Accepted Manuscript

Combining humans and machines for the future: A novel procedure to predict human interest

Tanveer Ahmed, Abhishek Srivastava

PII: S0167-739X(18)30135-3
DOI: https://doi.org/10.1016/j.future.2018.01.043
Reference: FUTURE 3950

To appear in: *Future Generation Computer Systems*

Received date : 31 August 2016
Revised date : 14 October 2017
Accepted date : 20 January 2018

Please cite this article as: T. Ahmed, A. Srivastava, Combining humans and machines for the future: A novel procedure to predict human interest, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2018.01.043

Combining Humans and Machines for the Future: A Novel Procedure to Predict Human Interest

Tanveer Ahmed, Abhishek Srivastava

Indian Institute of Technology Indore

{*phd12120101, asrivastava*}*@iiti.ac.in*

Abstract

This paper proposes a method to quantify interest. In common terminology, when we engage with an object, e.g. Online Games, Social Networking Websites, Mobile Apps, etc., there is a degree of interest between us and the object. But, owing to the lack of a procedure that can quantify interest, we are unable to tell by how 'much' of a factor are we interested in the object. In other words, can we find a number for someone's interest? In this article, we propose a method that uses the principle of Bayesian Inference to tackle this issue. We formulate the "interest estimation problem" as a state estimation problem to deduce interest (in any object) indirectly from user activity. Activity caused by interest is computed through a subjective objective weighted approach, then using indirect inference rules, we provide numerical estimates of interest. To do that, we model the dynamics of interest through the Ornstein-Uhlenbeck process. To further enhance the base performance, we draw inspiration from Stochastic Volatility models from Finance. Subsequently, drawing upon a self-adapting transfer function, we provide an *avant-garde* statistical procedure to model the transformation of interest into activity. The individual contributions are then combined and a solution is provided via Particle filters. Validation of the method is done in two ways. 1) Experimentation is performed on real datasets. Through numerical investigation we have found that the method shows good performance. 2) We implement the framework as a Web application and deploy it on an Enterprise Service Bus. The framework has been successfully hosted on a Cloud based Virtualized testbed consisting of several Virtual Machines constructed over XENServer as the underlying hypervisor. Through this experimental setup, we show the efficacy of the proposed algorithm in estimating interest, at much the same time, we demonstrate the viability of the method in practical cloud based deployment scenarios.

Keywords: Human Machine Systems, Data Analytics, Interest Modeling, Machine learning, Stochastic Volatility Models, Ornstein-Uhlenbeck Process.

1. Introduction

The last few years have witnessed a tremendous growth in the field of artificial intelligence. In this area, work has been trying to elevate a machine from a mere *mechanical device* to a full-fledged system capable of behaving intelligently like a human. This vision is indeed one of the most fascinating instances of researchers trying to induce human like intelligence in lifeless entities. Compelled by this foresight, there is a huge body of work dedicated to the study of stimulating human like factors in an artificial environment [1]. Moreover, the state-of-the-art developments in Data Analytics, Human Computer Interaction, and Cloud Computing have laid a foundation for these visions to become a reality, e.g. there are studies that have tried to analyze Human Relationship Dynamics [2], Dynamics of Betrayal [3], and so on. In this paper, we follow this particular line of research and focus our attention on estimating one of the variables closely linked to the human psyche. In particular, we address the issue of a machine automatically estimating the property of human interest.

Interest is an intangible mental variable that has attracted substantial research (First paper on interest was published in 1806/1965 [4]). According to [5], interest is an every day term that specifies a person's characteristic or perhaps an innate preference towards an entity, subject, or topic in the real world. It has further been specified that interest is a representative of the actions taken by an individual and is an outcome of the desire to engage with an object of one's interest [6]. Because of its relationship between the psychological and the physical being, interest has become one of most attractive topics of scientific investigation. Though, the initial days witnessed significant efforts in the discipline of Psychology, it grew from a mere mental variable to a concept of particular curiosity in Artificial Intelligence (AI). Work in this field (AI) ranges from analyzing interest in Communication (informal/formal speeches) [7], Video Watching [8], identification of a person's topic of interest [9] and so on (See section 2 for more details). Despite such a long history and a wide array of investigation, work is unable to answer one of the trivially formed questions that we have often come across in our social experience: *How much are you interested in any object, for example Instagram, Facebook, Mobile Games, TV series, and so on?* Simply put, we have raised the question that is to have a method that can quantify a person's in-

terest towards any entity at any given point of time, for example in any month, week, day, hour and so on. In lay terms, we want to find a number representing someone's interest. Moreover, we want to answer the question (i.e. estimate interest) irrespective of any application (or object). With respect to the question asked here, and if we think of the issue purely from a humanistic point of view, then we, as human beings, cannot precisely answer the question. Furthermore, it has often been speculated that interest can be felt and not quantified. It is after all a human emotion. Therefore, can a machine feel or rather understand the emotion of interest? The question we have raised here not only challenges the current state-of-the-art, but it also raises additional issues for AI. We can therefore say that to expect an answer (to the question raised here) from a *lifeless-mechanical* device is non-trivial.

Despite the significant nature of the problem, if we can device a method that can estimate a person's interest, the future research possibilities could be tremendous. For instance, computational systems could then monitor the interestingness of people toward online platforms (e.g. Facebook, Twitter etc.), employers could administer the interest of their employees in commercial projects. It is needless to say that the application of such a method could, at times, outperform our current imagination. However, and to face reality, such a procedure would require a conglomerate of multiple disciplines working together simultaneously. Moreover, it would first require us to solve the raised *Interest Estimation Problem* (IEP). This, however, is easier said than done. Numerically quantifying interest is a significant challenge. There are several issues that we have to face if we want to address the raised IEP. The issues are elaborated upon in the following points:

C1. Through our social experience, we have often noticed that interest in any object changes itself with time. For instance, consider a person interested in a Mobile game. Initially, say the person was very interested in playing the game, but, with time the person got bored and his/her interest decreased. This is a phenomenon of practical import that is experienced by most individuals. Although, the social circumstances that nourish versus forestall interest could be different, the inseparable link between interest and its dynamics being regulated as a function of everyday erratic circumstances, indeed cause several changes that so far cannot be modeled computationally. Therefore, the first challenge to the IEP is: **We have to devise a procedure that can artificially and statistically capture the long-term evolution of interest**. The method should be able to computationally capture the short-term as well as the long-term fluctuations (ups and downs) in interest. It is understood that the short-term fluctuations are caused by the daily erratic circumstances. At much the same time, the long-term evolution is owing to a person undergoing changes in his/her desire gradually over time.

C2. Literature has found out that stimulated by interest, a person is compelled to take actions towards the entity of his/her interest. For example, consider that a person is interested in playing football. It is therefore natural to expect that the person will take certain actions to express his/her interest. That is, he/she will play the game (in this case). This is one of the most extensively reported and commonly observed phenomenon in the real world [10], [11]. However, a brief insight into the term activity reveals its diverse nature. Activity is rarely a singular concept and spans a broad spectrum of viewpoints. Moreover, it is an abstact term and has a wide variety of viewpoints (depending on the context). For example, if a person is interested in playing mobile games, the perspectives of activity are: The number of hours spent playing, the number of gaming sessions in a day, the gap between two gaming sessions, and so on. Moreover, if the same person is interested in reading novels, the attributes of activity are totally different to that of mobile games. In this case (reading novels), the attributes could be: The number of pages read, number of hours spent reading, and several others. To generalize this behaviour, activity towards any object of interest spans a broad spectrum of perspectives, and therefore, such an abstract variable presents an obstacle that cannot be bypassed by considering only one point of view. Moreover, as is clear in the two use-cases, we have to consider the attributes of activity for every object of interest separately. As a result, the second challenge for the IEP is: ***We need a method that can combine the different perspectives of activity into a single and computationally operable concept. Further, this has to be done for each of object of interest separately.***

C3. The next issue is directly connected to the previous two challenges. We know that interest changes itself, moreover, it stimulates activity. However, the issue is: How does interest stimulates activity? Activity and interest are two intersecting concepts that do not occur in a vacuum. For instance, considering the first use case discussed in the previous point, where interest in the game compelled the person to play, the challenge is: How did interest compelled the person to play the mobile game? We know that the person is interested in playing the mobile game, however, how did interest dictated the actions of the individual? How to map this phenomenon (interest causing activity) computationally? In other words, we need to trace the statistical roots of interest dynamically transforming into activity. Consequently the third challenge is: ***we need a statistical procedure that can model the phenomenon of interest changing into activity.***

The issues highlighted in the above points are not only theoretically significant, but they are particularly important when it comes to computational agents capable of inducing human like intelligence in artificial environments. Moreover, we can understand the importance of the question raised here, and the potential societal impact if we can device an accurate interest estimation procedure. Though, engineering such a system is hard, in this article we make an attempt to address the issue of estimating interest through machine driven procedures. Further, in this paper our goal is also to present a potential roadmap that can solve the raised IEP. In doing so, we work at the Intersection between Psychology and AI to present a different picture that complements existing work and can model the long term evolution of interest. We present a set of mere statistical guidelines that can make a lifeless-machine automatically and systematically estimate the internal property of interest. Therefore, with this goal in mind, the contribution and the method of working

of the paper is highlighted in the following points.

1. The *modus operandi* for the IEP is formulated as a state estimation problem and interest is deduced indirectly via activity. *We use basic rules of Bayesian Inference and try to provide numerical estimates of interest indirectly from activity.*

2. To address C1, interest is modeled through the Ornstein Uhlenbeck process [12]. We use mean-reverting stochastic procedures and model the continuous evolution of interest by borrowing ideas from Physics. Moreover, we go one step ahead and draw inspiration from Finance to make the volatility component of the process stochastic, thereby making the procedure more accurate. In doing so, *we make an attempt to find order from the chaotic internal mental states of a human.*

3. To tackle the challenge highlighted in C2, *activity is computed through a subjective-objective approach* wherein the method comprises of a strategy that can incorporate the subjective and the objective nature of a person. We combine different perspectives of activity and present a numerically feasible solution to calculate activity.

4. As is expected in state estimation problems, and to address C3, we draw inspiration from Adaptive Filtering. We use the Normalized Least Mean Square Algorithm. *We present a self-configuring procedure that automatically adjusts its internal mechanisms to model the transformation of interest changing into activity.* In doing so, we follow a black box approach and let the system evolve itself automatically.

5. Lastly, to provide a computationally feasible solution to the IEP, we use particle filter. *By following the rules of Monte carlo simulations, we provide real time estimates of interest*. We filter numerical interest from numerical activity.

To demonstrate the efficacy of the proposed method, we conduct numerical tests on real datasets. Through extensive simulation studies performed on StackOverflow databases, we have found that the method shows good performance. The results are especially appreciable considering the fact that the paper has tried to address one of the significant problems of literature. Moreover, and to further validate the feasibility of the proposed method in practical cloud based deployment scenarios, we engineer a prototype. We implement the method as a Web application and deploy it on an Enterprise service bus. Furthermore, we host and execute the application on several virtual machines with XENServer as the base hypervisor.

Before we begin the discussion on the proposed method, we must reiterate here that the work in this paper has tried to model an internal mental state. To model and correspondingly quantify such a construct, especially through machines, is a challenge. In this paper, we have emphasized on simplicity, hence, we sacrifice on generality. It is not claimed here that the model covers all different aspects of interest. Interest is a multifaceted concept that is used in a variety of disciplines, contexts, and applications. We, henceforth, work along existing terms in Artificial Intelligence and Software Engineering [7], [13], [14], [15], and focus on the proposed framework as a mere starting point to let machines infer the property of interest. It has been specified in literature that "To be motivated (and interested) means to be moved to do something" [16]. In this paper, we work with the idea behind these lines. Further, the motive is to use model based approaches to assess and evaluate interest towards any entity via activity. It should be noted here that in this paper, we use advanced data analytics to estimate interest from activity.

The rest of the paper is organized as follows: In section 2, we present the necessary background on interest and highlight the related work. The proposed method is discussed in detail in Section 3. In Section 4, we present the results. Finally, we conclude with future research directions in Section 5.

2. Background and Related Work

2.1. Background

To find a plausible solution to the IEP, we explored literature across several disciplines. As we are trying to mix two distinct disciplines, we explored literature in Psychology and Artificial Intelligence. During our study on the discipline on the Human Psychology, we found that interest is categorized into "*situational*" interest and "*individual*" interest [5], [17], [18]. In the following subsections, we discuss some of the necessary theoretical foundation on these two different branches of interest.

2.1.1. Situational interest

Situational interest (it is also referred to as spontaneous interest) is the context dependent and extrinsically stimulated feeling of interestingness, for example a beautiful painting stimulating the temporary interest of the viewer. It is short lived, temporary, and stimulant dependent. In the words of the authors of [19], "*It is a kind of spontaneous interest that appears to fade as rapidly as it emerges, and it is almost always place-specific*". From these lines, we can understand that situational interest does not last long and is mostly context or place-specific. Further, studies in psychology have gone deeper and have categorized interest into many sub-branches, for a taxonomy on situational interest see [19]. Though there are several broad categories of interest and literature has focused on each of the sub-categories of situational interest in detail, in this paper, we focus only on the broad category of situational interest only. Further, we work with the convolution of situational and *individual* interest. We discuss the details of individual interest in the next subsection.

2.1.2. Individual Interest

In the previous subsection, we gave a brief account on situational interest. In this subsection, we discuss the details of individual interest. Individual interest is the feeling that is generated from one's within. It is caused by the application of the Internal desire and an individual's intrinsic factors. It is a long term, persistent, and an enduring predisposition to engage with

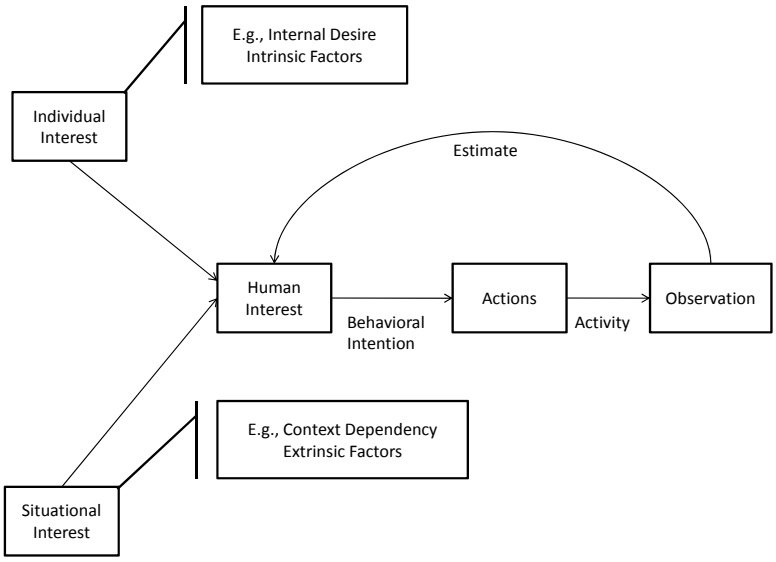

Figure 1: Conceptual Model for Interest.

a particular content or object or subject and it is an outcome of one's internal self. For example, a researcher working in the area of his/her interest for a long time. Individual interest is not affected by short term changes in the environmental factors. It should be noted here that whatever changes happen to individual interest happens in the long-term. This is in contrast to situational interest where changes occur on a moment-by-moment basis. In the previous subsection, we specified that situational interest is divided into several categories. In much the same way, research has sub-classified individual interest into several categories. They are discussed in [19]. It should be noted here that in this article, we work with the combination of the two broad categories interest (Situational & Individual) and proposed the solution by keeping in mind this broad framework of literature in psychology [5], [17], [18]. This is because both categories of interest have drawn significant attention in literature [20], [21]. In this regard, and to summarize the work in Psychology, we present the conceptual model in Fig. 1. As a result of the convolution of situational and individual interest, we propose a method that can monitor the moment-by-moment fluctuations and the long-term evolution simultaneously.

Apart from Situational and Individual interest, work has further gone deep and has found that though interest has cognitive components, it also include affective mechanisms that interact separately with the cognitive system of the human psyche [22], [23]. Furthermore, it has been specified that interest defines a person's perception and is a representative of the actions taken by a person to engage with an object of his/her interest [6]. With respect to these work, it is understandable that interest indeed has deep roots in the human mind, moreover, the core of this concept has been investigated rigorously from multiple points of view, but the inability to provide an answer to the question, How 'much' are you interested in any object, created several research gaps. More precisely, work is unable to answer the following questions: **1) How to mathematically model the long term evolution of interest? 2) How to find a 'number' for someone's interest at any given interval of time? 3) How to estimate interest towards any entity? 4) How to statistically define the conversion of interest into activity?** These few unanswered questions present an excellent opportunity to look into the matter and find a solution. Therefore, to address the issues, and to find the answers is the sole motivation behind the work presented in this paper. Moreover, we aim to address these questions through automatic data driven procedures.

2.2. Related Work

In literature, there are studies in the past that have tried to analyze interest via machine based algorithms. We found that there are two ways in which interest is analyzed. 1) Work has tried to classify one's *level of interest*. For example interest estimation has been treated as a classification problem, i.e. classifying interest as high, low, medium etc. 2) Literature has also dedicated significant efforts to determine one's *topic of interest* i.e. What type of movies does one like? For the former category, the work of Schuller et. al [7] recognizes interest in conversational speeches using Support Vector machines and the idea of 'bag of frames'. [13] proposes an automatic system for detecting one's level of interest in conversations. This is done via by combining different modes of stream analysis. There also exist proposals that focus on detecting interest through different modes of observation. For example, acoustic based interest estimation [24], video based interest detection [25], moreover, some studies have gone further and have combined multiple modes to detect one's level of interest [26]. Furthermore,

interest has also been investigated by analyzing eye movement and content presentation [27]. The studies discussed in [28], [29], [30] try to classify interest into different categories. Much like the work on detecting the level of interest, these papers focus on the idea of multiple modalities to classify interest. For the latter category, i.e. to infer one's *topic of interest*, work has dedicated its efforts to content recommendation (e.g. online product or content recommendation). In this respect, [14] tries to deduce one's topic of interest from contextual information for search and recommendation systems. The authors of [15], [31] use past history, consumption data, and search criterion to deduce a possible representation of a user's interest. [32] focuses on personalization of items using probabilistic models for web searching. The authors of [33] employ long-term and short-term user behavior to deduce one's interest in online searches. [34] uses the idea of pre-search activity to provide a richer set of information about a user's intentions, thereby offering several topics of interest in advance. Since, we are trying to deduce interest from activity, there exists work that focuses on the use of activity to deduce the dynamics of human interaction. For example, [35] tries to understand the collaborative activities at Wikipedia. The authors found the existence of a double-power law at the platform. [36] uses a multi-agent approach to understand the interaction patterns of individuals. Moreover, the authors go one step ahead and combines the study by considering the psychological and socio-cultural environments of the individual. In the analysis conducted in [37], the authors have focused on GitHub. Though, the authors have found several interesting points, however, the point of note is that the authors have found that elite-developers have long-term interest (individual interest) to participate in project on GitHub. In [38], the authors analyze posting patterns of users in news forums. In addition the authors have found symmetric inter-event time.

In sum, we must specify that the idea of analyzing interest through machine based algorithms is not new and there are many papers on the topic. However, work is not able to address the four question asked in the previous paragraph. Further, literature has dedicated its efforts to application specific and context dependent scenarios. In contrast, we focus on quantifying and modeling interest independent of any application. Moreover, we also tackle the issue of the *activity gap* i.e. we try to complement work in literature by presenting a method that can estimate interest even when there is no activity. In practical scenarios, if we are trying to estimate interest, the data about activity is not always available. Therefore, in such cases, interest estimation is out of scope. In this paper, we present a solution to this problem as well (see Section 3.6 for details on the issue).

Although the studies discussed in this section have similarity to the work proposed in this paper, the closest to our work perhaps is [39] and [9]. In [39], the authors have used an information theory framework to analyze the multidimensional concept of Intrinsic Motivation. We draw theoretical inspiration from this work, specifically the knowledge based procedures, to model and therefore infer interest in every day objects. In doing so, we go deep into data analytics. Moreover, we mix the approach with a technological framework, thereby creating a prototype psycho-techno model for interest. Although, [39] is one of the theoretical inspirations behind this paper, the closest to our work from a practical point of view is [9]. The authors of this paper have analyzed interest via data analytics. The paper is focused on three issues: 1) The duration for which interest lasts; 2) the probability distribution that model the return of a person to a previous topic of interest; 3) the transition and ranking of a user's topic of interest. With respect to these points, we must specify here that although we do not focus on any of these criteria, we have nevertheless tried to follow the initiative of the authors. Our work is similar to [9] as we have employed data analytics to study and analyze interest. However, our goal is different. In contrast to [9], where the authors focus on one's *topic of interest*, we are more interested in quantifying, at much the same time, modeling the long term evolution of interest. With respect to data analytics, the authors have specified *"As a branch of the science of "Big Data", the field of human-interest dynamics is at its infancy"*. We base the motivation of this paper along these lines and use data driven algorithms to model and estimate interest. It was specified in Section 1 that to solve the IEP, we need a conglomerate of disciplines working together. Therefore, we use current terms in literature and build upon existing work to present the proposed framework. The novelty of the paper lies in the fact that has made an attempt to bind existing work in a seamless framework, at much the same time, fix the shortcomings and has tried to bridge the gap between existing research and the practical issues in estimating interest. As a result, the paper complements the current state-of-the-art and shows future research a potential guideline to solve the issue of "numerically quantifying" interest towards "any entity".

With respect to the work in literature, we present a brief comparison of the proposed method with existing techniques in the following points:

1. Unlike the work in literature focusing on a specialized application area, e.g. speeches [7], conversations [13], videos [8], web [33], and others discussed in this subsection, we do not focus on any of these criterion. Our goal is to propose a general framework to estimate interest towards any entity. We propose an application independent interest estimation procedure.

2. Rather than formulating the problem as a classification problem [28], [29], [30] or a prediction problem [32], we formulate the problem as a "state estimation problem". We use Bayesian Inference and state estimation problem framework to estimate interest form activity.

3. We further propose a method to handle the activity gap problem. That is, we provide estimates of interest even when there is no activity. This is also one of the significant problem we have tried to address in this paper.

4. We draw heavily on the use advanced data analytics to model, and thus, estimate interest.

5. In contrast to the work in the two *distinct* disciplines of Psychology [5], [17], [18] and AI [7], [13], we work at

the intersection between Psychology and AI to propose a prototype framework to model the long-term dynamics of interest. Dynamics of interest includes temporary fluctuations and long-term gradual changes.

6. Contrary to the work in AI [7], [28], [29], [30], [33], [36], we propose a continuous time model for interest. The model is able to evolve itself automatically without any external intervention. Moreover, we provide a feedback mechanism to correct the estimated value of interest. This is done by feeding the system regular input data for activity.

In light of the points discussed above, and the work discussed in this section, we highlight the novel contribution of the paper in the following text:

i) We do not study a person's level of interest. Furthermore, we also do not infer one's topic of interest. The aim of this paper is to model the long term evolution of interest.

ii) We do not limit the scope of the work to a particular context or application. We propose a general framework to deduce and quantify interest towards any entity. Though, the limitations imposed by the current technology presents an obstacle, nevertheless, we have tried to present a generic method (See Section 5.2 for more details).

iii) We further propose a novel framework to compute activity (caused by interest) towards any entity.

iv) We focus on finding a number representing one's interest at any given time interval. For instance, we present a method to quantify a person's interest (towards WhatApp, Facebook, Twitter etc.) at any week, day, hour, and so on.

v) We present two novel statistical models in this paper. First, we present a method to capture the dynamics of interest. Second, we discuss a framework to define the conversion of interest into activity.

3. Methods

3.1. Workflow to Estimate Interest from Activity

In this subsection, we discuss the workflow used in the paper. The following points summarize the working of the method in short. It should be noted here that the details regarding each point is discussed in detail from the next subsection onwards. To understand the workflow, let consider that a person (say Alice) is interested in Facebook. Furthermore, lets also consider that we want to estimate interest of Alice on a daily basis.

1. The first process-step in the workflow is to formulate a strategy for the mode of working. That is: How should we estimate interest from activity? In this paper, we do this via Bayesian Inference. We formulate the IEP as a state estimation problem. We use model based procedures to estimate interest from activity. The details of Bayesian Inference and state estimation problem is discussed in Section 3.2.

2. The second step is to collect the attributes of activity. Considering Facebook, the possible perspectives of activity are: The number of comments, the number of edits, the number of likes, the number of hours Alice surfed Facebook and so on. These are the different perspectives (or attributes) of activity. The next objective of the system is to combine all these viewpoints into a single and computationally viable option. This is done via the proposed subjective-objective weighted approach. The method is elaborated upon in Section 3.3.

3. The next goal in the IEP is to formulate a strategy that could model the long term evolution of interest. This is because the paper proposes a model driven approach to estimate interest. Hence, the third process-step is the definition of a function that can model the evolution of interest. The model should be able to incorporate the short term fluctuations as well as the long term gradual change in interest. Furthermore, the method should have self-correcting features to automatically correct its numerical values. As specified, this is done via feeding data about activity in regular intervals. In other words, if the model deviates from its original path, new information about activity is fed to the system to make appropriate corrections. The detailed discussion on the method is presented in Section 3.4.

4. The fourth process-step in the workflow is the definition of a function that can convert interest into activity. We need this procedure because state estimation problems rely upon the foundation of this method (We discuss the details of state estimation problem in Section 3.2). We present the details of the method in Section 3.5.

5. Once we have the models for interest as well the procedure to transform interest into activity, subsequently, we need indirect inference rules to filter numerical estimates of interest from activity. In this article, this is done via Monte carlo simulations. In particular, we use particle filter. The method is discussed in Section 3.6.

6. Lastly, as we are trying to estimate interest from activity, there is an issue that activity is not always available. That is, there gaps in activity. This is called as the issue of activity gap (We present an example in Section 3.7 to better explain this issue). Therefore, to overcome this issue, we present a method in Section 3.7.

3.2. Estimating Interest: Application of Bayesian Statistics

In this subsection, we discuss the theoretical foundation of the proposed work. More specifically, we present the Bayesian perspective for the IEP. With respect to Bayesian Inference, the objective of the paper is: We have a series of activities measured for a person k over a time period $T \in (1, t)$ as: $\zeta_k = \{\zeta_k^1, \zeta_k^2, ..., \zeta_k^t\}$, here ζ_k^t is the numerical value of activity at the t^{th} unit of time. The aim of the method is to infer the Interest Vector ϱ_k, where ϱ_k is defined as: $\varrho_k = \{\varrho_k^1, \varrho_k^2, ..., \varrho_k^t\}$, here ϱ_k^t is the interest at the t^{th} unit of time.

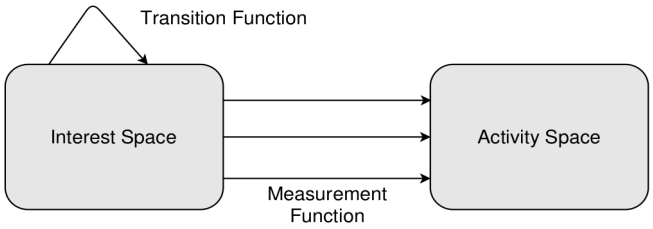

Figure 2: Bayesian Perspective For Interest Prediction Problem.

In Section 1, we pointed out that we have formulated the IEP as a state estimation problem. In this subsection, we discuss the necessary theory on state estimation problem. The theory for state estimation problems is available in many books as well as in literature [40]. In this subsection, we customize the theory and apply it to the IEP. In such type of problems, it is commonly known that the state is hidden, but the output is visible. This property is aligned with the IEP as interest (the state) is not directly observable, but, activity (the output) is. In this context, and with regard to Bayesian Inference, the foundation of the method is the definitions of the following two functions: 1) The transformation function and 2) The measurement function. In Fig. 2, we have presented the basic idea of the two functions. In the figure, we have shown two state spaces: I) The Interest space and II) The Activity space. To understand the idea in intuitive terms, interest space consists of all the possible interest values for a person. Similarly, activity space consists of all the possible activity values. With these two state spaces, and as per Fig. 2, we know that interest in any entity evolves with time. For instance, John was highly interested in playing a Mobile Game, but with time the desire to play the game decreased. Hence, we need a strategy that can simulate this real world phenomenon and can *evolve* John's *interest* inside his *interest space*. To do that, we need a *transformation function*. This is represented as:

$$\varrho_k = \hat{T}_k(\varrho_{k-1}, \Phi_k) \tag{1}$$

where, ϱ_k is the interest value at the k^{th} interval of time, Φ is the i.i.d process noise. From the figure, it is also visible that the interest space transforms into the activity space. In intuitive terms, stimulated by interest John played the Game, hence, there was activity. To model this practical phenomenon (interest stimulating activity), we need the so-called *measurement function* that can provide an injective as well as a surjective mapping of John's Interest space into his Activity space. This is represented as:

$$\zeta_k = \hat{m}_k(\varrho_k, \hbar_a) \tag{2}$$

where, ζ_k is the activity at the k^{th} interval of time, ϱ_k is the interest value, $\hbar_a$ is the i.i.d measurement noise.

As we have employed Bayesian Inference, hence, the goal in such system is to find the posterior probability density function $P(\varrho|\zeta)$ of interest with the information that is available at hand. That is, we have to estimate interest from visible activity. To do that, Bayesian statistics rely on two basic principles: 1) Predict and 2) Update. For the first step, the prediction step, the system uses the following equation to make a prediction:

$$P(\varrho_t|\zeta_{t-1}, \gamma) = \int_{\varrho} P(\varrho_t|\varrho_{t-1}, \zeta_{t-1}, \gamma) P(\varrho_{t-1}|\zeta_{t-1}, \gamma) d\varrho_{t-1} \tag{3}$$

where, γ is the parameter vector. Once the system has predicted interest, we feed new information about activity and update the predicted value. In terms of Bayesian statistics, we calculate the posterior density with the newly fed data. This is done through the following equations

$$P(\varrho_t|\zeta_t, \gamma) = \frac{P(\zeta_t|\varrho_t, \gamma) P(\varrho_t|\zeta_{t-1}, \gamma)}{P(\zeta_t|\zeta_{t-1}, \gamma)} \tag{4}$$

where the denominator is represented as:

$$P(\zeta_t|\zeta_{t-1}, \gamma) = \int_{\varrho} P(\zeta_t|\varrho_t, \zeta_{1:t-1}, \gamma) P(\varrho_t|\zeta_{1:t-1}, \gamma) d\varrho_t \tag{5}$$

The expression presented in equation (5) is relatively constant with respect to ϱ and is often ignored in practise.

Using the theoretical foundations discussed in this subsection, we now have an understanding of the method, and can proceed to find a solution to the IEP. However, to apply this theory in practise and to find real time numerical estimates of interest from numerical activity, we need computationally feasible definitions for several components. They are as follows:

i) We need a method to compute activity.

ii) We need a definition for the transformation function (equation 1).

iii) We need a computationally feasible definition of the measurement function (equation 2).

iv) Lastly, we need a Bayesian filter to filter interest from activity.

In the following sections, we shall consider each of these components in turn. We start with the method to calculate activity.

3.3. Measuring Activity

Algorithm (i). Computing Subjective Weights

Initialize. Feed the pairwise comparison matrix to the system. The matrix has to satisfy the following properties: $M_{kk} = 1$, $M_{jk} > 0$, $M_{jk} = \frac{1}{M_{kj}}$

where, M_{jk} is the preference towards a particular attribute a_n w.r.t the attribute a_m.

Problem Formulation. Weights are computed after minimizing the following objective function:

$$\min X = w_T bw = \sum_{i=1}^{n}\sum_{j=1}^{n}(m_{ij}w_j - w_i)^2$$

subject to

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

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

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

Applying Non-liner Solution. The solution is obtained as:

$$SM = B^{-1}I / I^T B^{-1} I$$

where, I is the identity matrix, and $SM > 0$, $\sum SM = 1$, SM is the subjective weight matrix.

In Section 1, it was pointed out that interest is estimated via activity. However, it was also specified (point C2) that activity is rarely a unitary variable. Consequently, the issue is: How to computationally measure activity? In this section, we present a method to handle this issue.

To understand the method of measuring activity, let's first consider the issue from a general point of view. Let us consider the case where a person is interested in a social networking application, say Twitter. If the person chooses to engage with this platform, the different perspectives of activity are: The number of tweets, the number of retweets, the number of tweets read, the number of tweets responded to, the number of times a user logged in a day, the length of each login session, and so on. A straightforward implication of this use case certifies that activity is not limited to a set of congenial attributes and has multiple points of view. Moreover, the attributes discussed for the use case were constrained to a particular platform (Twitter), for other platforms or objects, the dimensions of activity are expected to be different. For instance, in the case where a person is interested in the game of Basketball (for example), the attributes linked to activity will differ to the case of Twitter. The discussion here emphasizes on the fact that if we want to measure activity, *we have to consider these granular attributes for every object of interest separately*. In this direction, lets assume that for a specific object of interest O_i, we have κ dimensions of activity. Mathematically, activity is therefore a function of κ attributes. We represent this phenomenon as:

$$\zeta^{\delta}_{O_i} = \hat{\Theta}(a_1^{\delta}, a_2^{\delta}, ...a_{\kappa}^{\delta}) \tag{6}$$

where, a_{κ}^{δ} is the κ^{th} perspective of activity, ζ^{δ} is the activity at time interval δ. For example, if a person chooses to engage with Twitter, a_{κ}^{δ} can denote the number of tweets on any day.

After discussing the perspective or attributes of activity, the next objective is to find a computationally feasible definition for the function $\hat{\Theta}$. To that end, we select a weighted approach. We combine all the different facets of activity into a single number using the well-tested weighted approach. This is one of the most favored and respected methods across disciplines [41]. Using this approach, we get the definition of the function $\hat{\Theta}$ as:

$$\hat{\Theta}(a_1, a_2, ...a_{\kappa}) = \sum_{i=1}^{\kappa} w_i a_i \tag{7}$$

where, $w_{\kappa} \in \{0, 1\}$ is the weight of the κ^{th} attribute. Further, $\sum_{i=1}^{z} w_i = 1$. Before we proceed to compute activity based on the definition of the function $\hat{\Theta}$ in equation (2), we need to address one more issue. If we want to calculate activity using a weighted approach, we need a procedure to compute the value of the weights. This is because from a statistical point of view, weights specify the importance of the preference that a user has towards a particular perspective of activity. It is clear that we are dealing with a human dependent system. Naturally, we have to bring the human in the loop. For the IEP, it is vital that we incorporate typical human preferences while computing activity. The rationale here is backed by the fact that two people (say interested in the same object) need not show an equal alignment towards a particular perspective of activity. For example, lets reconsider the case of Twitter. It is possible that one person likes to send a lot messages (tweets) through his/her account, whereas the other one likes to read the tweets of other people. In short, the alignment of people across the different dimensions of activity need not be the same. As a result, the procedure must consider the '*subjective*' nature of humans while computing weights. However, going with the subjective and often judgmental nature of humans ***alone*** is not the best strategy. Literature has pointed out that subjectivity as a sole criterion is not the most ideal of approaches [42]. One should also incorporate an element of *'objectivity'* in one's decisions. This is owing to the fact that the subjective method is often limited by insufficient and incorrect information that compromises the judiciousness needed in a computational procedure. Thus, the structure of subjectivity often collapse under such circumstances. We therefore follow the analysis presented in [43] and complement the subjective approach with an objective method to compute weights. Consequently, we employ a subjective-objective weighted approach to calculate activity. The method is taken from the discussion presented in [43] (the method is summarized in Algorithms (i) and (ii)). The procedure to calculate subjective weights is an application of the least squares to solve a series of linear algebraic equations, whereas, the objective method uses an artificial programming model to minimize the distance between the ideal and several alternate solutions. We combine the two methods via the following equation:

$$\zeta^{\delta} = \xi \times SM \times A^{\delta} + (1 - \xi) \times OM \times A^{\delta} \tag{8}$$

where A is the attribute matrix, OM is the objective weight matrix, SM is the subjective weight matrix, $\xi \in (0, 1)$ is the bias parameter (An example of Activity calculation is presented in the Results section).

Algorithm (ii). Computing Objective Weights

Initialization. Feed the system normalized decision matrix $Z = (z_{ij})$, for $i = \{1, 2..., m\}$ and $j = \{1, 2, ..., n\}$.

Change Z into weighted normalized matrix $WZ = (z_{ij})w_j$, for $i = \{1, 2..., m\}$ and $j = \{1, 2, ..., n\}$.

Define z^*, WZ^*, and Y as

$WZ^*_j = \max\{WZ_{1j}, WZ_{2j}, ..., WZ_{mj}\}$, and

$z^*_j = \max\{z_{1j}, z_{2j}, ..., z_{mj}\}$, and

$Y = \{y_1, y_2, ..., y_m\}$, where

$y_i = \sum_{k=1}^{n}(WZ^*_k - WZ_{ik})^2$, for $i = \{1, 2, ..., n\}$

Problem Formulation. Compute weights by minimizing the following objective function:

$$\min w^R F w$$

subject to

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

where, F, a diagonal matrix, is defined as

$f_{ii} = \sum_{j=1}^{m}(z^*_k - z_{jk})^2$, for $k = \{1, 2, ..., n\}$

Solution by non-linear programming. Solution to the problem is as follows:

$$OM' = F^{-1}I/I^R F^{-1}I$$

where, I is the identity matrix, and

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

3.4. A Model for Interest

In line with the points specified in Section 3.1, we not discuss the next step in the procedure to estimate interest. We discuss the procedure that can model the long term evolution of interest. That is, we provide a computationally feasible definition for the transformation function. It should be noted here that owing to a lack of literature on statistically capturing the long term dynamics of interest, we formulate the phenomenon by employing everyday and common observations. Subsequently, we take a few assumptions. We begin the discussion by presenting the observations.

- The first observation is: **Interest is a stochastic procedure**. The rationale here is backed by work in analytical psychology, where research often studies internal human properties via non-deterministic methods, e.g. [44], [45]. Furthermore, the dynamic function that takes in the every day unpredictable circumstances as its input, creates several unforeseeable instances that forces interest to become chaotic. If, however, we want to refute this notion, we can predict human behavior, in fact every internal mental state can be estimated with certainty. This, as expected, is a contradiction. Therefore, using proof by contradiction, interest is a stochastic process.
- **Interest does not increase indefinitely with time**. Similar to the previous point, and owing to several erratic and uncertain circumstances in a person's life, the human routine goes through a cycle of unpredictability. Moreover, it is commonly observed that a person does not engage with his/her object of interest with the same rigor all the time. In fact the cycle goes through several ups and downs. Hence, interest does not increase indefinitely with time.

After outlining the starting observations, we move to the assumptions of this paper. These are explained in the following points.

1. **We assume that interest is Diffusion Process**. In simple terms, we assume that interest is Markov process without jumps. This is a standard assumption of literature [46], [44], [45]. Internal mental states do not have any abrupt gaps. In fact, mental properties exhibit continuous evolution [44]. Therefore, we assume interest to be a diffusion process.
2. **We assume that interest reverts to a particular numerical value in the long run**. When a person engages with an entity (say a Social Networking Website), interest is usually high in the beginning. However, over time the desire to engage stabilizes and interest fluctuates around a particular level in the long run. For some people, it is quite possible that the level of convergence might be high, whereas, for others it could be low, but an important point of note here is, interest does stabilize around a particular level. With this dialectical viewpoint and the social observations around us, we assume that interest fluctuates around a particular value in the long run. This observation is also in line with literature in Finance where work assumes that the fluctuations in the underlying asset (caused by crowd Psychology and the noise) reverts to its mean position [46], [47].

With respect to these properties, we model interest via the Ornstein-Uhlenbeck (OU) process in Physics [12]. In contrast to the similar procedures in its class, it is the only method that is Markovian, stationary, and follow normally distributed increments. According to the fundamental theorem discussed in [48], [49], the equation for the procedure is as follows:

$$d\varrho_t = \lambda(\mu - \varrho(t))dt + \sigma dW_t \tag{9}$$

where μ is the mean, σ is the volatility, ϱ is the interest, dW is the Weiner process. The Weiner process, in its most basic form, is described by the following equation:

$$dW = \sqrt{dt}\mathcal{N}(0, 1)$$

The physical description of the process is briefly explained in the following points.

1. The OU process describes the movement of a Physical particle, e.g. a molecule, in space. The motion of the particle is non-deterministic at each interval of time.

2. Although, the process moves randomly in space, it has an inherent tendency to converge to a particular point in space. This point is denoted by the parameter μ. The property is called as *mean reversion*. Owing to this phenomenon, the OU process is also called as mean reverting stochastic procedure.

3. The extent of stochasticity is controlled by σ. It is an indicative of the amount of randomness entering into the system. σ is called as volatility. This parameter also makes the process fluctuate around μ in an uncertain manner.

4. λ corresponds to the speed of the process. It is called as convergence speed.

5. At each instant of time, the drift represented by the term $\lambda(\mu - \varrho(t))$ is force that pulls the particle in the direction of the long term mean.

To find an analytical solution for equation (9), we substitute $f(\varrho_t, t)$ as $e^{\lambda t}\varrho_t$ and integrate with respect to time. We manipulate both sides and get the final expression as:

$$\varrho_t = e^{-\lambda t}\varrho_0 + \mu(1 - e^{-\lambda t}) + \int_0^t \sigma e^{\lambda(g-t)} dW_g \tag{10}$$

Through equation (9), we can now model interest. In statistical terms, the method can model the short term fluctuations (via the term σdW_t) and the long term change (via the $\lambda(\mu - \varrho(t))dt$ component) in the interest towards any object (for e.g mobile games). Though, the properties discussed so far have found the theoretical solution of the proposed model (of interest), to computationally simulate equation (10), we need to dig more deep. In particular, we need to find a discrete model. To do so, we explored literature and found that substantial effort has been expended to numerically simulate the OU process. We therefore followed the analytical discussion presented in [43], [44] and arrived at the following discrete model:

$$\varrho_t = e^{-\lambda t}\varrho_{t-1} + \mu(1 - e^{-\lambda t}) + \sigma\sqrt{(1 - e^{-2\lambda t})/2\lambda}\epsilon_t \tag{11}$$

where, $\epsilon_t \sim N(0, 1)$.

In light of the derivation and the discussion, we now have a base framework that can model interest. However, as we intend to model interest via the OU process, therefore, we have to look at the statistical properties of the OU process and the physical (perhaps even psychological) properties of interest simultaneously. As a result, we analyzed the OU process in tandem with interest and found two shortcomings. These are explained in the following points:

1. The first issue with the basic OU process is the constant value of σ. Interest is a construct that does not remain constant throughout the life of a person. This is owing to the fact that the human routine is characterized by a variety of factors that induces unpredictability, and this, brings in the factor of uncertainty. Under such conditions, it is expected that a person's routine is not monotonous, rather, it is a mixture of unforeseen elements that are, on occasions, beyond one's control. Furthermore, interest and its course is often regulated by factors that are not so easily preordained. Such situations could very well be the result of a variety of socially stimulated dynamic factors. As a result, it is impractical to expect that the amount of unpredictability entering into the system is constant. Therefore, if we expect to model interest via the OU process, we cannot assume a constant value of σ (Recall that σ controls the amount of randomness entering into the system). Hence, we have the first limitation of the OU process.

2. The second issue is: The OU process assumes a constant value for λ, the convergence speed. Much like the previous point, it is infeasible to expect that the process will maintain a constant convergence speed for the entire cycle of its existence. In simple words, we cannot say how slow or fast will interest converge to its mean position. As we are dealing with a stochastic system, we have to work with realistic conditions. A natural consequence to this dialectical point of view is that we need principles that can be applied to such circumstances. In short, we cannot assume a constant value of λ. This is the second shortcoming of the OU process.

To overcome the first shortcoming, we take inspiration from Stochastic Volatility models in Economics and make the volatility component of the OU process stochastic. Stochastic Volatility models were first introduced to overcome the limitations of the famous Black-Scholes formula. By following the ideas discussed in the work presented in [50], we use the same principle to fix the problem. However, in contrast to [50], we apply the concept to the OU process. Moreover, we work along the guidelines presented in [47], [51] and make the volatility of equation (9) itself follow the OU process. This is represented as:

$$d\sigma(t) = \hat{\lambda}(\hat{\mu} - \sigma(t))dt + \hat{\sigma} d\hat{W}_t \tag{12}$$

To overcome the second drawback, we follow a similar procedure, and let the convergence speed of the process follow the OU process. We express this phenomenon as:

$$d\lambda(t) = \hat{\lambda}'(\hat{\mu}' - \lambda(t))dt + \hat{\sigma}' d\hat{W}'_t \tag{13}$$

Hence, via equations (12) and (13), we have fixed the two discussed limitations of the OU process. In intuitive terms, if a person in interested in Football (for example), equation (12) takes care of the everyday unpredictability in a person's routine. Similary, equation (13) controls the speed at which interest will change itself. As a result, we have a method that can functionally model interest (The importance of equations (12) and (13) is discussed in the results section, further, we will also validate the choice of the OU process to model the variation in the parameters: σ and λ). It should be noted here that for the rest of this paper, we will call equations (9), (12), and (13) as the *Stochastic Parameters based OU process for Interest*.

By following the discussion in this subsection, we saw that interest is modeled through the stochastic parameters based OU process. Furthermore, from the above equations it is also visible that any OU process is dependent upon three crucial parameters (λ, μ, σ). We therefore need a method to estimate their values. In this respect, there is a huge body of work dedicated to the study of parameter estimation for Stochastic Volatility models in Finance. In this paper, we have chosen one of the most widely applied methods. We estimate the parameters via Maximum Likelihood Estimation. Parameter estimation, however, is out of scope for this paper (The method is summarized in the supplementary material[1]). It should be noted here that we have followed the method presented in [52].

3.5. Transforming Interest into Activity: Applying Function Approximation

In this subsection, we present a statistical function to dynamically transform interest into activity. To do that, we take inspiration from Function Approximation and Adaptive Filtering.

The central focus of this section is to find an approximation of $\hat{m}$ (equation (2)) from the Hypothesis space (It is a space containing all the definitions of the measurement function set), such that $\hat{m} : \varrho \rightarrow \zeta$. In this formulation, ϱ is the set of all interest values and ζ is the activity set. In other words, we have to find the functional map $L : Z^n \rightarrow \mathcal{H}$, where Z is the product space defined as $Z = \varrho \times \zeta$. Furthermore, it is expected that the function should look at the set of data points D, defined as $\{(\varrho_1, \zeta_1), (\varrho_2, \zeta_2), ..., (\varrho_N, \zeta_N)\}$, and produce a map so that $\hat{m}_s(\varrho) \approx \zeta$. In sum, we have to find an algorithm that looks at the data set D, containing numerical interest and activity, and selects the best map, $\hat{m}_D$, as

$$\hat{m}_D = \arg\min_{\hat{m} \in \mathcal{H}} l_s(\hat{m}) \tag{14}$$

where, l_s is the empirical risk with $l_s = \frac{1}{n} \sum G(\hat{m}, z_i))$, $G(\hat{m}, z_i)$ is the loss in prediction. With respect to these basic principles and ideas, and in context of the IEP, we need a procedure that should perform *online estimation*. That is, it should estimate the output as well as change it's internal mechanisms as soon as new data is fed to the system. With a few statistical changes, this can be easily accommodated, the issue, however, is the theoretical foundation of the potential method. In simple words, we need a strong theoretical foundation. To that end, we found the work presented in [53]. The authors of this paper have reviewed several existing approaches in literature and have specified that work assumes a positive correlation between curiosity and actions. We work along this widely accepted theoretical notion and aim to apply it in practice. As a result, we use Normalized Least mean square algorithm (NLMS) [54] to find the potential function that can predict activity from interest. It is the advanced version of the tradition LMS algorithm. To apply the theory of NLMS, equation (14) is rewritten as:

$$\min l_{ER}(\varrho \in \hat{m}', Z^N) = \sum_{j=1}^{N} (\zeta_i - \Omega(\varrho_i))^2 \tag{15}$$

where, l_{ER} is the empirical risk, $\hat{m}'$ is the dual of ϱ, Ω is the weight vector of NLMS. The above equation is the objective function that we have to minimize. The weight vector is an indicative of the previous prediction errors made by the algorithm. The importance of this vector is that it stores the information about past predictions and accordingly make corrections in its internal mechanisms to find the most optimal functional map. It is understood that to predict activity from interest is a non-trivial phenomenon. For instance, if a person is interested in playing a Mobile game, then it is hard to predict how long will the person play the game in the next gaming session. The weight vector stores the information about the importance of previous prediction errors and make changes to the future predicted values based on this critical knowledge. In this context, and to obtain the estimates of the weight vector, the method relies upon instantaneous estimates. Furthermore, the estimates are computed by following the procedure of *Stochastic Gradient*. This is expressed as:

$$\nabla J_{\Omega}(n) = 2E[\varrho_n - \Omega_{n-1}^T \varrho_n] = 2E[e_n \varrho_n] \tag{16}$$

where, e_n is the error. After this step, we approximate the values of $E[e_n X_n]$ as: $E[e_n X_n] = e_n^{\alpha} X_n$. This directly results in:

$$e_n = \zeta_n - \Omega_{n-1}(\varrho_n) \tag{17}$$

and,

$$\Omega_n = \Omega_{n-1} + \eta e_n \varrho_n \tag{18}$$

where, η is the step size also called as the learning step. NLMS modifies equation (18) and introduces the notion of normalization. Therefore, we rewrite the equation as:

$$\Omega_n = \Omega_{n-1} + \frac{\eta e_n^{\alpha} \varrho_n}{\|\varrho_n\|^2} \tag{19}$$

Thus, from equations (17) and (19), we get the estimated value of activity$\hat{\zeta}$ as:

$$\hat{\zeta} = \Omega_n(\varrho) \tag{20}$$

Using these basic formulations, we get an automatic, an iterative, and a reconfigurable measurement function that can define the transformation of interest into activity. The step-by-step procedure to transform interest into activity is summarized in Algorithm (iii).

Algorithm (iii): Procedure to Transform Interest into Activity

Input: Input Data (ζ_n, ϱ_n), $n \in \{1, 2, .. t\}$
Output: Predicted Activity Matrix $\hat{\zeta}$, Weight Matrix Ω.
1. Initialization: $\Omega(0) = 0$; choose η
2. for every element in $D \in (\zeta_j, \varrho_j)$ do
3. $e_j = \zeta_j$ - $\Omega(j-1)\varrho_i$

[1] In the supplementary material, we have presented the derivation of the method

4. $\Omega_n = \Omega_{n-1} + \frac{\eta e_n^{\alpha} \varrho_n}{\|\varrho_n\|^2}$
5. Predict Activity $\hat{\zeta}_j = \eta \sum_{i=1}^{j-1} e_i \, \varrho_i$
end for

3.6. Bayesian Filter for Estimate Interest from Activity: Monte Carlo Simulations for Estimating Interest

As per the points specified in the end of Section 3.1, we need the next component to computationally estimate interest from activity (point 4). In this subsection, we discuss the method to filter interest from activity. From the discussion in subsections 3.3 and 3.4, we have the definitions of the transformation function (point 2 in section 3.1) and the measurement function (point 3 in section 3.1). Therefore, the last step in the IEP entails filtering numerical interest values from activity. To do that, it is commonly known that Bayesian Inference problems rely on Markov Chain Monte Carlo simulations. In this paper, we have chosen one of the variants of this family. We provide real time estimates of interest via particle filters.

Particle filter (PF) is an example of the Recursive Bayesian filter that is frequently encountered in handling the so-called perceptually hard problems. PF rely upon the principle of repeated random sampling to obtain numerical samples from the defined state space (i.e. the interest space). The central idea of PF is to use randomness and filter good numerical estimates of the underlying asset by approximating the most appropriate posterior distribution. To do that, PF is provided with a set of *Z Particles*. The set of particles is denoted by $p^Z = (\varpi^k, w^k)$, where ϖ^k is the k^{th} possible numerical hypothesis of interest, w^k is the weight (or the importance factor) of the k^{th} particle. To determine the most optimal numerical value for interest, the set of the particles is sampled from a known distribution (equations (9), (12), (13)). Once the particles are sampled, then based on existing information stored inside the data structures of NLMS, we predict activity. In the next step, we compare the predicted activity with the actual activity available to the system (Step 5). Based on this comparison, we determine the importance (or the weight) of each hypothesis of interest. Note, in this step the particles are susceptible to the problem of weight collapse. This is because PF relies on random sampling (of particles) and there is a possibility that some of the particles are badly sampled. To overcome this issue, we use the idea of importance sampling [40]. In this step, we normalize the weight of all the particles, and select only the best samples. One way to do it is to use cumulative distribution. Following this method we let the best candidates propagate forward to the next iteration. The advantage here is: we do let poor approximations of interest proceed to the next iteration, consequently, we do not compromise on good numerical estimates. Once the iterations are complete, we take the mean of all the particles and get a potential estimate of interest. As a result, we get an automatic procedure that can estimate interest from activity. The method of the particle filter to estimate interest is presented in Algorithm (iv).

Algorithm (iv). Monte Carlo Simulations for Estimating Interest.

Input: Activity Vector, ζ_x, $x \in \{1, 2,...,n\}$.
Output: Interest Vector, ϱ_x, $x \in \{1, 2,...,n\}$.

I. At time=0, sample a set of Z particles δ, where $\delta = \{p^1, p^2, .., p^Z\}$. Express each particle as $p^m = (\varpi^m, w^m)$. Further, the particle's initial estimate is obtained as: $\varpi_0^m \sim \frac{1}{\sigma_L \sqrt{2\pi}} e^{-(\mu_L)^2/2\sigma_L^2}$.

II. For iterations, $j = 1, 2, ..., ITC$.

III. For $i = 1, 2, ..., Z$, sample $\varpi_t^i | \varpi_{t-1}^i$ using equations (9), (12), 13.

IV. Compute activity, $\hat{\zeta}$, for each particle using existing information stored in NLMS.

V. Set $\hat{\varpi}_{0:t}^i = (\varpi_{0:t-1}^i, \hat{\varpi}_t^i)$ and generate importance factor. Importance factor w_t^i is calculated as: $w_t^i = \frac{1}{\sigma_I \sqrt{2\pi}} e^{-(\zeta_t - \hat{\zeta}_t^i)^2/2\sigma_I^2}$.

VI. Calculate total weight of particles $S = \sum_{i=1}^{N} w_t^i$.

VII. Normalize. $\varpi_t^i = S^{-1} \times w_t^i$.

VIII. Resample.

IX. Take mean of all particles $p(\varpi_t) \in Z$, and compute interest.

X. end *i*.

Go to next Iteration.

3.7. Activity Gaps

So far we have discussed the details to estimate interest from activity. However, despite the discussion, the method is unable to resolve the issue of the *activity gap*. To understand this issue, lets take an example of the case where, the person Alice, is interested in playing an outdoor sport (say Hockey), and likes to play it everyday. But, owing to several erratic circumstances, Alice is unable to play on the "*current*" day. Such type of situations are quite ordinary and can happen owing to a variety of reasons. As a result, there is no activity, and hence, we cannot estimate interest on the "*current*" day. However, we understand that interest is not zero on the day when there is no activity. This use case exemplifies the problem of activity gap.

To overcome the problem of activity gap, we go back to Bayesian Inference and fix the issue by following the principle of K-step ahead prediction density [55]. We modify the base equation of the paper (equation (6)), and use the following theoretical equation:

$$P(\varrho_{t+k}|\zeta_{t-1}, \gamma) = \int_{\varrho} P(\varrho_{t+k}|\varrho_{t-1}, \zeta_{t-1}, \gamma) \, P(\varrho_{t+k-1}|\zeta_{t-1}, \gamma) d\varrho_{t+k-1} \tag{21}$$

	P1	P2	P3	P4	P5	PN1	PN2	PN3	PN4	PN5	Activity
Day 1	5	4	3	7	347	5	0.6818181818	5	4.2857142857	0.1493221894	3.3784682265
Day 2	3	2	1	1	157	2.5	0.2272727273	0	0	0.0524921007	1.0849771134
Day 3	2	1	2	3	987	1.25	0	2.5	1.4285714286	0.4754866986	1.1292164702
Day 4	1	4	1	1	54	0	0.6818181818	0	0	0	0.0839727273
Day 5	5	23	3	2	9865	5	5	5	0.7142857143	5	4.5283142857
Day 6	4	2	1	4	568	3.75	0.2272727273	0	2.1428571429	0.2619508715	1.8871836108
Day 7	3	1	2	1	654	2.5	0	2.5	0	0.3057792274	1.4598560442

Table 1: **Activity Calculation**. P1: Attribute 1,..., P5: Attribute 5. PN1: Normalized Attribute 1 ,..., PN5: Normalized Attribute 5. **Subjective Weight Matrix**: [0.1549, 0.133, 0.2127, 0.1944, 0.3047], **Objective Weight Matrix**: [0.5941, 0.1166, 0.0916, 0.0536, 0.1441], **Bias Parameter** $\beta = 0.4$. **Activity Calculation was done via Normalized Attributes**. **For instance, Day 1 Activity**: 0.4×(0.1549*5+0.133*0.6818+0.2127*5+0.1944*4.857+0.3047*0.1493)+(1-0.4)×(0.5941*5+0.1166*0.6818+0.0916*5+0.0536*4.857+0.1441*0.1493)=3.37846

To practically implement the strategy, we let interest evolve itself according to equations (9), (12), (13) in cases of the activity gap. To understand this, lets reconsider the use case discussed in this section. On the day when Alice is not able to play the game (hockey), the system automatically evolves her interest using equations (9), (12), (13). In this case, note, we can predict her interest's value, however, we cannot update it. But, as soon as new data about activity is fed to the system, we will use equations (9), (12), (13) and Algorithm (iv) to update her predicted interest value. Hence, by following this procedure, we get a method similar to the ***continuous time model for interest***. This property is especially important as we expect any human internal state to be a continuous time function.

4. Results

To validate the viability of the proposed model, we experiment with real datasets. We experiment on datasets provided by Stackoverflow. It is one of the most popular crowdsourcing based Q&A Technical discussion forums on the Internet. Owing to its popularity among software developers, work has found that users are addicted to participate in its regular activities [56], [57]. Therefore, this platform presents an excellent opportunity to test the feasibility of the model in practice. To this end, we collected the details of 300 users for a period of one year.

4.1. Data Description

It was specified that the data was obtained from StackOverflow databases. We also specified that StackOverflow is an open-crowd based Q&A discussion forum. The data generated daily at StackOverflow is been made public and is available to the common public for the purpose of experimentation & analysis. The data is spread across 27 tables and has been distributed in 191 different attributes. The attributes (of the tables) describe the detail all the posts, users, revisions, votes, comments, and tags. The dataset is indeed comprehensive, but, for our purpose, that is, to calculate activity and estimate interest, only 5 appropriate attributes are available. This is because StackOverflow cleans the data before it is made public. It is done to ensure that the privacy of users is not compromised. Therefore, owing to this reason, we collected the following attributes 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. We specified that we collected the data for 300 users. To do that, we wrote several SQL queries and ran them on StackOverflow's live query editor. This is available at the link[2]. The online query editor gave us *live data-feeds*. It is now understood that the necessary information was spread across multiple tables, hence, we wrote multiple cross-table SQL queries and did some necessary data processing. This is done because the returned data (from the online query editor) is in raw format. Hence, to clean the raw data, multiple independent python and linux scripts were written. Once the scripts were executed, we obtained the necessary information in the required format. This information was then fed to the system to calculate activity and estimate interest.

4.2. Prototype Development

To demonstrate the feasibility of the method in actual deployment scenarios, we have developed a prototype. The prototype was developed as a Web application, and deployed on an Enterprise Service Bus (A Cloud based application integration framework). We deployed the application on MULE ESB[3]. Furthermore, the entire software setup was hosted on several Virtual Machines (VMs) inside the Computing laboratory of the Institute. The base hypervisor to host the VMs was XenServer[4]. The base configuration of the machine was IBM Tower Server, 48 GB RAM, Xeon X5 processor, with 1.8Ghz processing capability. The proposed system was programed in Java and had encoded classes for the Particle filter and the information required for MULE ESB to deploy the application. The mathematical functions were implemented from Apache Math library[5].

4.3. Experimental Setup

The first step in the experimental setup is to compute the numerical values for activity. Recall that activity depends upon several attributes (equation (6)). We specified in Section 4.1 that we collected five different attributes from StackOverflow.

[2]http://data.stackexchange.com/
[3]https://www.mulesoft.com
[4]http://xenserver.org/
[5]http://commons.apache.org/proper/commons-math/

With those collected attributes, the procedure to compute activity is explained in Procedure I.

Procedure I. Calculating Activity
Input: Subjective-Objective Weights, Attribute Matrix, Bias Parameter
Output: Numerical Activity
Steps:

1. The first step to calculate activity is to obtain the subjective-objective weight matrices. Recall that activity was calculated via equation (8), and the necessary input to the equation was the subjective-objective weight matrices and the attribute matrix. To obtain the former (subjective-objective weight matrices), we followed the procedure discussed in algorithms (i) and (ii). The numerical values of the weights obtained after the procedure is presented in the caption of Table 1.

2. Once we obtained the weight matrices, the next step is to obtain the attribute matrix. The procedure to obtain the attribute matrix is explained in the next two points.

3. Table 1 presents an example of activity calculation for one random user with all the necessary data for 7 days. The five attributes, shown under the columns titled P1:P5, are collected from Stackoverflow (The five attributes were specified in Section 4.1). For the purpose of clarity, the data from Table 1 (P1:P5) is also shown in the following matrix:

$$\begin{bmatrix} 5 & 4 & 3 & 7 & 347 \\ 3 & 2 & 1 & 1 & 157 \\ 2 & 1 & 2 & 3 & 987 \\ 1 & 4 & 1 & 1 & 54 \\ 5 & 23 & 3 & 2 & 9865 \\ 4 & 2 & 1 & 4 & 568 \\ 3 & 1 & 2 & 1 & 654 \end{bmatrix}$$

4. From this data (P1:P5), the attributes were normalized between 0-5. The normalized attributes are shown under the columns PN1:PN5 in Table 1. The matrix containing these normalized attributes is called as the *attribute matrix*. The attribute matrix for the example presented in Table 1 is shown below. Please note, the matrix is taken from columns PN1:PN5 in Table 1.

$$AM = \begin{bmatrix} 5 & 0.6818 & 5 & 4.2857 & 0.1493 & 3.3784 \\ 2.5 & 0.2272 & 0 & 0 & 0.0524 & 1.0849 \\ 1.25 & 0 & 2.5 & 1.4285 & 0.4754 & 1.1292 \\ 0 & 0.6818 & 0 & 0 & 0 & 0.0839 \\ 5 & 5 & 5 & 0.7142 & 5 & 4.5283 \\ 3.75 & 0.227 & 0 & 2.1428 & 0.2619 & 1.8871 \\ 2.5 & 0 & 2.5 & 0 & 0.3057 & 1.4598 \end{bmatrix}$$

5. At this point, we have the subjective-objective weight matrices and the attribute matrix. Therefore, using this input data, we used equation (8) to obtain the numerical values of activity. The computed values of activity is represented under the column titled Activity in Table 1. Further, an example of Day 1 is also elaborated in the caption of the table.

Following the method discussed under Procedure I, we obtain the required activity vectors for all the 300 users in the dataset. The next step in the setup is to estimate interest from activity. The method to predict interest from activity is elaborated in Procedure II. The input to Procedure II is the numerical activity vector obtained from Procedure I. The output is the interest vector and the predicted activity vector. Note, in state estimation problems, we can estimate interest as well as predict activity. Hence, the output of Procedure II is also the predicted activity vector. the theory of estimation problems was discussed in Section 3.2.

Procedure II. Estimating Interest
Input: Numerical Activity
Output: Numerical Values of Interest, Predicted Activity
Steps:

I. The data obtained by following Procedure I was fed to the system.

II. The system comprises of Java Classes written for the Particle filter. This main class was encoded with the definition of the transformation function (Algorithm (iii)) and the measurement function.

III. Once the input data was fed to the system, we used Algorithm (iv) to obtain real time numerical estimates of interest.

IV. The procedure was followed for each of the 300 users in the dataset separately, thereby obtaining a total of 300 interest vectors.

We specified that in Bayesian Inference problems (specifically in state estimation problems) that the method can not only predict the state, but it can also predict the output [40]. In simple words, we can estimate interest and predict activity simultaneously. *This is a property of state estimation problems*. Therefore, by comparing the actual activity (obtained from Procedure I) and the predicted activity (obtained from Procedure II), we evaluate the performance of the proposed work. To do that, we have used the standard error metrics: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). As we dealing with a stochastic system, therefore, to qualitatively test the method, we conduct 50 test runs and present the average values. The procedure to obtain the error values is elaborated in Procedure

III. The following equations are used in Procedure III to calculate RMSE and MAE.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} \tag{22}$$

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |e_t| \tag{23}$$

where, $e_t = \zeta_t - \hat{\zeta}_t$, $\hat{\zeta}_t$ is the predicted activity (computed from Procedure II), ζ_t is the actual activity (computed from Procedure I) and $\hat{\zeta}_t$ is the predicted activity.

Procedure III. How to calculate RMSE and MAE.

1. We followed the method discussed under the title procedure II and obtained 300 Interest vectors. Moreover, following the framework of state estimation problems, we also obtained 300 "predicted" activity vectors. Subsequently, we used the basic rules of error calculation to compute the values for RMSE and MAE. This was done by comparing the actual activity vectors (obtained from Procedure I) with the predicted activity vectors (obtained from Procedure II).
2. We followed the above step for 300 users in the dataset.
3. Following the above two steps, we obtained 300 MAE and RMSE values. Note we got one MAE and RMSE value for each user. Subsequently, we took the mean of 300 numerical values to obtain a single number for both MAE and RMSE.
4. The previous three steps were repeated 50 times. Thus, the system had 50 MAE and RMSE values.
5. After obtaining the 50 RMSE and MAE values, we took the average of the available data. We present this final number in the paper. The number, presented in the paper, represents the overall predictive capability of the proposed work.

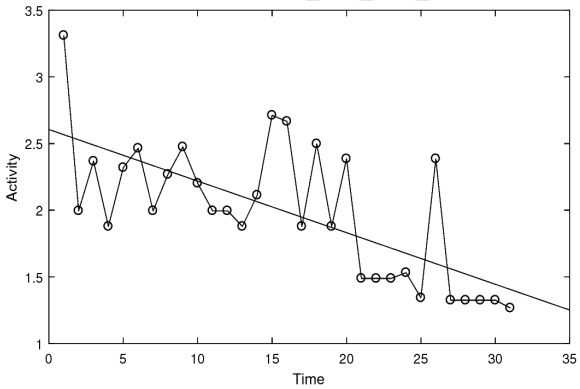

Figure 3: Activity for One Random User. Results for One Month are Presented.

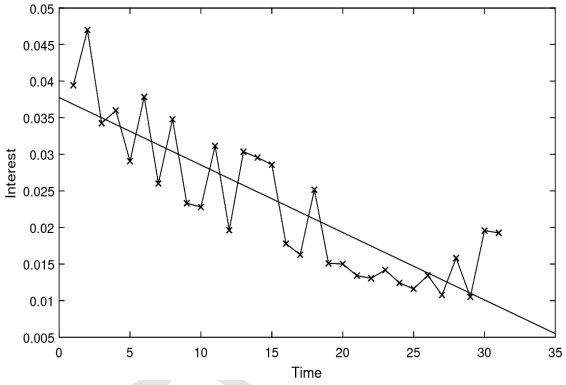

Figure 4: Interest for One Random User. Results for One Month are Presented.

4.4. Analyzing Interest and Activity

It was specified in Section 1 that interest is estimated indirectly from activity. In this subsection, we analyze interest and activity simultaneously. To that end, we have shown the calculated activity values for one random user in Fig. 3. Results for only one month are presented. The interest values estimated from activity is shown in Fig. 4. From Fig. 3 and 4, we can see that both interest and activity are going through several ups and downs. This result is expected as no user will engage (with any object) with the same intensity and rigour each day. The essence of the proposed method is that: 1) It is able to capture this phenomenon of daily import; 2) The method has made the procedure of interest quantification completely automatic. Moreover, and from this seemingly chaotic representation, we find a pattern. The trendlines for interest& activity are one and the same. That is, the trendline for the graph of interest and activity has a negative slope. The pattern discovered in this graph is intuitively as well as objectively acceptable. This is owing to the fact that we know that high interest implies high activity and vice-versa. The method proposed in this paper does not only live up to this theoretical expectation, but, it is also able to model these changing dynamics using objective and completely automatic features.

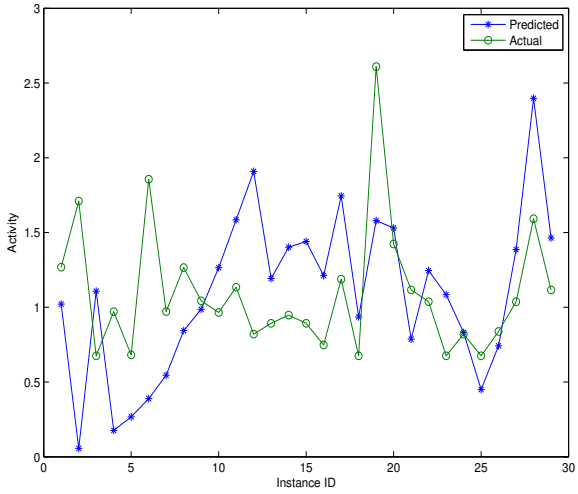

Figure 5: Results for Activity Prediction.

It should be noted here that for an exact model of interest, we expect interest and activity to exhibit *the exact same* pattern. However, as we are trying to model and computationally simulate a mental state of a human being, it is hard to get accurate and precise readings. In this regard, we specified that we test the method by comparing the predicted activity with the actual activity available to the system. Consequently, the results of activity prediction for one random user are presented graphically in Fig. 5. It is evident from the Figure that the results for activity prediction are not accurate. This nonetheless is along expected lines as we cannot precisely estimate the output. This is owing to the fact that we have a high dynamic system. The objective in such cases is to get close numerical estimates. Moreover, quantifying human mental states through machines is tough.

4.5. Recurrence Quantification Analysis of Interest and Activity

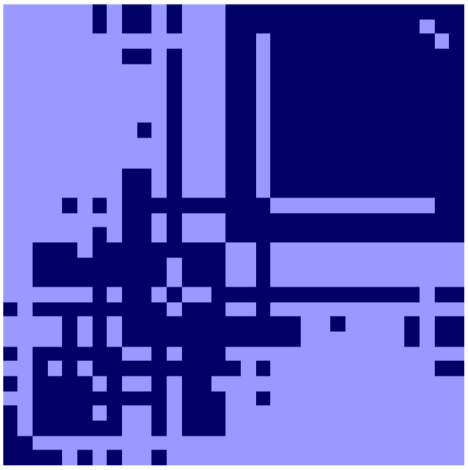

Figure 6: Binary Recurrence Plot for Interest.

In this subsection, we conduct Recurrence Quantification Analysis (RQA) of interest and activity. RQA is a non-linear data analysis technique that is used in high dynamical systems. It is often employed in Physics and Chaos Theory. In an attempt perform further investigation on interest and activity, we use this technique to extensively analyze the proposed framework. The purpose to use RQA is also to check whether a person interested in an entity goes through the same amount of interest he/she felt sometime before. For example, if a person had a certain level of interest at StackOverflow (say on 25^{th} December), the question is: Did the same level of interest repeated itself again in the next few days? That is, did the previous level of interest recurred? To answer this question, the binary plots for interest and activity are presented in Fig. 6 and 7 respectively. The graphs shown in Fig. 6 and 7 are the binary recurrence plots for the data shown in Fig. 3 and 4. The dark portions in the Fig. 6 and 7 represent values (for interest and activity) that recurred during the month. We can see from the figures that both interest and activity are recurring. In other words, some levels of interest are repeating themselves. This, in practical situations, could happen for a plethora of reasons. Moreover, this is expected as a part of our daily lives. To better explain the results, we have presented the necessary statistics of RQA in Table 2. In this table, there are several metrics. However, the most important observation is shown under the columns titled: Recurrence rate (RR) and Trapping Time (TT). The metric RR denotes the amount of time the variable under study (interest and activity) repeated it's previous state. From the results presented in Table 2, we can see that both activity and interest recurred ~47% of the time. This is a statistically feasible result. If activity is recurring, we expect a similar result for interest. The observations presented in the Table 2 as well in Fig. 6 and 7 is a practical realization of this social phenomenon. This analysis adds more value to the proposed method because we are able to realize this social observation in computational environments. In addition to RR, the second interesting fact is also shown under the column Trapping Time (TT). This metric represents the *total holding time* for a particular state. In this case, the amount of time interest and activity held a particular level. For example, and considering interest, the metric TT denotes: how many days did interest sustained itself? From the table we can see that the number for interest is ~six days, and that for activity, is ~five days. In intuitive terms, if a particular level of activity is sustaining itself, we expect the same of level of interest to sustain itself for the same amount of time. Though, the figures are not exact (six days and five days), they are close. This adds additional importance to the proposed method and makes it practically feasible.

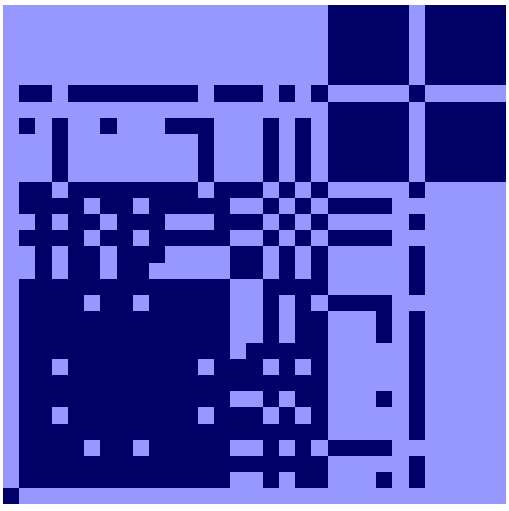

Figure 7: Binary Recurrence Plot for Activity.

4.6. Comparison with Similar Stochastic Procedures

In this subsection, we compare the performance of the proposed stochastic parameters based OU process. Though, there is a lack of literature on modeling interest via continues time models, in an attempt to highlight the superiority of the proposed framework, we compare performance with methods of

	Recurrence Rate (RR)	Determinism (DET)	Trapping time (TT)	Laminarity (LAM)
Interest	47.76	81.69	6.26	88.67
Activity	47.34	80.21	5.31	85.27

Table 2: Metrics for Recurrence Quantification Analysis for Interest and Activity.

Method Name	MAE	RMSE	Exec. Time (ms)
RW	0.6301	1.0609	26488
GBM	3.0568	3.8904	26204
Square Root Model	0.5341	0.9612	32674
Inverse Square Root Model	0.5002	0.9276	33426
Proposed Framework	**0.4037**	**0.8564**	55874

Table 3: Comparison with Random Walk and Geometric Brownian Motion. RW: Random Walk. GBM: Geometric Brownian Motion. Execution Time is in Milliseconds.

	MAE	RMSE
OU process	0.6485	1.0020
OU process with varying σ	**0.4037**	**0.8564**
OU process with varying λ	2.9376	3.8244
OU process with varying λ and σ	1.3240	1.8985

Table 4: Accuracy for different variations in the OU process. The parameters follow a mean reverting stochastic procedure.

a similar kind. Recall that we modeled interest via equations (9), (12), (13). Herein we compare the performance of the proposed method with Random walk, Geometric Brownian Motion, Square root Model and Inverse Square root Model (for details on the method refer to [58]). These methods are chosen as they are applied in a variety of practical stochastic systems, in many contexts, and in many disciplines [59], [60], [61], [58]. To compare the performance, we discard equations (9), (12), (13) and model interest using Random Walk. Next, we model interest via Geometric Brownian motion, Square root Model and Inverse Square root Model. Subsequently, with these methods, we estimate activity and calculate the error in prediction. With this setup, the results are presented in Table 3. From the evidence presented in the table, we can see that the performance of the proposed method appears superior to the methods highlighted here. Though, the methods are used extensively in literature, when we model interest using the two procedures, the performance is not up to the mark. This implies the inability of the methods to model the dynamic nature of interest closely. In contrast, the proposed method shows good performance. If we focus on the numbers, then MAE improved by 35.93%, 86.79%, 24.41%, and19.29% and the numbers for RMSE are better by 19.27%, 77.98%, 10.9%, and 7.67%.

From the results presented in the previous paragraph, we saw that the stochastic parameters based OU process showed good performance, but it should be noted here that this procedure has a high computation time. This, however, is expected, and is, acceptable. We have modeled interest using advanced interdisciplinary techniques of data analytics, therefore, it is logical to expect a high execution time. Further, it is a good choice that we compromise a little on execution time for accuracy.

4.7. Investigating the effect of varying parameters of the OU process

The next series of tests are concerned with the feasibility of the fix proposed in Section 3.3. Recall that we modeled interest using the OU process, equation (9). However, a detailed theoretical analysis revealed two shortcomings. It assumed a constant value of σ and λ. We therefore fixed the issues via equations (12) and (13). In this section, we evaluate the feasibility of the proposed fixes.

To begin with the experiment, and to evaluate the importance of the fixes, we model interest in the following four ways:

I. We model interest through the OU process with no changes to either σ or λ. In this case, note, interest is modeled via equation (9) only.

II. We fix the issue with constant σ and introduce the notion of stochastic volatility. Interest is therefore modeled via equation (9) and (12).

III. Third, we fix the issue with constant λ and introduce the idea of stochastic convergence speed. The evolution of interest is represented via equations (9) and (13).

IV. Lastly, we introduce the notion of stochastic volatility and stochastic convergence speed simultaneously. Interest is represented via equations (9), (12), and (13).

With this setup, we predict activity and present the error in estimation. The result for each of these test cases is presented in Table 4. As per this table, and if we take a look at Test Case 1, we can see that the performance of the basic OU process is acceptable. Although the numbers are a little high, they are nevertheless better than those for Random Walk and Geometric Brownian Motion presented in Table 3. In the next test case, by introducing stochastic volatility (varying σ), we see that the performance has improved. With respect to Test Case 1, the numbers have improved by 37.74% in terms of MAE and 14.53% for RMSE. Therefore, the motive to vary σ gave fruitful results. However, the objective to vary λ compromised performance. The numbers are poorer in comparison to Test Case 1. We expected the fix to the convergence speed (λ) to improve accuracy, however, the numbers presented in the table negate our initial belief. Therefore, we can say that this fix failed. Similarly, and in the fourth test case, the performance is once again compromised. We expected Test cases 3 and 4 to produce good results, but, the evidence in the table suggests otherwise. The numbers in the Table clearly points to the inference that varying the volatility (σ) component alone produces the best performance.

4.8. Why varying σ increases the accuracy

From the results presented in the previous subsection, we saw that by varying σ stochastically we improved the accuracy of the system. Therefore, we tried to find a reason for this behavior. To uncover this, we explored the data dynamics to look for an explanation. We calculated the volatility in the output observations. Volatility was computed according to the following equation:

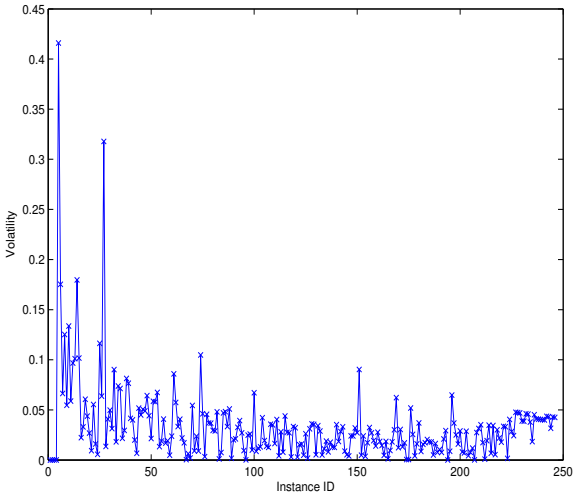

Figure 8: Volatility for One Random User.

$$Vol = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}(\zeta_j - \overline{\zeta})^2} \tag{24}$$

where Vol is the volatility, m is the number of observations, $\overline{\zeta}$ is the mean value. The results corresponding to this equation for one random user are presented in Fig. 8. As per the figure, we can see that for the entire duration, the volatility component is varying continuously with time. Furthermore, we also found that the pattern of evolution is different for different users (see supplementary material[6]). Therefore, assuming a constant value is not going to produce good results. In this paper, we not only varied the volatility component with time, but, we also used a mean reverting stochastic procedure to model the variation, thereby obtaining good results. Though, this motive increased the complexity of the system and increased the computation time (execution time increased by 19.5%), it also improved the performance.

4.9. Additional Investigation in Parameters

The efficiency brought in by introducing the notion of stochastic volatility in the OU process intrigued our curiosity to experiment further. Drawing inspiration from the analysis presented in the Sections 4.5 and 4.6, we wanted to try additional modes of variation (in the parameters). In particular, we developed the following hypothesis:

[6]In the supplementary material, we have presented multiple figures for the volatility

	MAE	RMSE
OU process	**0.6485**	**1.0020**
OU process with varying σ	0.6662	5.5814
OU process with varying λ	2.0531	14.8933
OU process with varying λ and σ	0.7556	7.9412

Table 5: Additional Mode of Variation 1. Accuracy for different variations in the OU process. The parameters follow Geometric Brownian Motion.

	MAE	RMSE
OU process	**0.6485**	**1.0020**
OU process with varying σ	0.6342	1.0434
OU process with varying λ	1.4241	2.1525
OU process with varying λ and σ	1.5981	2.3218

Table 6: Additional Mode of Variation 2. Accuracy for different variations in the OU process. The parameters follow Random Walk.

I. What is the effect of changing the mode of variation (in the parameters) on the performance?

II. Can other stochastic models outperform the proposed mode of variation in σ and λ?

Recall that we modeled the original variations in σ and λ via mean reverting stochastic procedure. Therefore, to test the two hypothesis, we will try two additional modes of variation. We model the stochastic nature of the parameters via i) Geometric Brownian Motion and ii) Random walk.

4.9.1. Experiment 1: Mode of Variation in λ and σ is Geometric Brownian Motion

In the first series of experiments, we model the stochasticity in σ and λ via Geometric Brownian Motion (GBM). To do that, we modify equations (12) and (13) and model the variation via GBM. It should be noted here that the base equation for interest, i.e. equation (9), is left intact. With this testbed, the results for the experiment are presented in Table 5. The evidence presented in this Table points to the fact that varying the parameters indeed affects the performance of the base model. In this case, however, the performance of the proposed method has decreased. This is in contrast to the result presented in Table 4 where varying σ produced the best performance.

4.9.2. Experiment 2: Mode of Variation in λ and σ is Random Walk

Similar to varying the parameters via Geometric Brownian Motion in the previous subsection, we next varied the parameters using Random walk. The results corresponding to this test are presented in Table 6. From the evidence shown in this table, we can see that unlike the numbers presented in Tables 4 and 5, the results have deteriorated, Test Cases 2, 3, and 4. Before the beginning the experiments, we expected the variation, however modeled, would yield greater accuracy. But, the analysis revealed a different picture. Though, the performance deterioration is marginal for Test Case 2 (varying σ), nevertheless, the error in prediction is high.

4.10. Lessons Learned

From the insights gained by following the discussion in Section 4.6 and 4.9, we learn a few valuable lessons. These are elaborated in the following points:

I. The evidence shown in Table 4 highlights the importance of varying the volatility component (σ) alone to improve accuracy. Though, for the last two methods of variation (Random Walk and Geometric Brownian Motion) the improvement margin is compromised, the evidence in Tables 4 shows that varying the volatility of the basic model (equation (9)) resulted in good performance.

II. To minimize the error in prediction, the choice of the modeling procedure plays an important role. *We modeled the variation in the parameters, specifically in σ, via Mean reverting stochastic procedure, hence, we obtained the best result*.

III. The combination of the parameters is also a criterion that must be given due importance. In this paper, we obtained the best performance via mean reverting stochastic *volatility* based OU process. However, it cannot be claimed that this combination as well as the method of variation is universal. We experimented with StackOverflow datasets. For other platforms, e.g. Facebook, WhatsApp etc., the combination of parameters, even the method of modeling could be different. One should not assume that the exact combination used in this paper has a broad scope and would be applicable across domains. Yes, the proposed work does provide the necessary guidelines, but, we do not claim here that this combination (of parameters) as well as the mode of the variation in universal. In lay terms, one should not forget the *No free lunch theorem* in Machine Learning [62]. That is, no solution (or method) offers a shortcut for every platform. For other practical encounters, it is recommended to explore and evaluate different models thoroughly.

5. Conclusion, Limitations, and Future Research Directions

In this paper, we presented a method to quantify and model interest. In doing so, we tried to answer the main question asked in this paper: How "*much*" are you interested in any object? To do that, we used Bayesian Inference and estimated interest via activity. First, activity was measured through a subjective objective approach. Then, interest was modeled as a Stochastic Volatility based Ornstein-Uhlenbeck process. Subsequently, interest was transformed into activity via Normalized Least Mean Square. All the individual contributions were combined and a solution was provided via particle filters. To validate the feasibility of the method, a prototype was built and experimentation was performed on real datasets. The analysis revealed superior performance of the proposed method. Further, via multiple test cases, we gained several insights and also learned a few valuable lessons .

5.1. Limitations

In this subsection, we discuss the limitations of the proposed work.

- The first important shortcoming of the method is: we cannot compute activity for every object. For example, consider the case where a person in interested in reading novels. In this case, it is tough to compute data about activity through computational methods. Though, the perspectives of activity could be: The number of pages read, the number of hours spent reading and so on. However, the big issue is: *How to capture these attributes computationally?* The current technology is not advanced enough to monitor activity towards every object in the real world. Consequently, in this use case and for similar use cases (swimming, playing hockey, thinking and so on), we cannot calculate activity. Hence, interest estimation, in those cases, is out of scope.

- Similar to the previous point, the proposed method needs data to work with. We can understand that it is imperative for any practical computational system to work on data. The absence of which leaves the system inoperable. For the IEP, if a person is interested in Facebook, he/she has to take certain actions that can be captured computationally. If interest is only in one's mind, we cannot model it through machine driven procedures. In fact, no system would be able to monitor, let alone model, a person's interest. The prevailing technology is not advanced to model and capture interest in one's mind. In short, for the proposed system to work, a person has to take certain actions towards his/her object of interest.

- Through the proposed method, we are getting a number for interest. However, the next big question is: what is the meaning of the number? At this point, human beings themselves do not completely understand the interest of other people. It is therefore hard for machines to feel or understand the human emotion of interest. At this point, a machine is an emotionless entity. Hence, it is not possible for a machine to feel the emotion of interest. This is the third drawback of the proposed work.

5.2. Future Research Directions

The paper presented a potential roadmap to approach the problem of quantifying interest through automatic machine driven procedures. However, we need additional efforts on multiple fronts. These are elaborated upon in the following points:

- In this paper, we empirically tested several procedures, and based on the findings, modeled interest via Stochastic Volatility based OU process. Though the numerical investigation showed acceptable performance, but we do not claim here that this procedure is accurate. To model the continuous and long term evolution of interest is a challenge. We used advanced data analytics and tried to tackle the issue via an Interdisciplinary approach. However, we

need additional efforts to understand and model interest in more detail. Furthermore, future efforts need not be limited to mean reverting stochastic procedures.

- We need more efforts to dynamically predict activity from interest. In this paper, we modeled the phenomenon through a self adjusting and an adaptable transfer function. It was illustrated that the function configured itself automatically in tandem with changing circumstances, but it is not claimed here that this statistical procedure is accurate. Much alike the previous point where we saw that to model the evolution of interest is a challenge, the phenomenon of converting interest into activity is similarly non-trivial. Therefore, it would not be wrong to say that we need more studies analyzing the mechanism that can convert interest into activity.

- In the above two points, we showed a few statistical research directions. However, future efforts have to be more adaptable with new insights about what is needed in novel practical and unforeseen circumstances, thereby allowing a seamless estimation of interest via machines. One such operating condition is *privacy*. The proposed method has a shortcoming in the sense that it needs access to the private data of an individual. In this regard, it is imperative that we preserve the privacy of the user. This issue especially in data analytics is not new though, and has been raised several times in literature [63]. Therefore, in addition to focusing on the statistical procedures, we need efforts to focus on methods that can preserve the privacy of an individual whilst estimating interest.

References

1. Chen, M., Zhang, Y., Li, Y., Hassan, M.M., Alamri, A.. Aiwac: affective interaction through wearable computing and cloud technology. *IEEE Wireless Communications* 2015;22(1):20–27.
2. Eagle, N., Pentland, A.S., Lazer, D.. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences* 2009;106(36):15274–15278.
3. Niculae, V., Kumar, S., Boyd-Graber, J., Danescu-Niculescu-Mizil, C.. Linguistic harbingers of betrayal: A case study on an online strategy game. *arXiv preprint arXiv:150604744* 2015;.
4. Herbart, J.. General theory of pedagogy, derived from the purpose of education. *Writings on education* 1965;2:9–155.
5. Schiefele, U.. Interest, learning, and motivation. *Educational psychologist* 1991;26(3-4):299–323.
6. Hidi, S., Renninger, K.A.. The four-phase model of interest development. *Educational psychologist* 2006;41(2):111–127.
7. Schuller, B., Rigoll, G.. Recognising interest in conversational speech-comparing bag of frames and supra-segmental features. In: *INTERSPEECH*. 2009:1999–2002.
8. Yeasin, M., Bullot, B., Sharma, R.. Recognition of facial expressions and measurement of levels of interest from video. *Multimedia, IEEE Transactions on* 2006;8(3):500–508.
9. Zhao, Z.D., Yang, Z., Zhang, Z., Zhou, T., Huang, Z.G., Lai, Y.C.. Emergence of scaling in human-interest dynamics. *Scientific reports, Nature* 2013;3.
10. Hidi, S.. Interest and its contribution as a mental resource for learning. *Review of Educational research* 1990;60(4):549–571.
11. Anderson, R.C.. Interestingness of children's reading material. *Center for the Study of Reading Technical Report; no 323* 1984;.
12. Uhlenbeck, G.E., Ornstein, L.S.. On the theory of the brownian motion. *Physical review* 1930;36(5):823.
13. Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., Konosu, H.. Being bored? recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing* 2009;27(12):1760–1774.
14. White, R.W., Bailey, P., Chen, L.. Predicting user interests from contextual information. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM; 2009:363–370.
15. Zhang, Y., Koren, J.. Efficient bayesian hierarchical user modeling for recommendation system. In: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM; 2007:47–54.
16. Ryan, R.M., Deci, E.L.. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology* 2000;25(1):54–67.
17. Hidi, S., Baird, W.. Strategies for increasing text-based interest and students' recall of expository texts. *Reading Research Quarterly* 1988;:465–483.
18. Hidi, S., Baird, W.. Interestingness—a neglected variable in discourse processing. *Cognitive Science* 1986;10(2):179–194.
19. Schraw, G., Lehman, S.. Situational interest: A review of the literature and directions for future research. *Educational psychology review* 2001;13(1):23–52.
20. Cordova, D.I., Lepper, M.R.. Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of educational psychology* 1996;88(4):715.
21. Renninger, K., Wozniak, R.H.. Effect of interest on attentional shift, recognition, and recall in young children. *Developmental Psychology* 1985;21(4):624.
22. LeDoux, J.. Cognitive-emotional interactions: Listen to the brain. *Cognitive neuroscience of emotion* 2000;:129–155.
23. Panksepp, J., Moskal, J.. Dopamine, pleasure and appetitive eagerness: An emotional systems overview of the trans-hypothalamic "reward" system in the genesis of addictive urges. *The cognitive, behavioral and affective neurosciences in psychiatric disorders* 2004;.
24. Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., et al. Combining efforts for improving automatic classification of emotional user states. *Proc IS-LTC* 2006;:240–245.
25. Ashraf, A.B., Lucey, S., Cohn, J.F., Chen, T., Ambadar, Z., Prkachin, K.M., Solomon, P.E.. The painful face–pain expression recognition using active appearance models. *Image and vision computing* 2009;27(12):1788–1796.
26. Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., et al. Combining efforts for improving automatic classification of emotional user states. *Proc IS-LTC* 2006;:240–245.
27. Hirayama, T., Dodane, J.B., Kawashima, H., Matsuyama, T.. Estimates of user interest using timing structures between proactive content-display updates and eye movements. *IEICE TRANSACTIONS on Information and Systems* 2010;93(6):1470–1478.
28. Kapoor, A., Picard, R.W., Ivanov, Y.. Probabilistic combination of multiple modalities to detect interest. In: *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*; vol. 3. IEEE; 2004:969–972.
29. Kapoor, A., Picard, R.W.. Multimodal affect recognition in learning environments. In: *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM; 2005:677–682.
30. Mota, S., Picard, R.W.. Automated posture analysis for detecting learner's interest level. In: *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*; vol. 5. IEEE; 2003:49–49.
31. Gündüz, Ş., Özsu, M.T.. Recommendation models for user accesses to web pages. In: *Artificial Neural Networks and Neural Information Processing—ICANN/ICONIP 2003*. Springer; 2003:1003–1010.
32. Sontag, D., Collins-Thompson, K., Bennett, P.N., White, R.W., Dumais, S., Billerbeck, B.. Probabilistic models for personalizing web search. In: *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM; 2012:433–442.
33. Bennett, P.N., White, R.W., Chu, W., Dumais, S.T., Bailey, P., Borisyuk, F., Cui, X.. Modeling the impact of short-and long-term behavior on search personalization. In: *Proceedings of the 35th interna-*

tional ACM SIGIR conference on Research and development in information retrieval. ACM; 2012:185–194.
34. White, R.W., Bennett, P.N., Dumais, S.T.. Predicting short-term interests using activity-based search context. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM; 2010:1009–1018.
35. Zha, Y., Zhou, T., Zhou, C.. Unfolding large-scale online collaborative human dynamics. *Proceedings of the National Academy of Sciences* 2016;113(51):14627–14632.
36. Wang, X., Zheng, X., Zhang, X., Zeng, K., Wang, F.Y.. Analysis of cyber interactive behaviors using artificial community and computational experiments. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 2017;47(6):995–1006.
37. Yan, D.C., Wei, Z.W., Han, X.P., Wang, B.H.. Empirical analysis on the human dynamics of blogging behavior on github. *Physica A: Statistical Mechanics and its Applications* 2017;465:775–781.
38. Sun, Z., Peng, Q., Lv, J., Zhong, T.. Analyzing the posting behaviors in news forums with incremental inter-event time. *Physica A: Statistical Mechanics and its Applications* 2017;479:203–212.
39. Oudeyer, P.Y., Kaplan, F.. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics* 2007;1:6.
40. Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T.. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *Signal Processing, IEEE Transactions on* 2002;50(2):174–188.
41. Wang, Y.M., Luo, Y.. Integration of correlations with standard deviations for determining attribute weights in multiple attribute decision making. *Mathematical and Computer Modelling* 2010;51(1):1–12.
42. Hastie, R., Dawes, R.M.. Rational choice in an uncertain world: The psychology of judgment and decision making. Sage; 2010.
43. Ma, J., Fan, Z.P., Huang, L.H.. A subjective and objective integrated approach to determine attribute weights. *European journal of operational research* 1999;112(2):397–404.
44. Ashby, F.G.. A stochastic version of general recognition theory. *Journal of Mathematical Psychology* 2000;44(2):310–329.
45. Siegel, D.J.. The developing mind: How relationships and the brain interact to shape who we are. Guilford Publications; 2015.
46. Vasicek, O.. An equilibrium characterization of the term structure. *Journal of financial economics* 1977;5(2):177–188.
47. Heston, S.L.. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of financial studies* 1993;6(2):327–343.
48. Doob, J.L.. The brownian movement and stochastic equations. *Annals of Mathematics* 1942;:351–369.
49. Karaoguz, C., Drix, E., Potapova, D., Huelse, M.. Curiosity driven exploration of sensory-motor mappings. In: *Deliverable for the IM-CLeVeR Spring School at the Capo Caccia Cognitive Neuromorphic Engineering Workshop*. 2011:1–7.
50. Hull, J., White, A.. The pricing of options on assets with stochastic volatilities. *The journal of finance* 1987;42(2):281–300.
51. Fouque, J.P., Papanicolaou, G., Sircar, K.R.. Mean-reverting stochastic volatility. *International Journal of theoretical and applied finance* 2000;3(01):101–142.
52. Chaiyapo, N., Phewchean, N.. An application of ornstein-uhlenbeck process to commodity pricing in thailand. *Advances in Difference Equations* 2017;2017(1):179.
53. Wu, Q., Miao, C.. Curiosity: From psychology to computation. *ACM Computing Surveys (CSUR)* 2013;46(2):18.
54. Liu, W., Principe, J.C., Haykin, S.. Kernel adaptive filtering: a comprehensive introduction; vol. 57. John Wiley & Sons; 2011.
55. Casarin, R.. Bayesian monte carlo filtering for stochastic volatility models. *Cahier du CEREMADE* 2004;(0415).
56. Bosu, A., Corley, C.S., Heaton, D., Chatterji, D., Carver, J.C., Kraft, N.A.. Building reputation in stackoverflow: an empirical investigation. In: *Proceedings of the 10th Working Conference on Mining Software Repositories*. IEEE Press; 2013:89–92.
57. Movshovitz-Attias, D., Movshovitz-Attias, Y., Steenkiste, P., Faloutsos, C.. 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; 2013:886–893.
58. Phillips, P.C., Yu, J.. Maximum likelihood and gaussian estimation of continuous time models in finance. In: *Handbook of financial time series*. Springer; 2009:497–530.
59. Noulas, A., Scellato, S., Lathia, N., Mascolo, C.. A random walk around the city: New venue recommendation in location-based social networks. In: *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE; 2012:144–153.
60. Han, B., Li, J., Srinivasan, A.. Your friends have more friends than you do: identifying influential mobile users through random-walk sampling. *Networking, IEEE/ACM Transactions on* 2014;22(5):1389–1400.
61. Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S.J., Chong, S.. On the levy-walk nature of human mobility. *IEEE/ACM transactions on networking (TON)* 2011;19(3):630–643.
62. Wolpert, D.H., Macready, W.G.. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation* 1997;1(1):67–82.
63. Fan, W., Bifet, A.. Mining big data: current status, and forecast to the future. *ACM sIGKDD Explorations Newsletter* 2013;14(2):1–5.

***Biographies (Photograph)**

Tanveer Ahmed is a Ph.D Candidate in Indian Institute of Technology Indore. He is working in the Interdisciplinary field of Psychology and Technology. He is trying to combine humans and machines for his thesis work. His area of Interest are: Computational Social Systems, Mobile Crowdsensing (with a focus on the Human Factor), Data Engineering, and Real world application of Data Analytics.

Abhishek Srivastava is an Assistant Professor in Indian Institute of Technology Indore. He has completed his Ph.D from University of Alberta, Canada. He is currently the Dean in the Indian Institute of Technology Indore. His area of Interest are: Mobile Crowdsensing, Crowdsourcing, and Service Oriented Architecture.

Highlights (for review)

1. A general method to predict interest towards any entity (e.g. Facebook, WhatsAPP, Twitter etc.) is proposed.
2. We aim to model the long-term evolution of interest. In lay terms, we want to find a number representing a person's interest
3. We model and quantify interest using data analytics.
4. Interest is modeled as a Stochastic volatility based Ornstein-Uhlenbeck process.
5. A method to dynamically transform interest into activity is proposed.
6. Validation is performed on real world datasets and a prototype is implemented.
7. A prototype is implemented and hosted on several cloud based virtual machines.