore, the crux of the matter is that a similar procedure is needed that can automatically manage these ever changing state of affairs.

5.3.5 Process Step IV: Finding Interest from Measurable Activity, An Application of Recursive Bayesian Filtering

As per the workflow specified in Section 5.3.1, the next goal (Point IV) is to define the procedure that can filter numerical estimates of interest from activity. To do this, state estimation problems rely on the so-called Monte Carlo methods. Such methods are used as their statistical properties are ideal for quantifying uncertainty. It can be deduced that we are dealing with Uncertainty Quantification with application in the internal mental states. Therefore, a variant of this family is employed. More specifically, particle filter is used to find numerical estimates of interest.

Particle filters are a variant of the Monte Carlo class of simulation algorithms wherein the objective of the system is to use the principle of random sampling to find good numerical estimates of the underlying variable (in this case interest). Furthermore, particle filters are good approximation algorithms that are frequently used in cases where the underlying structure of the model is not accurate [167]. This strategy is along the lines of the interest prediction problem as precise evolutionary dynamics of interest are unknown. To find numerical estimates of interest, the particle filter is provided with a set of X particles, where a particle is represented as (Υ^m, w^m), here Υ^m is the m^{th} hypothesis for interest, w^m is the weight, also called as the importance, of the m^{th} hypothesis. As particle filters use Monte Carlo simulations, therefore, the set X is *initialized* from an *initial probability distribution*. In the subsequent steps, every particle is sampled via the known distribution (equations (5.13), (5.16), (5.17)). Based on this sampling, and for every particle, activity is predicted and the importance of each hypothesis (of interest) is computed by comparing the predicted activity with the actual activity. To do this, the information stored inside RLS is employed. It was specified that particles are sampled randomly, hence, a few particles are susceptible to the problem of weight collapse. That is, the possible numerical hypothesis of interest is poor. To overcome this, the system discards the poorly sampled hypothesis and do not let their inefficiency compromise the procedure. This is usually done via cumulative distribution. Once the iterations are complete, we take the mean of the particles to get an estimate of interest. The method is summarized in Algorithm 4.

Algorithm 4. Particle Filter to Estimate Interest from Activity.

Input: Activity Vector, A_i, $i \in \{1, 2,...,n\}$.

Output: Interest Vector, I_i, $i \in \{1, 2,...,n\}$.

i. Initialization: Sample X particles, where $X = \{p^1, p^2, .., p^Z\}$. A particle, p^m, is expressed as: (Υ^m, w^m). Estimate initial values as: $\Upsilon_0^m \sim \frac{1}{\sigma_M\sqrt{2\pi}} e^{-(\mu_M)^2/2\sigma_M^2}$.

---

ii. for, $j = 1, 2, ..., k$.

iii. for $i = 1, 2, ..., |X|$, sample $\Upsilon_t^i | \Upsilon_{t-1}^i$ using equations (5.13), (5.16), (5.17).

iv. Compute activity, $\hat{A}_t^i$, via Algorithm A1.

v. Set $\hat{\Upsilon}_{0:t}^i$ = ($\Upsilon_{0:t-1}^i$, $\hat{\Upsilon}_t^i$). Compute importance, w_t^i, as: $w_t^i = \frac{1}{\sigma_B \sqrt{2\pi}} e^{-\left(A_t - \hat{A}_t^i\right)^2 / 2\sigma_B^2}$.

vi. Compute total weight H=$\sum_{i=1}^{|X|} w_t^i$.

vii. Normalize. w_t^i=$H^{-1} \times w_t^i$.

viii. Resample.

ix. Take mean of particles $p(\Upsilon_t) \in X$, and compute interest.

x. end i.

xi. end j.

Go to next Iteration.

5.3.6 The issue of Activity Gap

So far, the statistical considerations required to predict interest from activity are complete. However, despite tackling each of the individual issues highlighted in Section 5.3.1, there is an issue left. That is, the proposed method is unable to address the problem of *Activity gap*. If we look at the core of the discussion, we realize that the method can predict interest only when data about activity is fed to the system. However, in practical situations it is expected that information about activity would not always be present, and hence, interest prediction in those cases would be out of scope. To understand this issue, lets consider that John is interested in Facebook and logs-in on the platform everyday. Owing to certain unpredictable circumstances, e.g. he is working to meet certain work deadlines, John is unable to engage with Facebook on the "*current*" day. Hence, there are gaps in activity. Accordingly, and as per the proposed method, interest estimation is not possible at the "*current*" day. It is however understood that interest is not zero when there is no activity. This use case highlights the problem of activity gap.

To fix the issue, the base equation (5.3) is modified and a solution is found via the principle of K-Step ahead prediction density. To accommodate this principle, the system evolves interest

according to the following theoretical model:

$$p(I_{t+k}|A_{t-1},\gamma) = \int_I p(I_{t+k}|I_{t-1},A_{t-1},\gamma)p(I_{t+k-1}|A_{t-1},\gamma)dI_{t+k-1}. \tag{5.27}$$

To practically implement the strategy, the system automatically evolves interest via equations (5.13), (5.16), and (5.17) in cases of activity gap. To exemplify this, on the day that John was unable to engage with Facebook, the system uses equations (5.13), (5.16), and (5.17) to predict his interest value. In this case, it is expected that the method can predict interest, although, it is unable to update it. But, as soon as data about activity is fed to the algorithm, the system use equations (5.13), (5.16), and (5.17) and Algorithm 4 to predict as well as update John's interest. The essence of the idea is that a procedure similar to the continuous time model of interest is engineered. Further advantage of this idea comes from the fact that we expect any internal state of a human (not limited to interest) to follow a continuous time function.

5.4 Results

To validate the feasibility of the proposed framework, numerical simulations is performed on Datasets provided by StackOverflow. Owing to its popularity, research has found that users of StackOverflow are addicted to participate in its daily activities [168], [169]. Hence, based on these findings, the platform presents an excellent opportunity to test the feasibility of the method in practical scenarios. It is clear that interest is estimated via activity, therefore, the first step in the procedure is to compute activity. Recall that activity depends upon several attributes, therefore, the following attributes were collected from StackOverflow: 1) The number of comments. 2) The number of answers. 3) The number of questions. 4) The number of edits. 5) Time to Answer. Owing to privacy reasons, more attributes could not be included. Data of 250 users has been collected each day for a period of one year.

5.4.1 Prototype Development

To demonstrate the viability of the method in actual deployment scenarios, a prototype has been engineered. The prototype is implemented in JAVA. The mathematical functions used in the program were implemented from the Apache Common Math Library[1]. For the purpose of matrix manipulation, libraries provided by EJML[2] were used. Once the Java classes were

[1] http://commons.apache.org/proper/commons-math/
[2] http://ejml.org/

encoded, the method was deployed as a RESTful Web Service on Apache Tomcat v7.0.41. Lastly, to validate the viability of the method in Cloud based computing environments, the application was hosted on a virtualized testbed consisting of several Virtual Machines (VMs). The VMs were hosted on XEN[3] as the base hypervisor. The configuration of the underlying hardware was the IBM Tower Server with Detachable HDDs, Intel Xeon X5 Processor, 48 GB RAM, 1.8Ghz processing speed.

5.4.2 Dataset Description

StackOverflow has one of the largest public data repositories. The database has 27 tables with a total of 191 different attributes describing the details of every post, vote, user, comment, revision, tag and so on. Though, the dataset is extensive, however, for the purpose of activity calculation (especially for every user separately), only 5 relevant attributes are available. we specified in section 5.4.1that the attributes are: 1) The number of comments. 2) The number of answers. 3) The number of questions. 4) The number of edits. 5) Time to Answer. Further, it was specified that data for a total of 250 users was collected daily for one year. To do that, SQL queries were executed on *live* StackOverflow databases available at the link[4]. One can understand, the attributes of any dataset are scattered across different tables, moreover, sometimes to get an attribute that is not a part of the standard tables in the dataset, some cross-table SQL processing has to be done. Therefore, for this purpose, for example to get the number of comments from each user, several cross-table queries were written. It is understandable that the returned data after executing the queries is in raw format. Consequently, the next step is data cleaning and formatting. To do that, although there are many open source toolkits available, however, they do not give enough control and are limited for the purpose of data manipulation (it is not tailored for our requirement). As a result, several independent Linux and Python scripts were written and the necessary details (the attributes) of every user was then obtained.

5.4.3 Experimental Setup

What is Attribute Matrix?

In section 5.3.2, a method to computationally calculate activity was discussed. The final equation to compute activity is summarized below:

[3] http://xenserver.org/
[4] http://data.stackexchange.com/

	The no. of comments	The no. of answers	The no. of questions	The no. of edits	Time to Answer
Day 1	1	5	5	3	45
Day 2	3	4	2	8	510
Day 3	6	9	3	7	787
Day 4	1	4	9	5	45
Day 5	5	1	6	3	858
Day 6	8	3	7	1	55
Day 7	2	1	4	2	78

Table 5.1: An Example of the Attribute Matrix.

$$A^t = \gamma \times SM \times AM^t + (1 - \gamma) \times OM \times AM^t. \tag{5.28}$$

where AM is the attribute matrix, OM is the objective weight matrix, SM is the subjective weight matrix, $\gamma \in (0, 1)$ is the bias parameter, A^t is the activity value.

To compute the attribute matrix, we need different perspectives of activity. Recall that activity is dependent upon different attributes. The attributes were specified in section 5.4.2. With those attributes an example of the attribute matrix is discussed in the following paragraph.

In Table 5.2, a sample of the activity matrix is presented. The rows of the matrix corresponds to the numerical value of the attributes (or perspectives) and the columns corresponds to time. In context of the platform chosen for the paper, time is in days, hence the title Day 1 (for example in case of row one). In this example, data for only 7 days is presented. The matrix that contains the numerical values is the Attribute Matrix. For the convenience of the reader, the attribute matrix for the example presented in Table 5.2 is also presented here:

$$AM = \begin{bmatrix} 1 & 5 & 5 & 3 & 45 \\ 3 & 4 & 2 & 8 & 510 \\ 6 & 9 & 3 & 7 & 787 \\ 1 & 4 & 9 & 5 & 45 \\ 5 & 1 & 6 & 3 & 858 \\ 8 & 3 & 7 & 1 & 55 \\ 2 & 1 & 4 & 2 & 78 \end{bmatrix}$$

From the example presented in this section, one can therefore generalize the attribute matrix for any time unit (not limited to days), any platform, and for any number of attributes.

Attribute	Subjective	Objective
Answers	0.1549	0.5941
Questions	0.1333	0.1166
Comments	0.2127	0.0916
Edits	0.1944	0.0536
Time to Answer	0.3047	0.1441

Table 5.2: Subjective Objective Weights.

How to Calculate Subjective-Objective Weights.

In Algorithm 1 and 2, it was pointed out that the method to calculate the weight is the subjective-objective technique. It was specified in the in section 5.3.2 that the methods are taken from [156]. In this section, a few additional details are presented.

The first step, that is, the input to the two algorithms are the decision matrix (for the subjective method) and the normalized decision matrix (for the objective method). The decision matrix is the pairwise comparison matrix given by the decision maker [156]. In other words, the decision matrix contains the preference of the Decision maker towards every perspective of activity with respect to other perspectives. In a pairwise comparison matrix, two different attributes are evaluated on the basis of their relative importance. Values ranging from 1 - 10 are used. In this, if an attribute (say x) is exactly the same, that is, it is as important as any other attribute (say y), the pair gets a value 1. If, on the other hand, x is more important (assume very important) than y, the numerical value is 10. Rest of the ranking is in between 1-10. A good example and detailed explanations on the decision matrix (also called as Saaty's Matrix) can be found in [157]. The normalized decision matrix is the normalized (between 0-1) version of the decision matrix. Once the input data is fed to the algorithm, we use the optimization procedure as discussed in [156] and come with the subjective-objective weights. The final weight matrices after applying this procedure is shown in Table 2.

Activity Calculation, Interest Estimation, and Metrics Used in Evaluation

It is clear that interest is estimated via activity, therefore, the next step in the experimental setup is to compute activity. Recall that activity depends upon several viewpoints. Further it was specified that five different attributes were collected. With these attributes, the method to compute activity is explained in Module I.

Module I. Activity Calculation.

	V1	V2	V3	V4	V5	VN1	VN2	VN3	VN4	VN5	Activity
Day 1	1	5	5	3	45	0.125	0.5555555556	0.5555555556	0.375	0.0524475524	0.2506716453
Day 2	3	4	2	8	510	0.375	0.4444444444	0.2222222222	1	0.5944055944	0.4765237393
Day 3	6	9	3	7	787	0.75	1	0.3333333333	0.875	0.9172494172	0.7709347436
Day 4	1	4	9	5	45	0.125	0.4444444444	1	0.625	0.0524475524	0.3267072009
Day 5	5	1	6	3	858	0.625	0.1111111111	0.6666666667	0.375	1	0.6181169444
Day 6	8	3	7	1	55	1	0.3333333333	0.7777777778	0.125	0.0641025641	0.5954884615
Day 7	2	1	4	2	78	0.25	0.1111111111	0.4444444444	0.25	0.0909090909	0.2269494444

Table 5.3: Activity Calculation. Bias Parameter $\beta = 0.4$.

Input: Subjective-Objective Weights, Bias parameter, Attribute Matrix.

Output: Activity Vector

Steps:

I. The initial input requirement for this Module is the subjective objective weight matrices. For this purpose, the procedure discussed in [156] was followed and the necessary data (weight matrices) were obtained (They are presented in Table 5.3).

II. The subsequent step in the Module is to obtain the attribute matrix. Recall that activity depends upon the weight matrices and the attribute matrix (equation (5.28)). For the attribute matrix, the five attributes as collected from StackOverflow are shown in Table 5.3. They are represented under the column V1:V5. The data exemplifies the case of one random user for a period of 7 days.

III. The data shown under the column V1:V5 was then normalized to 0-1. This is represented under the columns titled NV1:NV5. The attributes were normalized to maintain uniformity across different perspectives (or attributes) of activity. The matrix containing these numerical attributes is now called as the attribute matrix.

IV. From the matrices obtained in Steps I and III, equation (1) was used to compute the final activity vector. The resulting activity vector is presented under the column titled activity in Table 5.3.

V. Steps I-IV were followed for every user.

From the data obtained under Module I, the next step is to estimate the interest vector. The procedure to obtain the interest vector is discussed in module II. It should be noted here that the procedure discussed under Module II was repeated for every user separately.

Module II. Estimating Interest from Activity.

Input: Activity Vector.

Output: Interest Vector, Predicted Activity Vector.

Steps:

1. It was specified that a prototype was engineered in JAVA, therefore, specific classes were written for Algorithms 3 and 4. Once the classes were implemented, the input data from Module I was fed to the system.

---

2. The main class of the system consists of the code written for Algorithm 4. This class had inbuilt functions to call the transformation function, equations (5.13), (5.16), (5.17), and the measurement function (Algorithm 3).

3. Once the necessary information was programmed, steps presented in Algorithm 4 were used to obtain numerical estimates of interest.

We know that through Bayesian Inference, we can not only predict interest but also activity. As there is an absence of a formal procedure in literature to model interest, we do not have a method that can find an exact number for interest. Hence, comparing the accuracy of the proposed model is an issue. To demonstrate that the obtained number (for interest) is a true representative of interest, we follow a method found in Bio-Medicine literature. Work in this discipline has to deal with constructs that are not directly measurable, for instance, the internal variables of the human heart [170], the human brain [171] and so on. In this field, state estimation models are utilized to deduce the values of the hidden variable. Our work has similar circumstances. Therefore, to find an approximate value of interest, we use indirect inference rules. If the proposed method can approximately quantify activity (the output variable), then we can, in an indirect way, say that the method can approximate interest as well. Hence, by comparing the actual activity with the predicted activity, we evaluate the performance of the framework. For this purpose, we use Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The methods for calculating RMSE and MAE are explained in the following text. Furthermore, 50 runs are conducted and the average values are presented. The equations for RMSE and MAE are described below:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} \tag{5.29}$$

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |e_t| \tag{5.30}$$

where, $e_t = A_t - \hat{A}_t$, here A_t is the actual activity (obtained from Module I) and $\hat{A}_t$ is the predicted activity (obtained from Module II). The procedure to obtain the RMSE and MAE values is described in the following Module.

Module III. How to calculate RMSE and MAE.

---

1. By following the discussion under Module II, a total of 250 Interest vectors were obtained. Further, as is expected in Bayesian Inference problems, a total of 250 activity

---

vectors were also obtained. Subsequently, basic principles of error calculation were used to estimate the numerical values of RMSE and MAE by comparing the predicted activity vectors with the actual activity vectors available to the system. Note, the predicted activity was obtained from the procedure followed under Module II and the actual activity was obtained by following the procedure discussed under Module I.

2. The above procedure was repeated for each of the 250 users in the dataset.

3. From steps 1-2, a total of 250 RMSE and MAE values were obtained (One for every user). Subsequently, mean of all the 250 numerical values was taken, thus, obtaining a single value for RMSE and MAE.

4. Steps 1-3 were repeated 50 times, thereby the system had a total of 50 RMSE and MAE values.

5. Once the system had 50 RMSE and MAE values, average of all the numerical data was taken and the final number is presented in the section. The number is a representative of the overall predictive capability of the model.

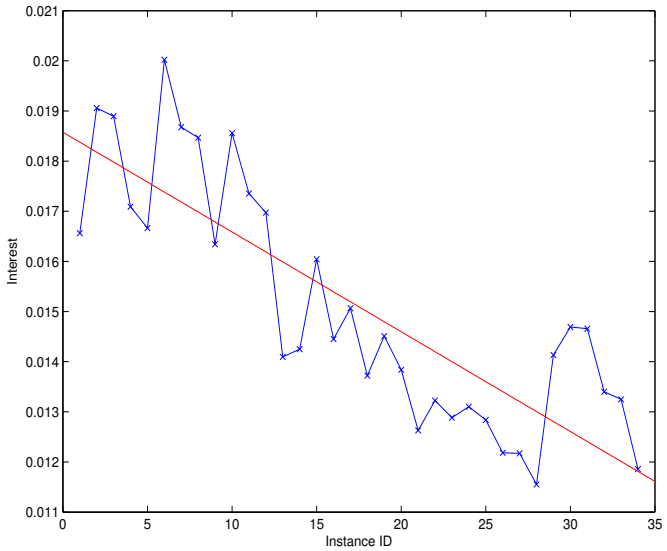

Figure 5.2: Evolution of Interest for One Random User.

5.4.4 Data analysis

To start with the data analysis, Fig. 5.2 show the results for the evolution of interest on a daily basis (for one month). The results for only one random user are presented. It is visible from the figure that the evolution follows several ups and downs. The pattern presented in the figure is a representative of the everyday fluctuations that interest goes through during one's lifetime. To exemplify this, consider the case where a person is interested in StackOverflow. Owing to

several erratic and unpredictable circumstances in one's daily routine, a person is expected to engage with StackOverflow in a dissimilar manner. That is, the person will not engage with StackOverflow with the same intensity and rigour every time. It is understandable that on a few occasion (days in this case) the interest will be high whereas on other days it could be low (there could be a plethora of reasons for this). The pattern shown in Fig. 5.2 is a computational realization of this real world phenomenon. With this particular observation in mind, it should be noted here that the idea of the proposed framework is to algorithmically model the evolution of interest. In general terms, the motive is to quantify: *How much does anyone enjoy any object (for instance StackOverflow)*? To do that, and to estimate the internal mental property of interest, several readily available attributes of activity are used (they were highlighted in the beginning of this section). Moreover, the benefit of the proposed framework comes from the fact that this attempt is the first where the estimation of interest has been done purely in an objective manner. Though, as will be discussed from the next subsection onwards that the method just scratches the surface, the chapter nevertheless has made an attempt to present a systematic direction towards answering the question asked in Chapter 1 of the thesis.

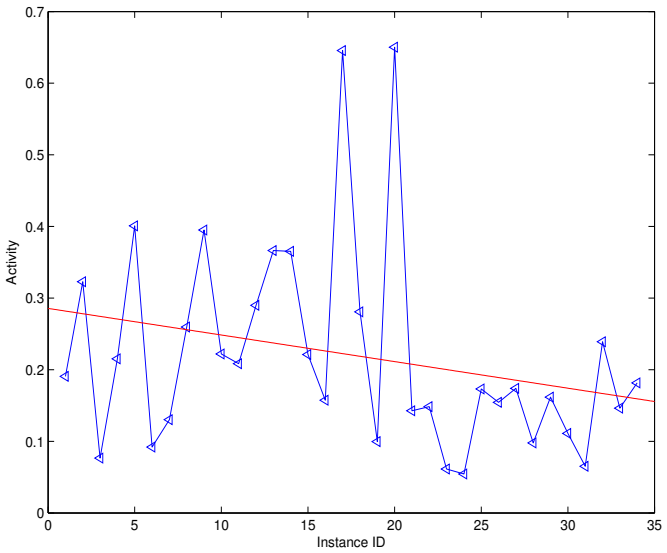

Figure 5.3: Activity for One Random User.

In much the same way, in Fig. 5.3 the graph of activity is presented. Results for a period of one month are presented. Note, the corresponding interest values for the same user are presented in Fig. 5.2. It is visible from the figures that both activity and interest goes through several ups and downs. This was discussed in the previous paragraph. However, an important point of note for the two graphs is the trend the results follow. The trend-line in both the graphs has a negative slope. That is, the activity is decreasing in tandem with interest. This result is intuitively expected and objectively feasible. High activity implies high interest and vice-versa. The proposed method is able to capture this phenomenon, hence, it can be said that the method lives up to the theoretical as well as to the intuitive expectations. It should be noted here that for

	MAE	RMSE	Exe. Time
RW	2.0939815	4.3490455	6157
GBM	0.71731357	3.6168484	7851
Proposed Framework	**0.0309669**	**0.1163177**	**9612**

Table 5.4: Comparison with Random Walk and Geometric Brownian Motion. RW: Random Walk. GBM: Geometric Brownian Motion. Execution Time is in Milliseconds.

an "*accurate*" system model, one expects interest and activity to follow the exact same pattern, however, as the chapter models and computationally simulates a complex mental state, it is hard to get accurate readings. In this respect, it was specified that the performance of the method is evaluated by comparing the predicted activity and the actual activity. In this context, the result of activity prediction for one random user is graphically demonstrated in Fig. 5.4. It is visible from the figure that the result of prediction is not accurate. In Uncertainty Quantification as well as in Stochastic Systems, it is hard to make accurate predictions. Moreover, it was also specified in this paragraph that finding an accurate system model to capture interest is difficult. Therefore, the objective in such problems is to approximate the most optimal values, thereby minimizing the error in prediction [167]. From the next subsection onwards, the performance of the method is analyzed in detail and a comparison with procedures of a similar kind is performed. In doing so, evidence will be presented that clearly signifies the advantage of the proposed framework.

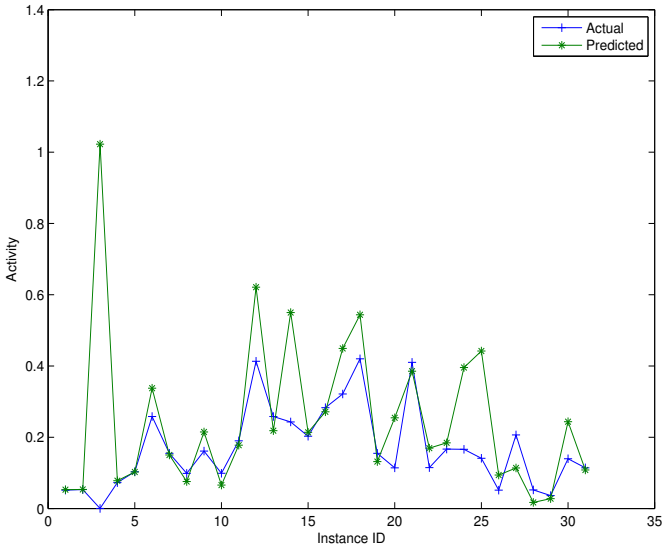

Figure 5.4: Predicted Activity vs Actual Activity.

	MAE	RMSE
OU process	0.9532092	2.0873524
OU process with varying λ	1.1709875	2.758046
OU process with varying σ	**0.0309669**	**0.1163177**
OU process with varying λ and σ	0.0629224	0.3565446

Table 5.5: Accuracy for Different Variations in Equation (5.13). The variation is modelled via Mean Reverting Stochastic Procedure.

5.4.5 Comparison with Random Walk and Geometric Brownian Motion

As there is a lack of literature on modelling the evolution of interest, in an attempt to demonstrate the efficacy of the proposed method, two additional models are considered. Specifically, the performance of the stochastic parameters based OU process is compared with Random walk (RW) and Geometric Brownian Motion (GBM). Although, there are several approaches for prediction in literature, these approaches are chosen as they have a particular affinity towards stochasticity, and are some of the more popular models of literature that are widely applied across disciplines. It is mainly owing to the fact that they have good discrete and computational properties. In this thesis, therefore, these approaches are chosen for the purpose of experimentation and comparison. To implement these methods, equations (5.13), (5.16), (5.17) are discarded, and interest is allowed to take a RW, subsequently, interest follows the GBM. Keeping the rest of the procedure intact (that is, Algorithms 3 and 4 are left untouched), activity is estimated and error in prediction is presented. With this setup, the results of the experiment are shown in Table 5.2. It is visible from the Table that the value for error, for the RW model, is high. The numbers signify that random walk is unable to approximate the dynamic nature of interest. Consequently, this inability results in poor approximations, hence, a high value of RMSE and MAE. Subsequently, when interest is modelled via GBM, the numbers improve (in comparison to RW). But, they are nevertheless high. In contrast, when interest follows the proposed method, the results witness a significant improvement. The best performance is achieved when interest follows the stochastic parameters based OU process. To be specific, MAE saw a reduction of 98.52% and 95.68%; RMSE values improved by 97.35% and 96.78% (w.r.t. RW and GBM). This increment in performance is most certainly noteworthy. Though, it is also visible from the Table that the method compromised on execution time, it improved accuracy.

5.4.6 Stochastic Volatility and Effect of Varying Convergence speed

In Section 5.3.3 interest was modelled via equation (5.13). However, it was specified that this equation assumes a constant value of σ and λ. The problem therefore was fixed by introducing stochastic variations in σ and λ. In this subsection, additional tests are performed to evaluate the validity of the two fixes. To do that, experiments are conducted in the following ways:

	MAE	RMSE
OU process	0.9532092	2.0873524
OU process with varying λ	2.3726269	6.5573857
OU process with varying σ	**0.9009318**	**2.0765609**
OU process with varying λ and σ	2.6934764	7.3547672

Table 5.6: Mode of Variation I. The parameters follow Random Walk.

1. Only the base OU process with no changes to volatility and convergence speed is considered. Interest is modelled via equation (5.13).
2. The convergence speed is varied. Interest is modelled via equation (5.13) and (5.17)
3. The volatility component of equation (5.13) is varied and the convergence speed is kept fixed. That is, interest is expressed via equations (5.13) and (5.16).
4. Lastly, simultaneous effect of varying speed and volatility is investigated. Interest is modelled via equations (5.13), (5.16), and (5.17).

Keeping the rest of the procedure as it is, that is, Algorithms 3 and 4 are untouched, the result for each of the test case is shown in Table 5.3. The numbers presented in the Table, specifically test case 1, points to the inference that the performance of the basic OU process with no variations in either σ or λ is compromised. This was theoretically expected as the volatility component (σ) and the speed component (λ) of the OU process is constant. Thus, the motive to fix the problems is justified. For the second test case, varying the speed of convergence (λ), the results have deteriorated even further (w.r.t Test Case 1). The numbers are poorer than the basic OU process. Before beginning the experiments, it was expected that varying the convergence will improve performance, but the results show a different picture. Therefore, it can be concluded that varying the speed of convergence did not result in any performance improvement. In the third case, however, it can be seen that the results have improved by a noticeable margin. It is evident from the Table that the performance of the model is best by applying the stochastic volatility model alone. Using this type of method, we get better results. More accurately, MAE improved by 96.75% and for RMSE the figures improved by 94.42% (w.r.t. Test Case 1). These numbers are noteworthy.

5.4.7 Additional Investigation in Parameters

The investigation conducted in this section so far is an indicative of practical feasibility of the stochastic parameters based OU process to model interest. As our motive is to focus on statistical algorithms to model interest, further exploration is performed. Drawing inspiration from the analysis presented in Section 5.4.7, two different hypotheses are formulated:

1. Are OU process based variations in the parameters (σ and λ) the most optimal?

	MAE	RMSE
OU process	0.9532092	2.0873524
OU process with varying λ	1.4207251	10.384024
OU process with varying σ	0.7527388	4.4423563
OU process with varying λ and σ	**0.2081836**	**0.6902227**

Table 5.7: Mode of Variation II. The parameters follow Geometric Brownian Motion.

2. What is the effect of other modelling procedures (in σ and λ) on the performance?

Recall that the proposed variation in the parameters were modelled via the mean reverting stochastic process (equations (5.16) and (5.17)). In this section, and to test the two specified hypotheses, experimentation is conducted with two different models. More accurately, the mode of variation in λ and σ is now changed from the mean reverting stochastic procedure (the OU process) to i) RW and ii) GBM.

In the first experiment, the underlying parameters (λ, σ) are varied via RW. In doing so, equations (5.16) and (5.17) are discarded and the variation in σ and λ is modelled via RW. It should be noted here that for the experiments, the base equation of the model, equation (5.13), is left untouched. Subsequently, using Algorithm 3 and 4 activity is predicted and error in prediction is presented. The results are shown in Table 5.4. It can be seen from the evidence presented in the Table that similar to the numbers in Table 5.3, we get the best performance by varying the volatility (σ) component alone, Test Case 3. Though, the improvement margin is not as good as that presented in Table 5.3, nevertheless, varying the parameter σ improved performance. It should be noted that the evidence in Table 5.4 shows that varying λ degrades the performance (Test Case 3). Moreover, varying σ and λ simultaneously results in the same degraded performance.

Similar to previous experiment, where the parameters were varied via RW, in the next experimental setup, the variations in σ and λ are now modelled via GBM. The result for this experiment is shown in Table 5.5. The numbers and the evidence in this table, however, tells a different story. This time the results are best when σ and λ are varied simultaneously. This is unlike the results presented in Table 5.3 and 5.4 where varying σ alone resulted in the best performance. Nevertheless, and however the combination, the motive to vary the parameters improved performance.

In light of the evidence presented in Tables 5.3, 5.4, 5.5, it can be concluded that when we vary the parameters via different models, the combination that produces the best performance is inconsistent. But, *the best results are obtained by varying σ alone and by modelling the stochasticity via the proposed procedure (equation (5.16))*. With respect to the experiments conducted in this section, a few lessons are learned. They are summarized in the following points.

I. The evidence in Tables 5.3, 5.4, 5.5 is a clear implication of performance improvement by varying the parameters (of equation (5.13)). Though, the combination and the numbers presented in the Tables are different, the idea and the motive to vary σ and λ gave good results.

II. To reduce the error margin and to improve the performance, the modelling procedure plays an important role. ***In this thesis, three different type of models were tried, however, the best performance was obtained via the proposed mean reverting procedure (the OU process).***

III. The combination of the parameters is also an important criterion. The best accuracy was obtained by varying σ only (Table 5.3). However, it was also seen that this combination is not universal (Results in Table 5.5).

IV. Lastly, we must not forget the *No free lunch theorem* in Machine Learning [172]. In other words, experimentation was performed with StackOverflow databases. For other platforms (e.g. Facebook, Twitter etc.), the mode as well as the combination of parameters could be different. It is therefore recommended to experiment and test multiple models on other platforms. In short, there is no shortcut for every potential encounter.

5.5 Discussion

So far a computational method has been discussed that can model and estimate the interest of an individual. Though, the framework is designed for a general purpose, considering the prevailing technology and the limitations imposed by the state-of-the-art, the work has several shortcomings. They are as follows:

1. The proposed method cannot measure interest for all possible circumstances and scenarios. For example, consider the case where a person is interested in reading novels (via paper back books). For this case, the perspectives of activity could be: The number of pages read, the number of hours spent reading, the number of reading sessions, the speed of reading, and so on. Though, the perspectives are feasible, the big issue however is: *How to measure them computationally*? More accurately, there is a lack of an omnipresent computational system that can observe activity for all possible contexts/applications/scenarios. In simple words, there is lack of data. The presence of data is imperative for a computational system to work properly. Therefore, for this use case and other similar use cases, interest estimation is out of scope.

2. Along the same direction of the previous point, interest cannot be quantified when it (interest) is only in one's mind. For instance, consider an individual is interested in learning new technologies, but has not taken any steps that could tell a third person about his/her interest. Similar to the previous point we can understand, a computational system needs tangible attributes and data. The current technology is not advanced enough to understand any internal mental state without making any observations. One has to take a few actions that can be

recorded by a medium (or media) that is computationally operable. Therefore, in situations where interest is only in one's mind, interest estimation is not possible.

3. With respect to estimating interest and comparing the interest of individuals, it should be noted here that we cannot compare the interest of all individuals on the same scale. Literature in Psychology has repeatedly specified that interest is a construct that has high "*personal*" definition [86]. It has been outlined that interest is directed at real entities, and has a personal meaning [173]. In simple words, every person has his/her own style of showing interest. It is therefore crucial not to compare the interest of all individuals on the same scale. Neither the proposed framework nor existing research allows comparing the interest of individuals on a common ground.

4. The work proposed in this Chapter has tried to find a number for interest. A natural question from the discussion arises: *What is meant by this number*? In simple words, how a make an emotionless machine analyze, rationalize, interpret, and then understand the number? This is a tough question to answer. At this point, human beings themselves do not completely understand the interest of another person (It can be considered fuzzy). This is one of the most frequently occurring questions that literature in Artificial Intelligence has to answer. To this, it must be pointed out here that the limitations imposed by the current state-of-the-art does not let a machine "*feel*" the human emotion of interest. A subsequent question closely related to this is: What are the psychological consequences of having a machine deduce the property of interest? Similar to the previous question, this question too is non-trivial. The two questions remain a problem that we intend to address in our future work.

5. The next issue is privacy. To estimate interest, the proposed method needs access to the private data of an individual. This raises concerns over privacy. We faced this situation while exploring datasets for experimentation. Prior to conducting the tests on StackOverflow, students studying in the Institute were approached to volunteer for the experiment and provide some private data. However, students raised concern over privacy and denied us access. Although, the motive of the work was purely academic, students felt reluctant to share their details, e.g. no one wanted to share the details of Facebook, WhatsApp etc. In this regard, concern over privacy is not new. There is a huge body of work dedicated to its study, e.g. [174], [175]. However, the potential of having an automatic method estimate human like interest is also considerable. If we can engineer a better and a privacy preserving interest estimation technique, the societal impact will be significant.

In this chapter, a technique that could computationally estimate interest was proposed. However, it is not claimed here that the method is accurate. From the discussion in the chaper, it is clear that the problem to accurately estimate interest is a challenge. Moreover, the discussion in the above points clearly adds additional complexity to the issue. Nevertheless, a systematic

sequence of steps has been shown that could help answer the simple yet powerful question: How “much” are you interested? Finding an exact solution to the problem is a long way to go. More efforts are needed and more cross disciplinary research is required to find an accurate answer.

5.6 Summary

In this chapter, a method to quantify interest was proposed. An important feature of the work was its ability to link data analytics with the virtual, the physical, and the mental simultaneously. To do that, Bayesian Inference was employed and interest was indirectly estimated via activity. First, interest was modelled through the Ornstein-Uhlenbeck process. Subsequently, concepts from Stochastic Volatility models were employed to improve the performance of the base model. Through a self adjusting transfer function, a prototype statistical procedure was proposed that dynamically transformed interest into activity. The contributions were then combined and numerical estimates of interest were provided via Particle filter. To test the feasibility of the work, experimentation was performed with real datasets. Further, the method was implemented as a RESTful Web Service and the entire application was hosted on several Virtual Machines with XENServer as the base hypervisor. The experimentation clearly showed the superior capability of the proposed method to model interest. At much the same time, the analysis also revealed a few useful insights. Lastly, a few shortcomings with the work were discussed in detail.

Chapter 6

Conclusion and Future Work

In this thesis, we explored and tried to study a few humanistic attributes. The aim of the work presented in this thesis was to understand them via computational methods. To do this, the central idea behind the work revolved around the intersection of psychology and machine learning. We looked into human behavior and made an attempt to quantify it using machine driven algorithms. The knowledge organized in this thesis can be classified as an application of data science and psychology to quantify and understand a few internal states of human beings. The expertise gained through the investigation conducted in this thesis enhances the existing knowledge base and augments the current computational understanding of literature. We tried to show that there is a different way of looking at people and understanding their internal properties. Though, the observations presented in the thesis are not perfect and, at this point, cannot be generalized to a large population and all platforms, the ideas showed that there is a possibility that an artificial agent can understand and interpret unique human properties. Furthermore, although the investigation conducted in this thesis is not cent-percent accurate, the ideas, the analysis, and the guidelines nevertheless can pave the way for future research in artificial man-machine systems.

6.1 Objectives Addressed

The objective of this thesis is to achieve an understanding of the human psyche at online systems. In this regard, the list of contributions of this thesis is summarized in the following points:

1. We focused our efforts on finding, devising, and engineering efficient means of addressing the challenge of motivating users, generating their interest, and enhancing their participation. We complemented work in the discipline of psychology by devising computational ways of simulating and understanding a few mental properties in artificial environments. Owing to the significant nature of the problem, the issue was broken down into two different parts. We did this step-by-step using the following two approaches:

- For the first part, we proposed a recruitment procedure that utilized concepts from statistics and engineered an automatic sequence of steps that made the procedure of candidate recruitment more reliable and efficient. We argued that a human being do not have any

obligation to accept a request, that is, the individual does not have any duty to respond to each and every request. Hence, we have to find efficient ways of maximizing user participation by respecting this particular human weakness/constraint. As a result, and to handle this issue, we presented a statistical perspective to look at the problem of crowd selection. The aim was to maximize user participation by selecting the most optimal and most reliable set of candidates. The method was validated via conducting simulations on StackOverflow datasets.

- For the second part of the same issue, we dug deep into human psychology. This was done to find alternate, efficient, and more productive means of enhancing user participation. We studied a few psychological properties and presented several arguments that showed how targeting specific human attributes can enhance user participation. We formulated different hypotheses, investigated several aspects of human psychology, and tried to support versus negate the original belief. The investigation on the original ideas also revealed a few stimulating patterns in the crowd's behavior. We attempted to back each observations by grounding the finding in well tested psychological theories. It was shown that we can indeed look beyond the curtain of conventional approaches and can approach the problem from a rather unprecedented stance. Based on the analysis performed in Chapter 4, we presented a set of recommendations that showed how attacking the psychological properties can get the job done in a more cost-effective and labour-saving way. The discussion in the Chapter concluded that we have to loose the ideal laboratory mentality and have to develop a more practical ideology at crowd systems. Lastly, the shortcomings of the work was discussed in detail.

2. The last proposal of the thesis was directly linked to the previous two contributions. For the previous issue, the motive was to device methods and find efficient techniques to generate interest and enhance user participation. Subsequently, the aim became: *How to quantify human interest?* To handle this problem, we proposed a novel method that could computationally model a person's interest. The method drew inspiration from multiple disciplines and made the procedure of interest quantification automatic. The research conducted on this challenging problem complemented and increased current knowledge in many ways. i) We proposed a computationally feasible definition to calculate activity. We were able to combine different perspectives of activity into a single and computationally operable construct. ii) We proposed a framework to model the dynamics of interest. We did this by drawing inspiration from Physics and Economics. iii) We presented a model to dynamically convert interest into activity. This was done by utilizing concepts from Adaptive filtering. iv) We presented a method that could quantify interest towards any object and at any point of time (second, minutes, hours etc). The method was validated by conducting numerical investigation on StackOverflow datasets. The anlysis and the experiences led us to investigate the model in more detail, thereby revealing a few useful insights.

6.2 Future Work

The work presented in this thesis merely scratched the surface on understanding the internal properties of human beings. Through the work presented in this thesis, we presented a fresh set of guidelines and discussed a few theoretical recommendations that could be applied to a variety of fields (not limited to crowd oriented systems). The discussion also took the notion of psychological computing [10] one step forward. This was done by complementing it with machine learning. In this regard, and with respect to the ideas presented in this thesis, we aim to extend the work at this lucrative and interesting intersection of psychology and machine learning in the following ways:

1. For the first issue handled in this thesis, we aim to improve the work in the following ways:
   - The analysis on the psychological properties showed that there other non conventional ways through which we can promote user participation. An important point of note about the study in Chapter 4 was that the ideas opened several new directions for researchers working in interdisciplinary disciplines. For example, we conducted our analysis on StackOverflow datasets (which is an example of technical crowdsourcing). Therefore, researchers in sociology and psychology can take the ideas further and into the domain of specialized professionals, preferably technical users, and explore their behavior online. We have to appreciate that the behavior in an online environment is different from that in an office environment. In our future work, we aim to put this idea under scrutiny.
   - Research efforts in crowdsourcing can benefit from this work by following the ideas to select the most appropriate set of candidates. Crowd workers usually comprise of a niche group that contributes to crowdsourcing tasks. The assumption is that workers are thorough professional and are neutral and unbiased in their approach. However, we highlighted a few issues that negate this notion. Building upon the observations discussed in the thesis, we aim to explore the humanistic point of view in more detail.

2. For the second problem addressed in this thesis, that is, to estimate interest using machine driven algorithms, there are a few additional challenges. We have identified the following areas to improve for our future work.
   - In Chapter 5, Interest was modeled via the Stochastic Volatility based OU process. However, it is not claimed here that the method is accurate. We must reiterate here that the study has tried to take one more step towards finding a solution to the interest estimation problem. However, more analysis is needed to refine the base model. Moreover, the efforts need not be limited to mean reverting procedures.
   - It is clear from the discussion in chapter 5 that we estimate interest via activity. An important point of note here is that in case there is no reward, activity may reduce, and interest may flag. Hence, there is a need to accommodate this type of feedback mechanism in estimating the property of interest. It must be specified here that the goal of the chapter was to merely estimate interest from activity. Furthermore, although we assumed that activity is already a consequence of various rewarding

mechanisms, we did not accommodate the feedback mechanism in estimating interest from activity. This is next area we intend to pursue in the future.

- The last area to work on is the way interest transforms into activity. The function was approximated dynamically using Adaptive filtering techniques, RLS. However, it is not claimed here that this procedure is accurate. It must be specified here that similar to modeling interest, the work presented in Chapter 5 (subsection 5.3.4) has tried to model a non-trivial phenomenon. Though, the method is acceptable from a machine's perspective, however, more refinements are required. It should be noted here that for this particular problem, simultaneous inputs from psychology are needed. This is because the problem deals at the intersection between Man and Machine, hence, we cannot approach the problem purely from a Machine's point of view.

Bibliography

[1] R. M. Ryan and E. L. Deci, "Intrinsic and extrinsic motivations: Classic definitions and new directions," *Contemporary educational psychology*, vol. 25, no. 1, pp. 54–67, 2000.

[2] E. L. Deci and R. M. Ryan, "Overview of self-determination theory: An organismic dialectical perspective," *Handbook of self-determination research*, pp. 3–33, 2002.

[3] R. M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being." *American psychologist*, vol. 55, no. 1, p. 68, 2000.

[4] R. Baheti and H. Gill, "Cyber-physical systems," *The impact of control technology*, vol. 12, pp. 161–166, 2011.

[5] D. E. Rumelhart *et al.*, *Parallel distributed processing*, vol. 1.

[6] H. Hong, "The effects of human interest framing in television news coverage of medical advances," *Health communication*, vol. 28, no. 5, pp. 452–460, 2013.

[7] J. M. L. Asensio, J. Peralta, R. Arrabales, M. G. Bedia, P. Cortez, and A. L. Peña, "Artificial intelligence approaches for the generation and assessment of believable human-like behaviour in virtual characters," *Expert Systems with Applications*, vol. 41, no. 16, pp. 7281–7290, 2014.

[8] I. Millington and J. Funge, *Artificial intelligence for games*. CRC Press, 2016.

[9] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[10] X. Bao, M. Gowda, R. Mahajan, and R. R. Choudhury, "The case for psychological computing," in *Proceedings of the 14th Workshop on Mobile Computing Systems and Applications*. ACM, 2013, p. 6.

[11] N. Kaufmann, T. Schulze, and D. Veit, "More than fun and money. worker motivation in crowdsourcing-a study on mechanical turk." in *AMCIS*, vol. 11, 2011, pp. 1–11.

[12] L. Micallef, P. Dragicevic, and J.-D. Fekete, "Assessing the effect of visualizations on bayesian reasoning through crowdsourcing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2536–2545, 2012.

[13] N. Immorlica, G. Stoddard, and V. Syrgkanis, "Social status and badge design," in *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015, pp. 473–483.

[14] H. Zheng, D. Li, and W. Hou, "Task design, motivation, and participation in crowdsourcing contests," *International Journal of Electronic Commerce*, vol. 15, no. 4, pp. 57–88, 2011.

[15] T. D. Wilson, J. G. Hull, and J. Johnson, "Awareness and self-perception: Verbal reports on internal states." *Journal of personality and Social Psychology*, vol. 40, no. 1, p. 53, 1981.

[16] J. Howe, "The rise of crowdsourcing," *Wired magazine*, vol. 14, no. 6, pp. 1–4, 2006.

[17] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: current state and future challenges," *IEEE Communications Magazine*, vol. 49, no. 11, 2011.

[18] T. Yan, M. Marzilli, R. Holmes, D. Ganesan, and M. Corner, "mcrowd: a platform for mobile crowdsourcing," in *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*. ACM, 2009, pp. 347–348.

[19] L. Kazemi and C. Shahabi, "Geocrowd: enabling query answering with spatial crowdsourcing," in *Proceedings of the 20th international conference on advances in geographic information systems*. ACM, 2012, pp. 189–198.

[20] V. Agarwal, N. Banerjee, D. Chakraborty, and S. Mittal, "Usense–a smartphone middleware for community sensing," in *Mobile Data Management (MDM), 2013 IEEE 14th International Conference on*, vol. 1. IEEE, 2013, pp. 56–65.

[21] G. Cardone, L. Foschini, P. Bellavista, A. Corradi, C. Borcea, M. Talasila, and R. Curtmola, "Fostering participaction in smart cities: a geo-social crowdsensing platform," *Communications Magazine, IEEE*, vol. 51, no. 6, pp. 112–119, 2013.

[22] X. Hu, T. Chu, H. Chan, and V. Leung, "Vita: A crowdsensing-oriented mobile cyber-physical system," *Emerging Topics in Computing, IEEE Transactions on*, vol. 1, no. 1, pp. 148–165, 2013.

[23] M.-R. Ra, B. Liu, T. F. La Porta, and R. Govindan, "Medusa: A programming framework for crowd-sensing applications," in *Proceedings of the 10th international conference on Mobile systems, applications, and services*. ACM, 2012, pp. 337–350.

[24] B. Guo, Z. Wang, Z. Yu, Y. Wang, N. Y. Yen, R. Huang, and X. Zhou, "Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm," *ACM Computing Surveys (CSUR)*, vol. 48, no. 1, p. 7, 2015.

[25] S. Reicher, "The psychology of crowd dynamics," *Blackwell handbook of social psychology: Group processes*, pp. 182–208, 2001.

[26] J. N. Baron and D. M. Kreps, *Strategic human resources: Frameworks for general managers*. Wiley New York, 1999.

[27] U. Schiefele, "Interest, learning, and motivation," *Educational psychologist*, vol. 26, no. 3-4, pp. 299–323, 1991.

[28] S. Hidi and K. A. Renninger, "The four-phase model of interest development," *Educational psychologist*, vol. 41, no. 2, pp. 111–127, 2006.

[29] P. A. Alexander and P. K. Murphy, "Profiling the differences in students' knowledge, interest, and strategic processing." *Journal of educational psychology*, vol. 90, no. 3, p. 435, 1998.

[30] D. I. Cordova and M. R. Lepper, "Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice." *Journal of educational psychology*, vol. 88, no. 4, p. 715, 1996.

[31] M. Mitchell, "Situational interest: Its multifaceted structure in the secondary school mathematics classroom." *Journal of educational psychology*, vol. 85, no. 3, p. 424, 1993.

[32] P. Qvarfordt, D. Beymer, and S. Zhai, "Realtourist–a study of augmenting human-human and human-computer dialogue with eye-gaze overlay," in *Human-Computer Interaction-INTERACT 2005*. Springer, 2005, pp. 767–780.

[33] K. Renninger and R. H. Wozniak, "Effect of interest on attentional shift, recognition, and recall in young children." *Developmental Psychology*, vol. 21, no. 4, p. 624, 1985.

[34] B. Schuller, R. Müeller, B. Höernler, A. Höethker, H. Konosu, and G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations," in *Proceedings of the 9th international conference on Multimodal interfaces*. ACM, 2007, pp. 30–37.

[35] J. Herbart, "General theory of pedagogy, derived from the purpose of education," *Writings on education*, vol. 2, pp. 9–155, 1965.

[36] A. Vespignani, "Predicting the behavior of techno-social systems," *Science*, vol. 325, no. 5939, pp. 425–428, 2009.

[37] E. L. Deci and R. M. Ryan, *Intrinsic motivation and self-determination in human behavior*. Springer Science & Business Media, 1985.

[38] R. M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being." *American psychologist*, vol. 55, no. 1, p. 68, 2000.

[39] T. Zhou, H. A.-T. Kiet, B. J. Kim, B.-H. Wang, and P. Holme, "Role of activity in human dynamics," *EPL (Europhysics Letters)*, vol. 82, no. 2, p. 28002, 2008.

[40] S. Hidi, "Interest and its contribution as a mental resource for learning," *Review of Educational research*, vol. 60, no. 4, pp. 549–571, 1990.

[41] R. C. Anderson, "Interestingness of children's reading material," *Center for the Study of Reading Technical Report; no. 323*, 1984.

[42] P. J. Hinds, "The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance." *Journal of Experimental Psychology: Applied*, vol. 5, no. 2, p. 205, 1999.

[43] S. A. Paul, L. Hong, and E. H. Chi, "Is twitter a good place for asking questions? a characterization study." in *ICWSM*, 2011.

[44] J. Nichols and J.-H. Kang, "Asking questions of targeted strangers on social networks," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 2012, pp. 999–1002.

[45] M. R. Morris, J. Teevan, and K. Panovich, "A comparison of information seeking using search engines and social networks." *ICWSM*, vol. 10, pp. 23–26, 2010.

[46] S. Reddy, D. Estrin, and M. Srivastava, "Recruitment framework for participatory sensing data collections," in *Pervasive Computing*. Springer, 2010, pp. 138–155.

[47] S. Hachem, A. Pathak, and V. Issarny, "Service-oriented middleware for large-scale mobile participatory sensing," *Pervasive and Mobile Computing*, vol. 10, pp. 66–82, 2014.

[48] L. G. Jaimes, I. Vergara-Laurens, and A. Chakeri, "Spread, a crowd sensing incentive mechanism to acquire better representative samples," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*. IEEE, 2014, pp. 92–97.

[49] M. Xiao, J. Wu, H. Huang, L. Huang, and C. Hu, "Deadline-sensitive user recruitment for probabilistically collaborative mobile crowdsensing," in *2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2016, pp. 721–722.

[50] H. Xiong, D. Zhang, L. Wang, and H. Chaouchi, "Emc 3: Energy-efficient data transfer in mobile crowdsensing under full coverage constraint," *IEEE Transactions on Mobile Computing*, vol. 14, no. 7, pp. 1355–1368, 2015.

[51] L. Wang, D. Zhang, and H. Xiong, "effsense: energy-efficient and cost-effective data uploading in mobile crowdsensing," in *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. ACM, 2013, pp. 1075–1086.

[52] H. Amintoosi and S. S. Kanhere, "A reputation framework for social participatory sensing systems," *Mobile Networks and Applications*, vol. 19, no. 1, pp. 88–100, 2014.

[53] ——, "A trust-based recruitment framework for multi-hop social participatory sensing," in *Distributed Computing in Sensor Systems (DCOSS), 2013 IEEE International Conference on*. IEEE, 2013, pp. 266–273.

[54] H. Amintoosi, S. S. Kanhere, and M. Allahbakhsh, "Trust-based privacy-aware participant selection in social participatory sensing," *Journal of Information Security and Applications*, vol. 20, pp. 11–25, 2015.

[55] L. Dabbish, C. Stuart, J. Tsay, and J. Herbsleb, "Social coding in github: transparency and collaboration in an open software repository," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 2012, pp. 1277–1286.

[56] P. Morrison and E. Murphy-Hill, "Is programming knowledge related to age? an exploration of stack overflow," in *Mining Software Repositories (MSR), 2013 10th IEEE Working Conference on*. IEEE, 2013, pp. 69–72.

[57] C. Treude, O. Barzilay, and M.-A. Storey, "How do programmers ask and answer questions on the web?: Nier track," in *Software Engineering (ICSE), 2011 33rd International Conference on*. IEEE, 2011, pp. 804–807.

[58] S. Wang, D. Lo, and L. Jiang, "An empirical study on developer interactions in stackoverflow," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*. ACM, 2013, pp. 1019–1024.

[59] S. M. Nasehi, J. Sillito, F. Maurer, and C. Burns, "What makes a good code example?: A study of programming q&a in stackoverflow," in *Software Maintenance (ICSM), 2012 28th IEEE International Conference on*. IEEE, 2012, pp. 25–34.

[60] A. Barua, S. W. Thomas, and A. E. Hassan, "What are developers talking about? an analysis of topics and trends in stack overflow," *Empirical Software Engineering*, vol. 19, no. 3, pp. 619–654, 2014.

[61] M. Linares-Vásquez, G. Bavota, M. Di Penta, R. Oliveto, and D. Poshyvanyk, "How do api changes trigger stack overflow discussions? a study on the android sdk," in *proceedings of the 22nd International Conference on Program Comprehension*. ACM, 2014, pp. 83–94.

[62] C. Rosen and E. Shihab, "What are mobile developers asking about? a large scale study using stack overflow," *Empirical Software Engineering*, vol. 21, no. 3, pp. 1192–1223, 2016.

[63] E. C. Campos, L. B. Souza, and M. d. A. Maia, "Searching crowd knowledge to recommend solutions for api usage tasks," *Journal of Software: Evolution and Process*, 2016.

[64] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Discovering value from community activity on focused question answering sites: A case study of stack overflow," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2012, pp. 850–858.

[65] B. Bazelli, A. Hindle, and E. Stroulia, "On the personality traits of stackoverflow users." in *ICSM*, 2013, pp. 460–463.

[66] M. Asaduzzaman, A. S. Mashiyat, C. K. Roy, and K. A. Schneider, "Answering questions about unanswered questions of stack overflow," in *Proceedings of the 10th Working Conference on Mining Software Repositories*. IEEE Press, 2013, pp. 97–100.

[67] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann, "Design lessons from the fastest q&a site in the west," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2011, pp. 2857–2866.

[68] D. Movshovitz-Attias, Y. Movshovitz-Attias, P. Steenkiste, and C. Faloutsos, "Analysis of the reputation system and user contributions on a question answering website: Stackoverflow," in *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*. IEEE, 2013, pp. 886–893.

[69] B. Vasilescu, A. Capiluppi, and A. Serebrenik, "Gender, representation and online participation: A quantitative study," *Interacting with Computers*, p. iwt047, 2013.

[70] D. Schenk and M. Lungu, "Geo-locating the knowledge transfer in stackoverflow," in *Proceedings of the 2013 International Workshop on Social Software Engineering*. ACM, 2013, pp. 21–24.

[71] A. Marder, "Stack overflow badges and user behavior: an econometric approach," in *Proceedings of the 12th Working Conference on Mining Software Repositories*. IEEE Press, 2015, pp. 450–453.

[72] D. Posnett, E. Warburg, P. Devanbu, and V. Filkov, "Mining stack exchange: Expertise is evident from initial contributions," in *Social Informatics (SocialInformatics), 2012 International Conference on*. IEEE, 2012, pp. 199–204.

[73] A. Pal, F. M. Harper, and J. A. Konstan, "Exploring question selection bias to identify experts and potential experts in community question answering," *ACM Transactions on Information Systems (TOIS)*, vol. 30, no. 2, p. 10, 2012.

[74] C. Santos, G. Kuk, F. Kon, and J. Pearson, "The attraction of contributors in free and open source software projects," *The Journal of Strategic Information Systems*, vol. 22, no. 1, pp. 26–45, 2013.

[75] M. Hossain, "Users' motivation to participate in online crowdsourcing platforms," in *Innovation Management and Technology Research (ICIMTR), 2012 International Conference on*. IEEE, 2012, pp. 310–315.

[76] K. W. Spence, "Behavior theory and conditioning." 1956.

[77] O. Toubia, "Idea generation, creativity, and incentives," *Marketing Science*, vol. 25, no. 5, pp. 411–425, 2006.

[78] T. M. Amabile, B. A. Hennessey, and B. S. Grossman, "Social influences on creativity: The effects of contracted-for reward." *Journal of personality and social psychology*, vol. 50, no. 1, p. 14, 1986.

[79] B. L. Bayus, "Crowdsourcing and individual creativity over time: The detrimental effects of past success," *Available at SSRN 1667101*, 2010.

[80] M. Hossain and I. Kauranen, "Crowdsourcing: a comprehensive literature review," *Strategic Outsourcing: An International Journal*, vol. 8, no. 1, pp. 2–22, 2015.

[81] D. Geiger, S. Seedorf, T. Schulze, R. C. Nickerson, and M. Schader, "Managing the crowd: Towards a taxonomy of crowdsourcing processes." in *AMCIS*, 2011.

[82] N. F. Noy, A. Chugh, and H. Alani, "The ckc challenge: Exploring tools for collaborative knowledge construction," *Intelligent Systems, IEEE*, vol. 23, no. 1, pp. 64–68, 2008.

[83] D. Easley and A. Ghosh, "Incentives, gamification, and game theory: an economic approach to badge design," in *Proceedings of the fourteenth ACM conference on Electronic commerce*. ACM, 2013, pp. 359–376.

[84] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 61–70.

[85] S. Hidi and W. Baird, "Strategies for increasing text-based interest and students' recall of expository texts," *Reading Research Quarterly*, pp. 465–483, 1988.

[86] ——, "Interestingness—a neglected variable in discourse processing," *Cognitive Science*, vol. 10, no. 2, pp. 179–194, 1986.

[87] A. Juarrero, *Dynamics in action: Intentional behavior as a complex system*. Citeseer, 1999, vol. 31.

[88] K. Renninger and R. H. Wozniak, "Effect of interest on attentional shift, recognition, and recall in young children." *Developmental Psychology*, vol. 21, no. 4, p. 624, 1985.

[89] R. J. Davidson, "The neuroscience of affective style," *The new cognitive neurosciences*, vol. 2, pp. 1149–1159, 2000.

[90] J. LeDoux, "Cognitive-emotional interactions: Listen to the brain," *Cognitive neuroscience of emotion*, pp. 129–155, 2000.

[91] J. Dewey, *Interest and effort in education*. Houghton Mifflin, 1913.

[92] S. M. Fulmer, S. K. D'Mello, A. Strain, and A. C. Graesser, "Interest-based text preference moderates the effect of text difficulty on engagement and learning," *Contemporary Educational Psychology*, vol. 41, pp. 98–110, 2015.

[93] M. Wöllmer, F. Weninger, F. Eyben, and B. Schuller, "Acoustic-linguistic recognition of interest in speech with bottleneck-blstm nets." in *INTERSPEECH*, 2011, pp. 77–80.

[94] J. H. Jeon, R. Xia, and Y. Liu, "Level of interest sensing in spoken dialog using multi-level fusion of acoustic and lexical evidence." in *INTERSPEECH*, 2010, pp. 2802–2805.

[95] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 677–682.

[96] A. Kapoor, R. W. Picard, and Y. Ivanov, "Probabilistic combination of multiple modalities to detect interest," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3. IEEE, 2004, pp. 969–972.

[97] S. Mota and R. W. Picard, "Automated posture analysis for detecting learner's interest level," in *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, vol. 5. IEEE, 2003, pp. 49–49.

[98] T. Hirayama, J.-B. Dodane, H. Kawashima, and T. Matsuyama, "Estimates of user interest using timing structures between proactive content-display updates and eye movements," *IEICE TRANSACTIONS on Information and Systems*, vol. 93, no. 6, pp. 1470–1478, 2010.

[99] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *Multimedia, IEEE Transactions on*, vol. 8, no. 3, pp. 500–508, 2006.

[100] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing*, vol. 27, no. 12, pp. 1760–1774, 2009.

[101] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *International journal of human-computer studies*, vol. 65, no. 8, pp. 724–736, 2007.

[102] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous *et al.*, "Combining efforts for improving automatic classification of emotional user states," *Proc. IS-LTC*, pp. 240–245, 2006.

[103] ——, "Combining efforts for improving automatic classification of emotional user states," *Proc. IS-LTC*, pp. 240–245, 2006.

[104] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon, "The painful face–pain expression recognition using active appearance models," *Image and vision computing*, vol. 27, no. 12, pp. 1788–1796, 2009.

[105] R. W. White, P. Bailey, and L. Chen, "Predicting user interests from contextual information," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009, pp. 363–370.

[106] W. Lam, S. Mukhopadhyay, J. Mostafa, and M. Palakal, "Detection of shifts in user interests for personalized information filtering," in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1996, pp. 317–325.

[107] W. Lam and J. Mostafa, "Modeling user interest shift using a bayesian approach," *Journal of the American society for Information Science and Technology*, vol. 52, no. 5, pp. 416–429, 2001.

[108] D. H. Widyantoro, T. R. Ioerger, and J. Yen, "Learning user interest dynamics with a three-descriptor representation," *Journal of the American Society for Information Science and Technology*, vol. 52, no. 3, pp. 212–225, 2001.

[109] P. N. Bennett, R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk, and X. Cui, "Modeling the impact of short-and long-term behavior on search personalization," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2012, pp. 185–194.

[110] F. Qiu and J. Cho, "Automatic identification of user interest for personalized search," in *Proceedings of the 15th international conference on World Wide Web*. ACM, 2006, pp. 727–736.

[111] R. W. White, P. N. Bennett, and S. T. Dumais, "Predicting short-term interests using activity-based search context," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 1009–1018.

[112] Z. Ma, Q. Dai, and N. Liu, "Several novel evaluation measures for rank-based ensemble pruning with applications to time series prediction," *Expert Systems with Applications*, vol. 42, no. 1, pp. 280–292, 2015.

[113] Z. Ma and Q. Dai, "Selected an stacking elms for time series prediction," *Neural Processing Letters*, pp. 1–26, 2016.

[114] N. K. Kasabov and Q. Song, "Denfis: dynamic evolving neural-fuzzy inference system and its application for time-series prediction," *IEEE Transactions on fuzzy systems*, vol. 10, no. 2, pp. 144–154, 2002.

[115] A. Aue, D. D. Norinho, and S. Hörmann, "On the prediction of stationary functional time series," *Journal of the American Statistical Association*, vol. 110, no. 509, pp. 378–392, 2015.

[116] Y. Zha, T. Zhou, and C. Zhou, "Unfolding large-scale online collaborative human dynamics," *arXiv preprint arXiv:1507.05248*, 2015.

[117] O. Kwon, W.-S. Son, and W.-S. Jung, "The double power law in human collaboration behavior: The case of wikipedia," *Physica A: Statistical Mechanics and its Applications*, vol. 461, pp. 85–91, 2016.

[118] Z.-D. Zhao, Y.-C. Gao, S.-M. Cai, and T. Zhou, "Dynamic patterns of academic forum activities," *Physica A: Statistical Mechanics and its Applications*, vol. 461, pp. 117–124, 2016.

[119] A. Kan, J. Chan, C. Hayes, B. Hogan, J. Bailey, and C. Leckie, "A time decoupling approach for studying forum dynamics," *World Wide Web*, vol. 16, no. 5-6, pp. 595–620, 2013.

[120] J. Chan, C. Hayes, and E. M. Daly, "Decomposing discussion forums and boards using user roles." *ICWSM*, vol. 10, pp. 215–218, 2010.

[121] F. B. Viégas and M. Smith, "Newsgroup crowds and authorlines: Visualizing the activity of individuals in conversational cyberspaces," in *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*. IEEE, 2004, pp. 10–pp.

[122] R. Xiong and J. Donath, "Peoplegarden: creating data portraits for users," in *Proceedings of the 12th annual ACM symposium on User interface software and technology*. ACM, 1999, pp. 37–44.

[123] Y. Wu, C. Zhou, J. Xiao, J. Kurths, and H. J. Schellnhuber, "Evidence for a bimodal distribution in human communication," *Proceedings of the national academy of sciences*, vol. 107, no. 44, pp. 18 803–18 808, 2010.

[124] P.-Y. Oudeyer and F. Kaplan, "What is intrinsic motivation? a typology of computational approaches," *Frontiers in neurorobotics*, vol. 1, p. 6, 2007.

[125] Z.-D. Zhao, Z. Yang, Z. Zhang, T. Zhou, Z.-G. Huang, and Y.-C. Lai, "Emergence of scaling in human-interest dynamics," *Scientific reports, Nature*, vol. 3, 2013.

[126] G. E. Uhlenbeck and L. S. Ornstein, "On the theory of the brownian motion," *Physical review*, vol. 36, no. 5, p. 823, 1930.

[127] O. Vasicek, "An equilibrium characterization of the term structure," *Journal of financial economics*, vol. 5, no. 2, pp. 177–188, 1977.

[128] F. E. Benth, J. Kallsen, and T. Meyer-Brandis, "A non-gaussian ornstein–uhlenbeck process for electricity spot price modeling and derivatives pricing," *Applied Mathematical Finance*, vol. 14, no. 2, pp. 153–169, 2007.

[129] P. Lánskỳ and S. Sato, "The stochastic diffusion models of nerve membrane depolarization and interspike interval generation." *Journal of the peripheral nervous system: JPNS*, vol. 4, no. 1, pp. 27–42, 1998.

[130] S. Ditlevsen and P. Lansky, "Estimation of the input parameters in the ornstein-uhlenbeck neuronal model," *Physical review E*, vol. 71, no. 1, p. 011907, 2005.

[131] J. M. Beaulieu, D.-C. Jhwueng, C. Boettiger, and B. C. O'Meara, "Modeling stabilizing selection: expanding the ornstein–uhlenbeck model of adaptive evolution," *Evolution*, vol. 66, no. 8, pp. 2369–2383, 2012.

[132] S. L. Heston, "A closed-form solution for options with stochastic volatility with applications to bond and currency options," *Review of financial studies*, vol. 6, no. 2, pp. 327–343, 1993.

[133] J.-P. Fouque, G. Papanicolaou, and K. R. Sircar, "Mean-reverting stochastic volatility," *International Journal of theoretical and applied finance*, vol. 3, no. 01, pp. 101–142, 2000.

[134] Y. L. Sun, W. Yu, Z. Han, and K. Liu, "Information theoretic framework of trust modeling and evaluation for ad hoc networks," *Selected Areas in Communications, IEEE Journal on*, vol. 24, no. 2, pp. 305–317, 2006.

[135] G. E. Box and G. C. Tiao, *Bayesian inference in statistical analysis*. John Wiley & Sons, 2011, vol. 40.

[136] B. S. Clarke and A. R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," *Journal of Statistical planning and Inference*, vol. 41, no. 1, pp. 37–60, 1994.

[137] R. M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being." *American psychologist*, vol. 55, no. 1, p. 68, 2000.

[138] R. J. Vallerand and R. Blssonnette, "Intrinsic, extrinsic, and amotivational styles as predictors of behavior: A prospective study," *Journal of personality*, vol. 60, no. 3, pp. 599–620, 1992.

[139] R. M. Ryan and E. L. Deci, "Intrinsic and extrinsic motivations: Classic definitions and new directions," *Contemporary educational psychology*, vol. 25, no. 1, pp. 54–67, 2000.

[140] E. G. Clary, M. Snyder, R. D. Ridge, J. Copeland, A. A. Stukas, J. Haugen, and P. Miene, "Understanding and assessing the motivations of volunteers: a functional approach." *Journal of personality and social psychology*, vol. 74, no. 6, p. 1516, 1998.

[141] K. Liu and X. Li, "Finding nemo: Finding your lost child in crowds via mobile crowd sensing," in *Mobile Ad Hoc and Sensor Systems (MASS), 2014 IEEE 11th International Conference on*. IEEE, 2014, pp. 1–9.

[142] N. Gantayat, P. Dhoolia, R. Padhye, S. Mani, and V. S. Sinha, "The synergy between voting and acceptance of answers on stackoverflow, or the lack thereof," in *Proceedings of the 12th Working Conference on Mining Software Repositories*. IEEE Press, 2015, pp. 406–409.

[143] X. Hu, T. Chu, H. Chan, and V. Leung, "Vita: A crowdsensing-oriented mobile cyber-physical system," *Emerging Topics in Computing, IEEE Transactions on*, vol. 1, no. 1, pp. 148–165, 2013.

[144] E. Bradner and G. Mark, "Why distance matters: effects on cooperation, persuasion and deception," in *Proceedings of the 2002 ACM conference on Computer supported cooperative work*. ACM, 2002, pp. 226–235.

[145] E. Locke and G. Latham, *Goal-setting theory*, 1994.

[146] Z. Kunda and S. H. Schwartz, "Undermining intrinsic moral motivation: External reward and self-presentation." *Journal of Personality and Social Psychology*, vol. 45, no. 4, p. 763, 1983.

[147] C. D. Batson, J. S. Coke, M. L. Jasnoski, and M. Hanson, "Buying kindness: Effect of an extrinsic incentive for helping on perceived altruism," *Personality and Social Psychology Bulletin*, vol. 4, no. 1, pp. 86–91, 1978.

[148] J. P. Meyer and N. J. Allen, "A three-component conceptualization of organizational commitment," *Human resource management review*, vol. 1, no. 1, pp. 61–89, 1991.

[149] P. Bateman, P. Gray, and B. Butler, "Community commitment: How affect, obligation, and necessity drive online behaviors," *ICIS 2006 Proceedings*, p. 63, 2006.

[150] M. Sunnafrank, "Predicted outcome value during initial interactions a reformulation of uncertainty reduction theory," *Human Communication Research*, vol. 13, no. 1, pp. 3–33, 1986.

[151] T. Althoff, C. Danescu-Niculescu-Mizil, and D. Jurafsky, "How to ask for a favor: A case study on the success of altruistic requests," *arXiv preprint arXiv:1405.3282*, 2014.

[152] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts, "A computational approach to politeness with application to social factors," *arXiv preprint arXiv:1306.6078*, 2013.

[153] Y.-M. Wang and Y. Luo, "Integration of correlations with standard deviations for determining attribute weights in multiple attribute decision making," *Mathematical and Computer Modelling*, vol. 51, no. 1, pp. 1–12, 2010.

[154] R. Hastie and R. M. Dawes, *Rational choice in an uncertain world: The psychology of judgment and decision making*. Sage, 2010.

[155] T.-C. Wang and H.-D. Lee, "Developing a fuzzy topsis approach based on subjective weights and objective weights," *Expert Systems with Applications*, vol. 36, no. 5, pp. 8980–8985, 2009.

[156] J. Ma, Z.-P. Fan, and L.-H. Huang, "A subjective and objective integrated approach to determine attribute weights," *European journal of operational research*, vol. 112, no. 2, pp. 397–404, 1999.

[157] T. L. Saaty, "What is the analytic hierarchy process?" in *Mathematical models for decision support*. Springer, 1988, pp. 109–121.

[158] F. G. Ashby, "A stochastic version of general recognition theory," *Journal of Mathematical Psychology*, vol. 44, no. 2, pp. 310–329, 2000.

[159] P. L. Smith and D. Vickers, "The accumulator model of two-choice discrimination," *Journal of Mathematical Psychology*, vol. 32, no. 2, pp. 135–168, 1988.

[160] P. E. Kloeden, E. Platen, and H. Schurz, *Numerical solution of SDE through computer experiments*. Springer Science & Business Media, 2012.

[161] P. C. Phillips, "The structural estimation of a stochastic differential equation system," *Econometrica: Journal of the Econometric Society*, pp. 1021–1041, 1972.

[162] P. C. Phillips and J. Yu, "Maximum likelihood and gaussian estimation of continuous time models in finance," in *Handbook of financial time series*. Springer, 2009, pp. 497–530.

[163] J. Hull and A. White, "The pricing of options on assets with stochastic volatilities," *The journal of finance*, vol. 42, no. 2, pp. 281–300, 1987.

[164] L. Valdivieso, W. Schoutens, and F. Tuerlinckx, "Maximum likelihood estimation in processes of ornstein-uhlenbeck type," *Statistical Inference for Stochastic Processes*, vol. 12, no. 1, pp. 1–19, 2009.

[165] Q. Wu and C. Miao, "Curiosity: From psychology to computation," *ACM Computing Surveys (CSUR)*, vol. 46, no. 2, p. 18, 2013.

[166] S. S. Haykin, *Adaptive filter theory*. Pearson Education, 2008.

[167] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *Signal Processing, IEEE Transactions on*, vol. 50, no. 2, pp. 174–188, 2002.

[168] A. Bosu, C. S. Corley, D. Heaton, D. Chatterji, J. C. Carver, and N. A. Kraft, "Building reputation in stackoverflow: an empirical investigation," in *Proceedings of the 10th Working Conference on Mining Software Repositories*. IEEE Press, 2013, pp. 89–92.

[169] D. Movshovitz-Attias, Y. Movshovitz-Attias, P. Steenkiste, and C. Faloutsos, "Analysis of the reputation system and user contributions on a question answering website: Stackoverflow," in *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*. IEEE, 2013, pp. 886–893.

[170] A. D. Kaplan, J. A. OrSullivan, E. J. Sirevaag, P.-H. Lai, and J. W. Rohrbaugh, "Hidden state models for noncontact measurements of the carotid pulse using a laser doppler vibrometer," *Biomedical Engineering, IEEE Transactions on*, vol. 59, no. 3, pp. 744–753, 2012.

[171] X. Hu, V. Nenov, M. Bergsneider, T. C. Glenn, P. Vespa, and N. Martin, "Estimation of hidden state variables of the intracranial system using constrained nonlinear kalman filters," *Biomedical Engineering, IEEE Transactions on*, vol. 54, no. 4, pp. 597–610, 2007.

[172] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE transactions on evolutionary computation*, vol. 1, no. 1, pp. 67–82, 1997.

[173] J. Dewey, *Interest and effort in education*. Houghton Mifflin, 1913.

[174] W. Fan and A. Bifet, "Mining big data: current status, and forecast to the future," *ACM sIGKDD Explorations Newsletter*, vol. 14, no. 2, pp. 1–5, 2013.

[175] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T. M. T. Do, O. Dousse, J. Eberle, and M. Miettinen, "From big smartphone data to worldwide research: The mobile data challenge," *Pervasive and Mobile Computing*, vol. 9, no. 6, pp. 752–771, 2013.