Comments by
Mário Alvim

ANÁLISE E MODELAGEM DA CURIOSIDADE DE USUÁRIOS EM SERVIÇOS DE INFORMAÇÃO ONLINE

ALEXANDRE MAGNO SOUSA

ANÁLISE E MODELAGEM DA CURIOSIDADE DE USUÁRIOS EM SERVIÇOS DE INFORMAÇÃO ONLINE

Projeto de tese apresentado ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação.

Orientador: Jussara Marques de Almeida
Coorientador: Flavio Vinicius Diniz de Figueiredo

Belo Horizonte
Novembro de 2022

ALEXANDRE MAGNO SOUSA

ANALYZING AND MODELING USER CURIOSITY IN ONLINE INFORMATION SERVICES

Dissertation project presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Advisor: Jussara Marques de Almeida
Co-Advisor: Flavio Vinicius Diniz de Figueiredo

Belo Horizonte
November 2022

Resumo

Na atual era da economia da informação, a curiosidade, componente fundamental do comportamento humano, exerce um papel de extrema importância, uma vez que diversos serviços de informação (e.g., busca, recomendação) são projetados para atrair e manter a atenção do usuário. Recentemente, muitos esforços têm surgido para introduzir modelos de curiosidade em serviços de informação online visando melhorar a personalização de seus resultados e, em última instância, a satisfação de seus usuários. Um modelo de particular interesse, fundamentado em teorias da Psicologia, é a curva de Wundt, que expressa como a curiosidade de uma pessoa varia em função do estímulo ao qual ela é exposta. O estímulo, por sua vez, é definido como uma combinação de variáveis colativas, tais como novidade, complexidade, conflito e incerteza, que capturam fatores que governam como a curiosidade humana é estimulada.

Esta tese tem por objetivo investigar a hipótese de que modelos que capturam múltiplas facetas da curiosidade humana podem ser usados para revelar perfis de comportamento humano relevantes para um entendimento mais fundamental do processo de disseminação de informação online assim como para o projeto de serviços de informação mais eficazes. Para tanto, propõe-se explorar múltiplas variáveis colativas que compõem o estímulo de curiosidade, um esforço ainda pouco explorado na literatura, em especial no contexto de serviços de informação online. A investigação é feita a partir de dois estudos de caso relacionados à disseminação de informação. O primeiro aborda o papel da curiosidade no consumo de informação, especificamente músicas na plataforma LastFM. O segundo foca na produção de informação e aborda o papel da curiosidade no compartilhamento de conteúdo em grupos do WhatsApp.

São propostas várias métricas que capturam diferentes variáveis colativas associadas à estimulação de curiosidade, e essas são usadas para identificar e caracterizar diferentes perfis de curiosidade dos usuários. Mais ainda, argumenta-se pela necessidade de se considerar a influência social como uma outra variável colativa que compõe o estímulo de curiosidade, em especial em plataformas de mídias sociais. Para tanto, são propostas novas métricas para representar esta variável, para indivíduos e grupos,

e é mostrado que essas novas métricas capturam aspectos ortogonais àqueles capturados por outras variáveis colativas tradicionalmente estudadas. Complementarmente, também são investigadas abordagens alternativas de modelagem que capturam o comportamento dinâmico e heterogêneo da curiosidade do usuário para os diferentes perfis de curiosidade encontrados. Por fim, como próximo passo, pretende-se incorporar as métricas e os modelos de curiosidade já desenvolvidos em sistemas de recomendação a fim de quantificar os possíveis benefícios.

Palavras-chave: Curiosidade, Influência social, Métricas de teoria da informação, Comportamento do Usuário, Disseminação da informação, Serviços de informação.

Abstract

In the current era of information economy, curiosity, a fundamental component of human behavior, plays a key role as various information services (e.g., search, recommendation) are designed to attract and maintain user attention. Recently, many efforts have emerged to introduce curiosity models in the design of online information services, aiming to improve result personalization and, ultimately, user satisfaction. A model of particular interest, based on theories of Psychology, is the Wundt's curve, which expresses how a person's curiosity varies depending on the stimulus to which she is exposed. Stimulus, in turn, is defined as a combination of collative variables, such as novelty, complexity, conflict and uncertainty, which capture factors governing how human curiosity is stimulated.

This dissertation aims to investigate the hypothesis that models capturing multiple facets of human curiosity can be used to uncover relevant user behavior profiles for the sake of enabling a more fundamental understanding of the online information dissemination process as well as designing more effective personalized information services. To that end, we propose to explore multiple collative variables as components of curiosity stimulation, an effort that has been little explored in the literature, especially in the online information services domain. The investigation is based on two case studies related to information dissemination. The first one addresses the role of curiosity in the consumption of information, specifically music on the LastFM platform. The second one focuses on information production and addresses the role of curiosity as a driving force behind content sharing in WhatsApp groups.

We propose a number of metrics that capture different collative variables associated with curiosity stimulation, and use those metrics to identify and characterize different profiles of user curiosity. Furthermore, we here argue for the need to consider social influence as another component of curiosity stimuli, especially on social media platforms. As such, we also propose new metrics to capture such aspect, both at the individual and group levels, and show that they capture aspects of curiosity stimulation that are not covered by other traditionally studied collative variables. Complemen-

tarily, we investigate alternative modeling approaches to more accurately capture the dynamics and heterogeneity of user curiosity for the different curiosity profiles uncovered. Finally, as a next step, we intend to incorporate the metrics and models developed in the design of a recommendation system in order to quantify their potential benefits.

Palavras-chave: Human curiosity, social influence, information theory metrics, user behavior profiles, information dissemination, information services.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Our personality traits are expressed in our social relationships and daily activities and, naturally, are registered in the digital world, mainly in social media as we interact with each other on the various existing platforms [Singer, 2014; Kosinski, 2014; Santos, 2017]. Indeed, the pervasive use of such platforms has changed the way people interact with their peers, generating massive quantities of online data records. Such expressive volume of data capturing different facets of human behavior has motivated many research efforts to analyze and model psychological aspects behind the observed behavioral patterns [Youyou et al., 2015; Singer, 2016; Santos and Sebastiá, 2016; Perez et al., 2018; Zurn and Bassett, 2018]. For example, the rising of the Big Five personality test [John and Srivastava, 1999] as a powerful tool for measuring and characterizing the human personality traits has opened new opportunities for computing researchers around the world [Santos, 2017; Tovanich et al., 2021]. By exploring models of human personality traits, such as those used as components of the Big Five test [Youyou et al., 2015; Kosinski, 2014], it has been possible to optimize various services and technologies for particular types of human personalities, ultimately increasing user satisfaction [Santos and Sebastiá, 2016; Monção et al., 2021]. Examples of such efforts are the use of personality models to improve search [Millan-Cifuentes and et al, 2014], recommender systems [Zhao and Lee, 2016; Loewenstein, 2017], matchmaking services [Kosinski, 2014], educational technologies, [Wu et al., 2014] as well as mobile devices [Thilakarathna et al., 2017] and services based on the Internet of Things (IoT) [Akyildiz and et al, 2016].

Several studies of human personality have already shown that some types of personality traits are strongly associated with human curiosity [Oudeyer et al., 2016; Santos, 2017; Kosinski, 2014; Youyou et al., 2015]. Indeed, curiosity has been recognized as a critical factor that influences human behavior in positive and negative ways at all

stages of life: from a driving force in child development to the main driver of all basic scientific and philosophical inquiries. Curiosity has also been linked to many behavior disorders and non-sanctioned behaviors, such as use of drugs and criminal activity [Hsee and Ruan, 2016; Niehoff and Oosterwijk, 2020; Loewenstein, 1994]. Similarly, curiosity has also been seen as an intrinsically motivated desire or appetite for information, or a passion for learning [Loewenstein, 1994]. As such, the modeling and analysis of human curiosity as a driving force behind online user behavior deserves further investigation. This dissertation pursues this avenue of research, tackling the challenges of modeling and analyzing user curiosity in information services.

1.1 Motivation

As argued in prior work, an individual's curiosity is essentially driven by external stimuli [Berlyne, 1960; Loewenstein, 1994] which are captured by various collative variables. These variables refer to different factors that govern how one's curiosity is stimulated [Berlyne, 1960]. Novelty, complexity, uncertainty and conflict are examples of collative variables capturing aspects related to curiosity stimulation that have been reported in the literature [Loewenstein, 1994]. According to Berlyne, novelty refers to how new the stimulus experienced is with respect to one's past experience, whereas complexity refers to the degree of diversity in a stimulus [Berlyne, 1960]. Uncertainty, in turn, arises when an individual has difficulty in deciding how to respond to a stimulus, whereas conflict occurs when the same stimulus triggers two or more incompatible responses in an individual. All these factors ultimately affect curiosity arousal. As a matter of fact, one can argue that a stimulus is as a combination of multiple collative variables and its intensity is related to the arousal of curiosity in an individual.

In that direction, the Wundt's curve has been proposed to model, in general terms, the curiosity of an individual as a function of the intensity of the stimulus the individual is subjected to [Wundt, 1874; Berlyne, 1960]. This is a bell-shaped curve illustrating that as the stimulus intensity increases from low to moderate, its effect on curiosity is pleasant and rewarding (i.e., curiosity increases). However, as it increases to higher levels, its effects becomes unpleasant and even painful (i.e., curiosity decreases). In other words, too little stimulus leads to an individual state of boredom, whereas too much stimulus leads to one's anxiety. Otherwise, there is an intermediate level of stimulus that leads to maximum curiosity. Each individual has her own Wundt's curve as the intensity thresholds defining the three zones (boredom, curiosity and anxiety) vary from person to person.

Curiosity models based on the Wundt's curve have already been applied in the design of artificial creativity systems, to control the behavior of non-player characters in digital games and in reconfigurable robots [Grace and Maher, 2015; Merrick and Maher, 2009; Gottlieb and et al, 2013; Li et al., 2020]. In the particular domain of online information services, there have been recent efforts to formally introduce curiosity models into the design and evaluation of recommendation systems [Zhao and Lee, 2016; Chen et al., 2019; Xu et al., 2021; Wang et al., 2020; Shandhilya and Srivastava, 2020].

As an example, Zhao and Lee developed a recommendation system framework which considers that each recommended item denotes a stimulus to the user. Such stimulus, in turn, was related to curiosity scores using the Wundt's curve. The authors showed that the integration of such curiosity model into different recommendation mechanisms provides personalized recommendations with improved accuracy [Zhao and Lee, 2016]. Yet, the authors have explored a single collative variable, namely novelty, as stimulus to curiosity and, as such, does not cover all components of human curiosity stimulation. Similarly, a few more recent efforts have also integrated curiosity models into service design, but exploring a single collative variable, often novelty [Wu et al., 2016; Niu and Al-Doulat, 2021; Shrestha et al., 2020; Abbas and Niu, 2019]. One notable exception is the work by Xu *et al.*, which explored two different collative variables, namely novelty and conflict [Xu et al., 2021]. However, the use of other collative variables, jointly or in isolation, remains to be investigated. Indeed, Wu *et al.* argued that computation models exploring psychological theories often consider just one (or a few) collative variable, disregarding yet how multiple collative variables affect the level of stimulation, individually or collectively. Nevertheless, the quantitative analysis of collative variables can help form a deeper understanding of the working mechanism of computational curiosity [Wu and Miao, 2013a].

In addition to focusing mostly on a single collative variable, most previous studies of curiosity stimulation in the domain of online information services have neglected an important aspect that may impact one's curiosity, namely, *social influence*. Social influence has been widely studied as a key component in various behavioral phenomena, from information dissemination to opinion adoption (e.g., [Bakshy et al., 2012; Kloumann et al., 2015]). These prior studies can be broadly grouped into two major categories: those that aimed at quantifying social influence, often with the intention to identify the most influential users on the system [Sun and Tang, 2013; Tang and Sun, 2014; Ivanov et al., 2017; Coró et al., 2021; Adamic and Adar, 2003; Zhu et al., 2020], and those that proposed models of social influence diffusion [Bin et al., 2020; Li et al., 2019; Logins and Karras, 2019; Hung et al., 2016; Schoenebeck and Tao, 2020; Min

and San Miguel, 2018; Centola and Macy, 2007; Guilbeault et al., 2018; Centola, 2019; Guilbeault and Centola, 2021]. Our present effort is closer to the former group but has a key difference in purpose: we here aim at proposing metrics to quantify social influence as a component of an individual's curiosity stimulation driving her towards behaving in certain ways, and further exploring such metrics to improve information services. To our knowledge, prior efforts to model social influence as a component of one's curiosity stimulation are still quite scarce and recent [Wu et al., 2016, 2017; Xu et al., 2021]. Similarly, some few studies have explored principles derived from social influence in the design of recommendation systems [Shokeen and Rana, 2020], but without explicitly quantifying it as part of a curiosity model.

Despite such lack of prior efforts to explicitly model social influence as a component of curiosity stimulation, the general concept of *social curiosity* has already been widely discussed. Social curiosity, a facet of curiosity, has been defined as the general interest in gaining new social information motivating exploratory behaviors [Hartung and Renner, 2013; Renner, 2006; Litman and Pezzo, 2007]. Specifically, social curiosity entails two different aspects: a general interest in obtaining new information about how others think, behave or act as well as an interest in interpersonal information that is obtained through exploratory behavior. For instance, social information is acquired through interpersonal conversations, i.e., through how one interprets daily events, hearing about each other's lives and observing each other's behaviors and expressions. As such, social influence can be seen as yet another collative variable that stimulates one's curiosity and ultimately drives one's behavior. Indeed Loewenstein [Loewenstein, 2017, 1994] asserts that an important reference point for individuals is the attainments of others, implying that an individual who adopts other people's information sets as their own informational reference points become curious to acquire all the knowledge they possess.

In sum, efforts to operationalize social influence as a component of curiosity stimulation (i.e., as another collative variable) is a research avenue that remains mostly unexplored, offering the potential to push the state-of-the-art in curiosity modeling even further. In particular, one may expect social influence to play a particularly strong role in curiosity stimulation on social media platforms, where user interactions (e.g., posts, comments, likes) drive the dynamics of human behavior in general.

1.2 Problem Statement

This dissertation aims at investigating the following hypothesis:

Hypothesis *Models capturing multiple facets of human curiosity can be used to uncover relevant user behavior profiles for the sake of enabling a more fundamental understanding of the online information dissemination process as well as designing more effective personalized information services.*

Towards addressing this hypothesis we aim at modeling and analyzing user curiosity as a driving force behind online behavior, notably information consumption and sharing. Building upon existing literature, we intend to advance the state-of-the-art by considering multiple collative variables as components of curiosity stimulation. In particular, we here argue for the importance of capturing social influence as a component of curiosity stimulation, notably in social media platforms. Moreover, we also investigate the use of the Wundt's curve as a model of user curiosity stimulation driving online information dissemination.

1.3 Objectives

We envision our investigation of the hypothesis posed in the previous section by tackling four complementary questions, as described next:

- **Research Question 1 (RQ1)**: *Are there distinct user behavior profiles in terms of curiosity stimuli, as captured by multiple collative variables?* Towards answering this question, we aim at designing metrics capturing different collative variables associated with curiosity stimulation, notably those identified by Berlyne [Berlyne, 1960], that is, novelty, uncertainty, complexity and conflict. We also intend to use these metrics to characterize and uncover user behavior profiles that can help understand the online information dissemination process.

- **Research Question 2 (RQ2)**: *How can we capture social influence as a component of human curiosity stimulation driving online information dissemination?* As mentioned, prior efforts to capture social influence as part of curiosity models have been very timid [Wu et al., 2016, 2017; Xu et al., 2021]. Yet, we speculate that such aspect may play an important role in the composition of the user curiosity driving the information dissemination process, at least for some users, particularly in the social media domain.

  Thus, towards tackling this question, we intend to cast social influence as another collative variable of curiosity stimulation, and design novel quantitative metrics to capture ~~such~~ that variable. Moreover, we intend to show that and this variable,

as represented by the proposed metrics, captures aspects of curiosity that are not covered by other traditionally studied collative variables (e.g., novelty, uncertainty, etc). We then intend to use the proposed metrics to characterize user behavior and reveal relevant patterns and profiles driving information dissemination.

- **Research Question 3 (RQ3)**: *To which extent, user curiosity driving online information dissemination can be accurately modeled by a Wundt's curve?* This question is motivated by observations that human behavior (curiosity in particular) can be very dynamic and heterogeneous [Loewenstein, 1994]. Such dynamics and heterogeneity have already been widely reported with respect to online behavior [Liu et al., 2015; Chen et al., 2015; Wang et al., 2016; Wu et al., 2019a]. As such, it is unclear whether the Wundt's curve, which has been analyzed mostly from a theoretical perspective, can indeed be a reasonably accurate representation of human curiosity stimulation for real users of online information services. Indeed, to our knowledge, no prior work has investigated the adequacy of such model for capturing human curiosity stimulation driving information dissemination.

  We intend to address this question by exploring alternative strategies to operationalize the general ideas behind the Wundt's curve (i.e., two thresholds delimiting the regions of maximum curiosity stimulation from boredom and anxiety) and assessing the fitting of the resulting models against different user profiles.

- **Research Question 4 (RQ4)**: *Can the curiosity models be explored to improve the effectiveness of online information services, specifically content recommendation?* Towards answering this question, we aim at exploring the proposed metrics and models of human curiosity as a component of an information service, notably a recommendation service, to offer personalization. We intend to assess the extent to which such modification improves the recommendation effectiveness for individual users, comparing it against recently proposed alternative strategies [Zhao and Lee, 2016; Abbas and Niu, 2019; Shrestha et al., 2020; Xu et al., 2021].

We intend to address the aforementioned research questions by exploring different case studies in the information dissemination domain. We illustrate how these four research questions complement each other, towards exploring our posed hypothesis, in Figure 1.1.

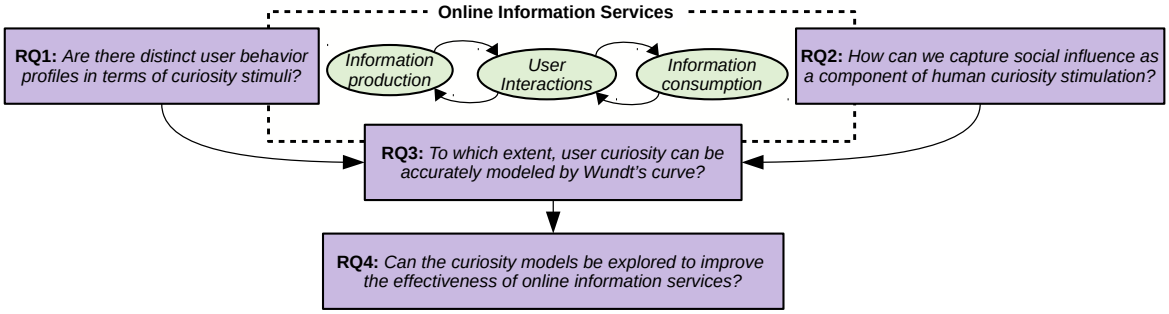


Figure 1.1: Pictorial representation of research questions of dissertation.

1.4 Contributions

The contributions achieved so far are concentrated in the first three research questions, i.e., RQ1, RQ2 and RQ3. Towards answering these questions, we explored two case studies related to information dissemination. In our first case study, focused on the LastFM platform [Schedl, 2019], we investigate the role of human curiosity in the *consumption* of online information, specifically online music. Our second case study, in turn, tackles the other end of the information dissemination process, namely information *production*. To that end, the focus is shifted to a platform where social interactions are a fundamental driver for user behavior, namely WhatsApp groups.

Specifically, our contributions so far are:

- Proposal and analysis of seven metrics capturing different collative variables associated with curiosity stimulation in the context of information consumption (RQ1). As mentioned, most prior work has focused on a single collative variable, notably novelty (and its three main components, recency, frequency and dissimilarity). We here take a step further by analyzing different variables (novelty, uncertainty, complexity and conflict) and investigating the relationships among them in terms their redundancy or complementarity in capturing aspects of curiosity stimulation. We proposed metrics to the particular scenario of online music consumption (first case study) and use them to uncover different profiles of user curiosity stimulation in LastFM.

- Proposal and analysis of novel metrics capturing social influence as a collative variable associated with human curiosity stimulation (RQ2). The metrics were designed to capture curiosity stimulation as a driver behind information production. Moreover, we focus on an communication environment, i.e., WhatsApp, where user behavior is expected to be mostly driven by their social interactions while sharing and exchanging messages (second case study). Specifically, we pro-

pose five metrics of social influence at an individual level as well as one metric capturing social influence at the (WhatsApp) group level. We then use the metrics to show that social influence is indeed complementary to to other (priorly studied) collative variables of novelty, conflict, complexity and uncertainty. In other words, our metrics are able to capture aspects of curiosity stimulation not covered by the traditionally studied variables, thus reflecting a novel and important component of the curiosity stimulation process. Moreover, we also use the metrics to offer an extensive characterization of user-level and group-level of curiosity stimulation.

- We model and analyze curiosity stimulation, as a driver behind online information dissemination, by focusing on the two ends of the process, that is, information production (and sharing) and information consumption, in two complementary case studies. In our first case study, we analyze curiosity stimulation as a driver for online information consumption, specifically music content. To that end, we use a large dataset containing records of user listenings to different musics on LastFM, one of the largest online music platforms and currently is a personal metadata repository of user's historical listenings [Odom et al., 2019; Schedl, 2019]. In our second case study, in turn, we focus on the information production and sharing end. Specifically, we delve into the role of social influence as a force that stimulate user's curiosity to share content. To that end, we focus on a currently very popular communication platform, notably WhatsApp, and look into the curiosity as a driver of user sharing content in different groups.

- Adaptations of the general Wundt's curve as models of curiosity stimulation to capture the diversity and complexity of mixed patterns observed in real users. Based on extensive analysis of real datasets capturing user listenings to different pieces of music (our first case study), we find that the curiosity stimulation curve for a large fraction of users is indeed multimodal, suggesting much more complex curiosity stimulation patterns than those expressed by the simple Wundt's curve. Nevertheless, we also find that, for most of such cases, a combination of two or three Wundt's curve is able to capture reasonably well such mixed patterns as well as the dynamics of user curiosity over time.

1.5 Outline

The rest of this dissertation is organized as follows: Chapter 2 discusses background and prior work in areas closely related to the topic of this dissertation. Chapter 3

presents our problem statement along with key assumptions and associated notations. Chapters 4 and 5 present our exploration of the three research questions, namely RQ1, RQ2 and RQ3, in two different case studies. Finally, Chapter 6 provides our final considerations, summarizes the contributions achieved so far and discusses the next steps planned to conclude this dissertation.

Chapter 2

Background and Related Work

In this chapter, we present a summary of background knowledge and related work that are essential to the understanding of this dissertation. Figure 2.1 shows a mind mapping of the the main domains to which the research developed in this dissertation relates. In light of this mapping, this Chapter is organized as follows:

- Section 2.1 provides the main concepts and theories of Psychology of curiosity, notably different dimensions of curiosity, the information gap theory and its relationship with optimal level of curiosity stimulation and social curiosity;
- Section 2.2 presents computational curiosity models, notably models used in online information systems, and discusses prior work on different applications of such models, including recommendations systems, which are more closely related to the scope of this dissertation;
- Section 2.3 reviews prior work on social influence and its important role in the information dissemination process;

2.1 Psychology of Curiosity

Curiosity is a powerful force that drives our daily occupations [Marvin and Shohamy, 2016]. In particular, human beings dedicate a lot of time in activities related to information seeking and consumption (e.g. browsing the Internet) [Kidd and Hayden, 2015]. All these activities are driven by our need for information, that is, by *curiosity*. Curiosity has been recognized as the desire for information which facilitates learning, promote new discoveries and enriches life of knowledge [Oudeyer et al., 2016]. In contrast, Hsee and Ruan highlight that the desire to resolve curiosity can lead humans to

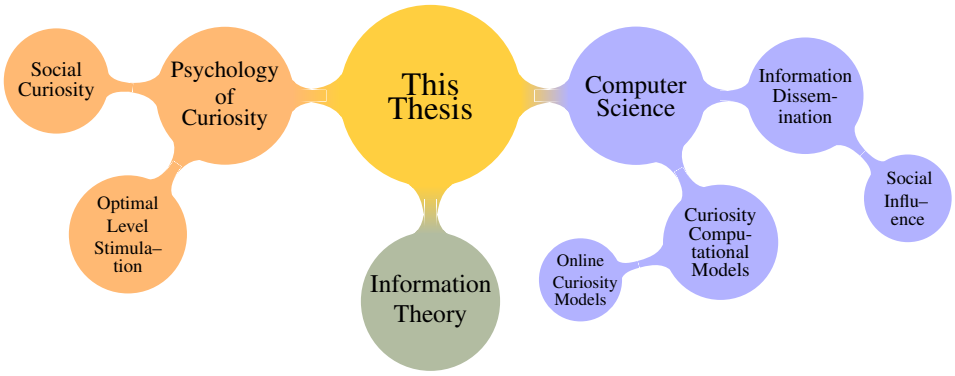


Figure 2.1: Mind mapping of the areas which compose this dissertation.

seek information despite predictably undesirable consequences [Hsee and Ruan, 2016]. This means that they do so to satisfy their curiosity without paying attention to potential negative consequences. Yet, curiosity has been seen as an intense, intrinsically motivated appetite for information, that is, an experience analogous to physiological appetites which may produce painful feelings of deprivation if not satisfied [Loewenstein, 1994].

Curiosity is not singular, it has many forms and can direct itself in a number of directions [Shankar, 2018]. Berlyne has defined two dimensions of curiosity: *perceptual* curiosity and *epistemic* curiosity [Berlyne, 1960]. The former relates to a driving force evoked by a novel stimuli and is reduced by continuous contact with these stimuli, such as, for example, the exploratory behavior in animals [Berlyne, 1960; Loewenstein, 1994]. *Epistemic* curiosity, in turn, concerns a desire for knowledge and is generally associated to humans beings. This is the type of curiosity that is arisen by complex ideas or conceptual ambiguities (e.g. scientific theories, intellectual enigmas), which motivate us to ask questions or test hypotheses in order to acquire knowledge [Collins et al., 2004]. Another commonly studied characterization of curiosity has been that distinguishing between *state*, which is curiosity in a particular situation, and *trait*, which is individual differences in capacity to experience curiosity [Loewenstein, 1994; Boyle, 1989]. Our focus in this dissertation is on the *state* of *epistemic curiosity* of human beings and its role in online information dissemination.

According to the Information Gap Theory [Loewenstein, 1994], one's curiosity arises due to a discrepancy between *what one knows*, fairly objective, and *what one wishes to know*, highly subjective. This implies that an individual should have awareness of information gap for experiencing curiosity [Arnone and Small, 1995; Golman and Loewenstein, 2016]. Thus, curiosity should be positively related to one's knowledge in a particular domain [Golman and Loewenstein, 2018]. According as Berlyne,

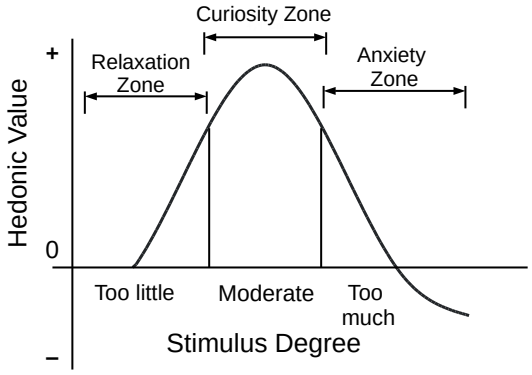


Figure 2.2: Wundt's curve and its zones of curiosity stimulation.

one's curiosity is induced by external stimuli, more specifically "stimulus conflict" or "incongruity" [Berlyne, 1960]. This involves properties such as complexity, novelty, and surprise. So, the arousal of curiosity on an individual depends on the appropriate level of stimulation that can be induced on her by a given stimulus.

Wundt introduced the concept of *optimal level of stimulation* and postulated a bell-shaped relationship between the stimulation level and the hedonic response [Wundt, 1874]. This hedonic response is directly associated with curiosity as it represents the level of pleasure one feels while satisfying her curiosity triggered by the input stimulus. The optimal level of stimulation is also called arousal potential [Berlyne, 1960]. An example of the Wundt's curve is shown in Figure 2.2, where the axis x denotes the level of stimulus received by an individual and the axis y denotes the individual's hedonic response to such stimulus. Lower hedonic values (unpleasantness) are achieved if the user receives too much (i.e. anxiety zone) or too little (i.e. boredom zone) stimulation, whereas maximum hedonic value (thus maximum curiosity) is triggered by moderate stimulus (i.e. *curiosity zone* or *arousal tonus*). The two ends of spectrum in the Wundt's curve reflect two rules in stimulus selection: *avoidance of boredom*, which is related to the boredom zone, and *avoidance of anxiety* regarding the anxiety zone [Wu and Miao, 2013a]. The intermediate region determines the *optimal level of stimulation*. Each individual has her own Wundt's curve as the thresholds defining the boredom, curiosity and anxiety zones vary across different people.

According to Berlyne [Berlyne, 1960], different *collative variables* govern the curiosity stimulation process, notably:

- *Novelty*, which is inversely related to frequency and dissimilarity in the stimulus pattern;

- *Uncertainty*, which is related to the difficulty in deciding how to respond to a

stimulus;

- *Conflict*, which occurs when the same stimulus triggers multiple incompatible responses; and

- *Complexity*, which refers to the diversity in a stimulus pattern.

Those are called collative variables because their evaluation involves an analysis of similarities and differences between existing elements in a stimulus pattern. In general, these variables refer to different external factors that govern how curiosity is stimulated. Thus, a stimulus is indeed a combination of multiple collative variables.

In his seminal work [Berlyne, 1960], Berlyne argues that properties of such collative variables can be discussed in an information-theory related language, and proposes a methodology to quantify such variables based on information theoretical metrics. Following a similar argument, Silvia [Silvia, 2006] states that, with some mathematical manipulation, all collative variables associated with curiosity stimulation can be expressed by formulations using metrics from information theory. As we discuss in the next section, these arguments inspired some recent studies [Zhao and Lee, 2016; Xu et al., 2021; Al-Doulat, 2018; Wu et al., 2016, 2017; Shrestha et al., 2020; Abbas and Niu, 2019] in Computer Science to propose metrics capturing different collative variables, notably those originally investigated by Berlyne (i.e., novelty, conflict, uncertainty and complexity).

We here argue that, in addition to the four aforementioned collative variables, *social influence* should also be considered as part of the stimulus to one's curiosity, especially in online social networks and social media applications, where, to a large extent, one's behavior is driven by connections, common interests with others and social observation.

Our argument is aligned with the concept of *social curiosity*, recently introduced in [Kashdan et al., 2018], as a key component of a five dimensional curiosity model. The proposed five dimensions, namely joyous exploration, deprivation sensitivity, stress tolerance, thrill seeking and social curiosity are related but independent and can be distinguished by links to personality, emotion and well-being. According to the authors, social curiosity denotes the individual skills to tackle the interpersonal world, that is, the arising of interest in obtaining new information about how others think, behave or act as well as an interest in interpersonal information which is obtained through exploratory behavior.

Others have also offered definitions and perspectives on the role of social curiosity on human behavior. For instance, Hartung and Renner define social curiosity as an

interest in gathering new social information motivating exploratory behaviors [Hartung and Renner, 2013], whereas Litman *et al.* argue that social curiosity includes individual's experiences and states in the social world (e.g., for social comparisons, for forming friendships and for attacking adversaries) [Litman and Pezzo, 2007]. Renner, in turn, declares that social curiosity is an essential element for building (and using) social networks and relationships, which, in turn, is a central human task [Renner, 2006]. More recently, Kashadan *et al.* distinguish between *overt* and *covert* social curiosity [Kashdan et al., 2020a]. The former refers to being curious about how others think, behave or act, and is related to healthier outcomes (e.g., interpersonal competencies). The latter refers to observing others in a indirect and secretive way to obtain new information and regards to unhealthy outcomes (e.g., gossiping and social anxiety). Despite all such prior studies, to our knowledge, efforts to *quantify* social curiosity, notably in the online world, are quite scarce.

In general, curiosity plays an increasingly large role in the economics of information, especially because modern technology is geared towards attracting, maintaining and redirecting user attention [Wojtowicz and Loewenstein, 2020]. Therefore, it is arguably one of the dominant forces driving many sources of economic growth and educational attainment. Indeed, thanks to the unprecedented availability of large amounts of data, there has been a recent surge of studies on psychological aspects of online behavior, notably curiosity and personality traits [Singer, 2014; Li et al., 2014; Lambiotte and Kosinski, 2014; Youyou et al., 2015; Singer, 2016; Santos and Sebastiá, 2016; Perez et al., 2018; Zurn and Bassett, 2018]. By characterizing one's curiosity and personality traits, it has been possible to tune and enhance services and technologies (e.g., search and recommendation services, educational and mobile technologies) to better meet one's individual personality, ultimately improving user satisfaction [Zhao and Lee, 2016; Akyildiz and et al, 2016; Loewenstein, 2017; Wu et al., 2014]. We further elaborate on such efforts in the next section.

2.2 Computational Models of Curiosity

Most existing computational models of curiosity are grounded in a general appraisal process that can be abstracted into a uniform two-step model as follows [Wu and Miao, 2013a]:

- ***Step 1***: Evaluation of the *stimulation level* based on collative variables;
- ***Step 2***: Evaluation of the *curiosity level* based on the principle of intermediate arousal potential, i.e., the optimal level of stimulation.

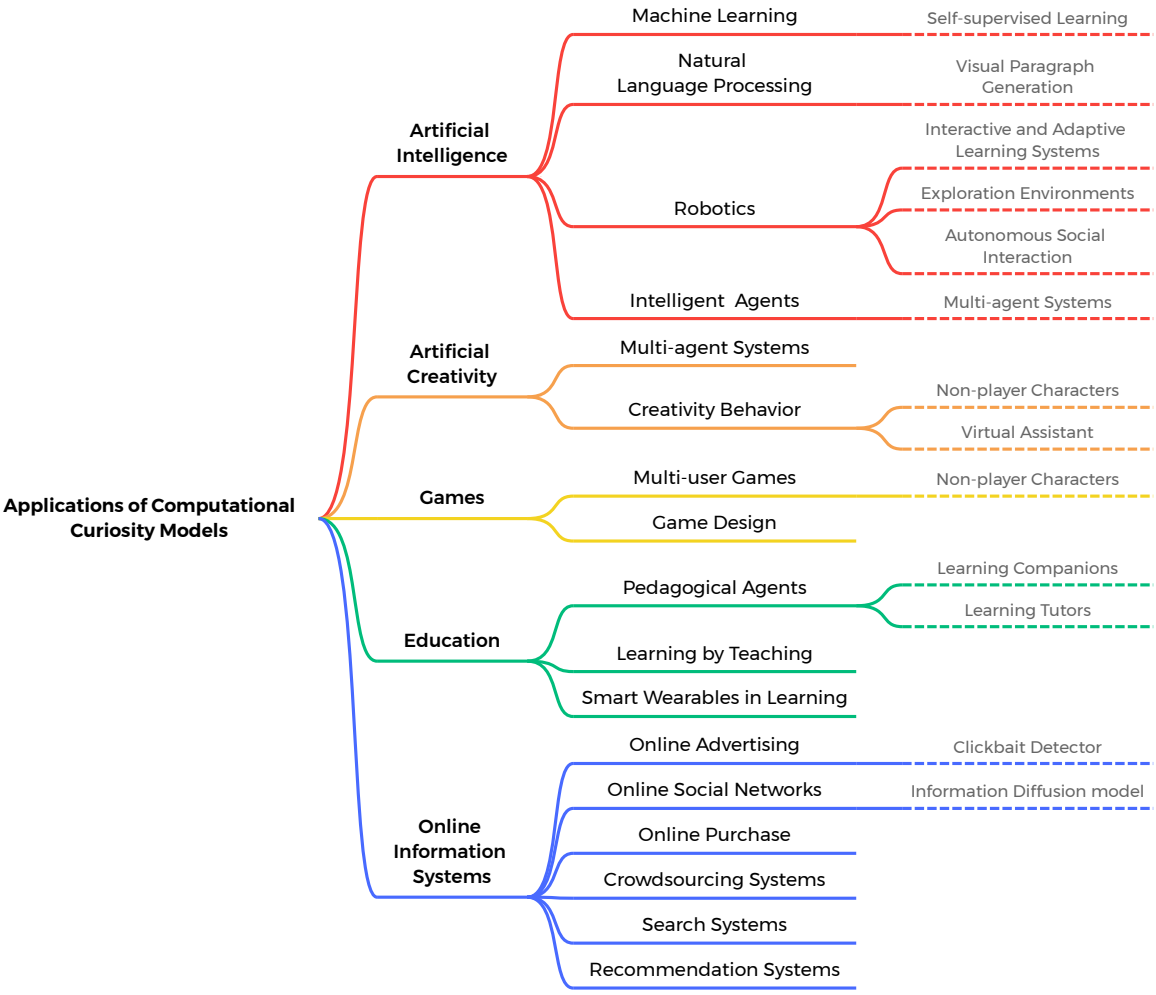


Figure 2.3: Applications of computational curiosity models.

According to the aforementioned general appraisal process, the input data is used to estimate the stimulation level, based on (one or a combination of) collative variables, as defined by the computational model. Next, the inferred stimulation level is mapped onto a curiosity level following, for instance, the *optimal level of stimulation* and the Wundt's curve discussed in Section 2.1. Finally, the curiosity level (and at times, even the stimulus level directly) are then used as component of a computational application.

In general, curiosity models have been broadly utilized in several computational areas, from artificial intelligence and education to online information services. Figure 2.3 provides an overview of various applications of computational curiosity models in several areas of Computer Science. As shown in the figure, the most important areas [Wu and Miao, 2013a] are artificial intelligence, artificial creativity, games, education and online information systems.

The most frequent use of curiosity models is, by large, in *artificial intelligence*

[Sun et al., 2022], including machine learning, natural language processing, intelligent agents and robotics. In machine learning, curiosity models have been used in the design models to tackle self-supervised learning problems such as visual categorization task [Twomey and Westermann, 2018]. In natural language processing, for example, those models have been employed in the automatic description of images using curiosity-driven reinforcement learning for visual paragraph generation [Luo et al., 2019]. Other uses of curiosity model in artificial intelligence include the design of agents that can explore environments [Macedo and Cardoso, 2005; Gottlieb and et al, 2013; Li et al., 2020], the autonomous social robots that interact with children in educational settings [Gordon et al., 2015], as well as interactive and adaptive learning system for shopkeeper robot [Doering et al., 2019], among others [Ceha et al., 2019; Bechtle et al., 2019].

In the *artificial creativity* area, it is worth mentioning the seminal work of Saunders and Gero describing an approach to developing computational models of creativity using agents that incorporate a hedonic function based on the Wundt Curve [Saunders and Gero, 2002]. Also, Maher *et al.* utilize a creativity behavior model with a combination of curiosity and learning in an agent-based system to focus attention of learning agents [Maher et al., 2008]. Grace and Maher, in turn, proposed a framework where a creative systems can intentionally exhibit transformational creativity, that is, it captures the notions of unexpectedness, surprise and specific curiosity [Grace and Maher, 2015].

In *games*, Merrick and Maher proposed the use of reinforcement learning together with psychologically inspired motivation models for creating curious non-player characters in multi-user games [Merrick and Maher, 2009]. In another approach, Schaekermann *et al.* used the identification of curiosity as player's motivation to improve game design [Schaekermann and et al, 2017].

In *education*, curiosity models based on the collative variables of novelty, surprise and uncertainty have been explored in the design of pedagogical agents (i.e., tutors or companions) to improve human learning in virtual environments [Wu et al., 2014]. Similarly, Wu *et al.* explored curiosity models to simulate learning companions that use knowledge of users and environments to identify interesting objects [Wu et al., 2012]. Also, Wu and Miao proposed a curious peer learner in a virtual learning environment to simulate the student's behavior [Wu and Miao, 2013b]. Similarly, Law *et al.* proposed a learning by teaching platform in which students work individually or in groups to teach a conversational agent how to classify objects [Law et al., 2020]. In a different perspective, Rani *et al.* designed an approach to aid a curiosity-inspired learning in every day life [Rani et al., 2021].

Finally, in the context of *online information systems*, curiosity models have been

employed in online advertising, online social networks, crowdsourcing systems, recommendations systems, search systems and online purchase. As these applications are closer to our present focus on online information dissemination, we further elaborate on prior efforts in that domain in the following sections.

2.2.1 Curiosity Models in Online Information Systems

In this section, we focus on applications of curiosity models in *online information systems* (blue branch at the bottom of Figure 2.3). As shown in the figure, those models have been employed in various subareas of interest.

For instance, in the particular context of online advertising, a seminal work is that of Menon and Soman, in which the authors studied the power of curiosity arousal in online advertising by tracking click-stream variables (e.g., records of screens visited and exposure time) as well as behavioral intention data [Menon and Soman, 2002]. More recently, Venneti and Alam proposed the use of curiosity in a clickbait detector by assessing the curiosity stimulus on readers exposed to several types of news headlines [Venneti and Alam, 2018].

Vega-Oliveros *et al.*, in turn, incorporated the notion of social curiosity into epidemiological models of information diffusion in online social networks [Vega-Oliveros and et al, 2017]. In particular, the authors considered that each user's curiosity is unique and, unlike most prior studies, dynamically evolves over time. Santos and Sebastiá, in turn, investigated the prediction of a user's curiosity based on features gathered from Facebook accounts [Santos and Sebastiá, 2016]. In a different context, Arif *et al.* analyzed the role of curiosity in user's purchase intention or buying in e-commerce websites [Arif et al., 2020]. To that end, they created a survey to evaluate the involvement, engagement and satisfaction of user and how much its behavior is influenced by curiosity.

Millan-Cifuentes *et al.* studied online search by curiosity, that is, search driven by an hedonistic need, rather than by a specific information need. To that end, the authors designed a social media search application in which the user's search is triggered by her spatial-temporal context (and not by the traditional query-response based information retrieval) [Millan-Cifuentes and et al, 2014]. The authors simulate a casual-leisure search scenario in which the user's context serves as stimulus to arise her curiosity and trigger the search.

There has also been efforts to study curiosity in crowdsourcing systems. For instance, Law *et al.* studied the potential of curiosity as a novel kind of intrinsic motivation to encourage crowd workers [Law and et al, 2016]. The authors grounded

their study on *the information gap theory* and applied curiosity-inducing designs in crowdsourcing task interfaces to improve engagement and performance.

The aforementioned studies offer a sample of recent efforts to apply curiosity models in the study of online information systems. In this dissertation, we focus on employing such models to study information consumption and production in social media applications. Additionally we also explore their use to improve user experience in recommendation systems, yet another area of application that has been explored in recent literature. Given such close connection to our present effort, we delve deeper into those efforts in the next section.

2.2.2 Curiosity Models in Recommendation Systems

Prior efforts to introduce curiosity models in recommendation systems can be generally grouped into two large sets depending on how user curiosity is inferred. One set of studies incorporates curiosity into the recommendation models by manual intervention through the gathering of user questionnaires to learn their curiosity levels. Another group of studies infer user curiosity from behavioral data (e.g., user traces) by first employing specific metrics to quantify collative variables based on the input data and then incorporating such metrics (and the inferred curiosity) directly into the recommendation process. We discuss prior studies in each category next.

2.2.2.1 Inferring User Curiosity from Questionnaire Data

Santos proposed a hybrid recommendation system that considers the curiosity level of each user as a factor in site recommendation [Santos, 2015]. User curiosity is inferred based on questionnaire data gathered by a survey test. Later, the same author and colleagues proposed CURUMIM, an online system that generates serendipitous, personalized and novel recommendations for tourist places [Santos et al., 2017]. The curiosity profile model is intrinsically based on informational data from online social network which is highly correlated with items of a survey test.

Maccatrozzo and Everdingen, in turn, proposed a model, called SIRUP, to operationalize user curiosity in content-based recommendation [Maccatrozzo and et al, 2017a,b]. Inspired by Silvia's [Silvia, 2006] and Berlyne's [Berlyne, 1960] theories on curiosity stimulation, discussed in the previous section, the authors focused on one particular collative variable, namely novelty. The authors used a metric to quantify the novelty of items in a TV program, applied a questionnaire to assess how users coped with novelty, and used the survey data to infer user curiosity profiles. They then used such profiles to create personalized recommendations.

Chen *et al.* conducted a large-scale user survey in Mobile Taobao, a popular app of e-commerce in China, to measure user perception regarding serendipity, novelty, diversity, user satisfaction and intention of purchase [Chen et al., 2019]. The authors also concluded that serendipity-oriented recommendation algorithms produce better results, as perceived by users and reported by them in the survey, compared to relevance- and novelty-oriented strategies. The same authors also analyzed the correlations between different item features and user characteristics (e.g., personality traits) the user perceptions of serendipity in recommendations, offering insights into how to improve serendipity-oriented recommender systems [Wang et al., 2020].

In a different direction, Mohseni *et al.*, in turn, proposed Pique, a personalized method to recommend sequences of scientific papers to students customized to their background and interest aiming at encouraging curiosity [Mohseni et al., 2019].

In common, all aforementioned studies relied on user surveys and questionnaires to infer user curiosity and its effect on recommendation. In this dissertation, we aim at proposing metrics to quantify different collative variables associated with curiosity stimulation from behavioral data (e.g., traces of usage of particular applications). As such, our work is more closely related to those discussed in the next section.

2.2.2.2 Inferring User Curiosity from Behavioral Data

One of the earliest efforts to infer user curiosity from behavioral data to support the design of a recommendation system is that by Zhao and Lee [Zhao and Lee, 2016]. The authors proposed the *Curiosity-based Recommendation System* (CBRS) framework to generate personalized recommendations by explicitly modeling the curiosity stimulation curve of each user, i.e., the user's Wundt's curve, from the user's access history. To do so, the authors introduced metrics to quantify the curiosity stimulated on a user by a given item based on three aspects associated with *novelty* (one of the collative variables introduced in [Berlyne, 1960]), namely *frequency*, *recency* and *dissimilarity*. These metrics were used as input to a probability distribution function mapping stimulus to curiosity score, approximated by a Beta distribution. The curiosity score, in turn, was then used as component of the recommendation function. In a similar strategy, Shrestha *et al.* also explored metrics associated with the same three aspects of novelty [Shrestha et al., 2020]. But unlike [Zhao and Lee, 2016], the authors used a Normal distribution to model the user's curiosity curve. Based on these curves, they proposed the Curiosity Inspired Personalized Recommendation Re-ranking Framework (CIR) aiming tackle overcome the tradeoff between recommendation accuracy and novelty.

Abbas and Niu also proposed a recommender system framework grounded on

the use of novelty as a source of curiosity stimulation [Abbas and Niu, 2019]. Yet, unlike [Zhao and Lee, 2016; Shrestha et al., 2020], the authors explored metrics related to frequency and surprise (a supplementary variable of novelty) to quantify the stimulus intensity. Also, they approximated the Wundt's curve of each user by a Beta distribution, parameterized according to the input data.

More recently, Xu *et al.* proposed the *Curiosity-drive Recommendation Framework* (CdRF) which also incorporates a curiosity stimulation mechanism into the recommendation process [Xu et al., 2021]. To that end, the authors explored metrics associated with two collative variables, namely novelty and conflict. In particular, the authors proposed metrics to quantify *social conflict*, defined as the closeness between positive and negative responses (i.e., ratings) on the same item received from the user's close social peers. They also modeled the Wundt's curve of each user as a predictive curiosity function which maps a stimulus intensity (estimated by a combination of the proposed metrics) into a curiosity score, which, in turn, is used as a component of the recommendation function.

In common, all the aforementioned studies proposed metrics to quantify curiosity stimuli, used them as input to explicit models of the user's curiosity stimulation curve (inspired by the Wundt's curve) and, in turn, used those models as part of the recommendation process. In contrast, some other authors opted to directly introduce metrics associated with curiosity stimulation into the recommendation function, thus avoiding the mapping from stimulus to curiosity score. In that direction, some studies focused on surprise as the main driver of curiosity stimulation [Al-Doulat, 2018; Niu and Al-Doulat, 2021]. For example, in [Al-Doulat, 2018], the authors explored alternative approaches to introduce surprise into a recommender system: one based on statistical co-occurrence likelihood and another based on user feedback. In contrast, the authors of [Niu and Al-Doulat, 2021] explored an existing computational model of surprise, aiming at improving satisfaction and inspiring curiosity of users of a health news recommender.

Surprise was also the main concept explored in [Wu et al., 2016], though in a somewhat different context. Specifically, the authors proposed a recommendation model grounded on *social curiosity* based on the idea that unexpected ratings of friends with respect to a given item triggers the user's curiosity towards the item. The proposed model combines user preferences (i.e., predicted ratings) and curiosity (as captured by the surprise with respect to items), taking into account the historical ratings of the user and of his friends. Predicted ratings and curiosity scores are then combined into an interest score used to rank the recommended items. In a follow-up study, the same authors turned their focus to another dimension of curiosity stimulation, that

is, uncertainty [Wu et al., 2017]. They modeled the feeling of uncertainty elicited on the user when his/her friends give different ratings to the same item. They then proposed a personalized recommendation strategy based on the aggregation of the rankings of items based on user preferences and on curiosity, here estimated with respect to uncertainty.

Table 2.1 offers a summary of the aforementioned efforts to introduce curiosity on recommendation systems by exploring behavioral data. For each study, the table lists the key components of the two steps of the general appraisal process of curiosity modeling – the evaluation of the stimulation level based on collative variable(s) and the estimate of the curiosity level as a function of the stimulation level. It also briefly presents some limitations of each work: in particular note that all those studies considered either a single or at most two collative variables associated with curiosity stimulation. One contribution of this dissertation is to propose metrics to capture a larger set of collative variables and assess to which extent these variables complement each other as component of the curiosity stimulation process. We do so for two different case studies related to online information consumption and production, discussed in Chapters 4 and 5, respectively.

Table 2.1: Summary of prior efforts to introduce curiosity into recommendation systems based on *behavioral data*.

Reference	**General Appraisal Process**		**Limitations**
	Collative Variable (Step 1)	**Stimulus Selection Process (Step 2)**	
[Zhao and Lee, 2016]	Novelty	Wundt's curve using *Beta distribution*	assumes data follows a Beta distribution and does not consider collative variables beyond novelty.
[Shrestha et al., 2020]	Novelty	Wundt's curve using *Normal distribution*	assumes data follows Normal distribution and does not consider collative variables beyond novelty
[Abbas and Niu, 2019]	Novelty and Surprise	Wundt's curve using *Beta distribution*	assumes data follows Beta distribution and it does not consider collative variables beyond novelty (surprise is supplementary of novelty).
[Xu et al., 2021]	Novelty and Social Conflict	Wundt's curve using *histogram function* and *sigmoid functions*	it does not consider other collative variables, namely, uncertainty and complexity.
[Al-Doulat, 2018]	Surprise	Stimulus intensity	considers only an additional aspect of novelty from collative variables (surprise) and it does not mapping the stimulus to curiosity score (Wundt's curve).
[Niu and Al-Doulat, 2021]	Surprise	Stimulus intensity	considers only an additional aspect of novelty from collative variables (surprise) and it does not mapping the stimulus to curiosity score (Wundt's curve).
[Wu et al., 2016]	Surprise	Linear combination of user preference and curiosity stimulus (novelty)	considers a single collative variable and does not model human curiosity stimulation process which is nonlinear.
[Wu et al., 2017]	Uncertainty	Combination of preference and curiosity stimulus (uncertainty)	considers a single collative variable and does not model human curiosity stimulation process which is nonlinear.

2.3 Social Influence

In the previous section, we discussed prior efforts to model curiosity in computational applications, notably recommender systems. As argued, these studies are grounded on traditionally studied collative variables associated with curiosity stimulation, most often novelty and surprise (which is related to novelty). Yet, one of the hypotheses investigated in this dissertation is that *social influence* may also be a component of (social) curiosity stimulation, notably in social media applications where user actions are often influenced by the peers. Therefore prior studies of social influence constitute an important related area of exploration, briefly reviewed in this section.

To our knowledge, prior efforts to model social influence as (part of) a stimulus to user curiosity, as is one of our present goals, are quite rare in the literature. Yet, there is a rich body of prior analyses of social influence in social media platforms from different (though related) perspectives [Sun and Tang, 2011; Li et al., 2018; Guo et al., 2019; Ribeiro et al., 2020; Samanta et al., 2021]. Those studies can be broadly grouped into two major collections: those that aimed at quantifying social influence, often aiming at identifying the most influential users on the system [Sun and Tang, 2013; Tang and Sun, 2014; Ivanov et al., 2017; Coró et al., 2021], and those that proposed models of social influence diffusion (such as epidemic models [Bin et al., 2020], cascade models [Li et al., 2019; Logins and Karras, 2019], linear threshold model [Hung et al., 2016; Schoenebeck and Tao, 2020] and complex contagious models [Min and San Miguel, 2018; Centola and Macy, 2007; Guilbeault et al., 2018; Centola, 2019; Guilbeault and Centola, 2021]). One of the contributions of this dissertation is to propose metrics to quantify social influence as a component of curiosity stimulation, a topic that is closer to the former group. Thus, we review studies in that category in this section.

Some early attempts to quantify social influence relied only on network properties such as degree distributions, diameter, clustering coefficient, community aspects and small world effect, often considering a single network model capturing explicit connections among users (e.g., friendship links) [Tan et al., 2010; David and Jon, 2010a; Kumar et al., 2016; Guo et al., 2019]. However, the limitations of considering only topological measures to estimate user influence have already been pointed out by prior studies [Zhang et al., 2018; Ver Steeg and Galstyan, 2012]. For instance, Zhang *et al.* argued that node degree (e.g, number of followers on Twitter) cannot be used alone to assess user influence as users with more connections (i.e., larger degrees) may not be the ones who more often forward content [Zhang et al., 2018]. Steeg and Galstyan [Ver Steeg and Galstyan, 2012] also argued that structural measures of influence can lead to misleading assessments of influence. For instance, higher popularity does not

imply necessarily higher influence. Moreover, topological properties are inherently dynamic [Ferreira et al., 2019], which implies that their temporal evolution, which cannot be captured by a single overall network structure, must be considered.

Also, most prior studies assume that user influence propagates over explicit links that connect users on the platform under analysis (e.g., friendship links) [Nassar et al., 2019; Negi and Chaudhury, 2016; Wu et al., 2019b]. However, in some platforms such as group communication applications (e.g., WhatsApp, Telegram), no such explicit links exist. Similarly, some prior analyses of user influence relied on explicit models such as epidemic models [Srivastava et al., 2014], topic models [Tan et al., 2010], probabilistic graphic models [Tang et al., 2009] and statistical models [Bonchi et al., 2018]. Instead, our present effort to model social influence as a component of curiosity adopts a more general *model independent* approach, grounded on information theory metrics. Indeed the metrics we propose are measured in bits, which facilitates straightforward comparisons between systems [Timme and Lapish, 2018]. Moreover, prior model-based studies of user influence often relied on linear models [Goyal et al., 2010; Tan et al., 2010]. However, online social networks are known to present *non-linear* relationships influencing information spread [David and Jon, 2010a; Matsubara et al., 2017]. In that sense, our information-theoretic metrics are more robust as they are capable of capturing linear and *non-linear* interactions [Timme and Lapish, 2018; Schreiber, 2000; Shorten et al., 2021].

Another body of studies relied on statistical or machine learning models to estimate user influence. For example, Goyal et al. [Goyal et al., 2010] proposed models of probabilistic influence among users and developed algorithms for learning the parameters of these models. Luceri *et al.*, in turn, proposed a deep neural network based model to estimate social influence by exploiting the history of user actions propagated among friends [Luceri et al., 2019]. Li and Xiong [Li and Xiong, 2017] presented measures to capture the social influence at both microscopic (based on explicit user actions such as comments and mentions) and macroscopic (based on number of followers) levels, while Zhang *et al.* [Zhang et al., 2016] used regression techniques to study the role of triadic patterns in user behavior prediction. Kumar *et al.* [Kumar and Schrater, 2017], in turn, analyzed the probability of an individual to be socially influenced based on several machine learning approaches and a rich set of features including personal network exposure, structural diversity, locality, retweet time delay as well as size and path length of cascades.

In contrast to the aforementioned studies, Bonchi *et al.* [Bonchi et al., 2018] adopted a *causal* approach to the analysis of social influence from propagation data. Their goal – retrieving a minimal causal topology from the data – is somewhat com-

plementary to ours. Instead, we here want to estimate how a user's curiosity may be stimulated by others in a group communication platform. Other studies of social influence relied on topic modeling to assess the social influence between users in specific topics [Tang et al., 2009] and factor graphs, a probabilistic graphic model to learn the effects of social influence, action correlation between users and time-dependency of user actions [Tan et al., 2010].

Finally, the work by Steeg *et al.* [Ver Steeg and Galstyan, 2012, 2013] is arguably the closest to our present effort. They also proposed to estimate social influence based on information theoretical measures, but they used a different metric – information transfer[1], which is based on mutual information. They used information transfer to measure the *direct* social influence between users on Twitter and identify influential users based on their capacity to predict the behavior of other users. Though with a similar goal, our work described in Chapter 5 differs from Steeg *et al.*'s by considering a completely different platform – WhatsApp, using somewhat different metrics, and considering both direct and indirect social influence as potential sources of curiosity stimulation.

More broadly, our work presented in Chapter 5 has some key differences in purpose from the aforementioned prior efforts, and, as such, has a complementary goal compared to them. Whereas most prior studies discussed above aimed at quantifying the influence of a user on others in general, often over a particular time window and focusing on the user who influences the others (origin of influence), we here aim at estimating how social influence from others may be driving a user towards sharing content (i.e., focus on the destination of influence). As such, social influence has to be computed at much finer granularity, i.e., each time a user shares a piece of content, since human curiosity is considered highly dynamic and contextual (see discussion in Section 3.2).

Aiming at distinguishing our work on social influence from prior studies, Table 2.2 presents a direct comparison of them along several dimensions. Some of these dimensions, such as whether the proposed solution is based on a *non-linear method* and whether it is *model independent*, refer to the level of generality of the proposed approach. For example, as mentioned, a non-linear model, as ours, should be more robust to capture non-linear effects impacting information diffusion in online social networks [Matsubara et al., 2017]. Similarly, by not exploring any specific model, as done by others [Tan et al., 2010; Srivastava et al., 2014; Tang et al., 2009; Bonchi et al.,

[1]Information transfer is also called of transfer entropy, information transfer between two stochastic process characterizes the reduction of uncertainty in a process due to knowledge of the other process [Schreiber, 2000].

2018], we offer a more general approach, where the result is not a model parameter but rather a number quantifying some relationships that exist in the data [Timme and Lapish, 2018], and free from particular assumptions and specificities that may constrain such models. Also, our study, unlike some others, does explicitly consider the *temporal dynamics* of social-influence driven curiosity, and by doing so, recognizes that this is a highly transient, dynamic and contextual human behavior trait [Loewenstein, 1994; Kashdan et al., 2020a; Lau et al., 2020].

Other dimensions listed in Table 2.2 relate to the particular domain of study – WhatsApp groups. As the first to propose metrics of curiosity stimulation to such environment, we had to address several challenges. For example, a *network free approach* was chosen due to the absence of explicit links connecting users. Instead, all group members may interact with each other at any time, unlike in other setups where there are explicit links connecting users through which social influence may propagate. This implies that, in the absence of such explicit links, a strategy to capture the heterogeneity and the dynamics of social influence in our metrics is required. Moreover, by developing a network free approach, we offer a more general solution that does not rely on existing links to estimate social influence and thus may be adapted to other setups.

Similarly, the target environment motivates the development of metrics to capture other effects of social influence, in addition to the often studied direct influence of one user on others. On one hand, we propose a metric to estimate *social curiosity at the group level*, which can be used to characterize collective behavior in different groups. Since WhatsApp allows only for small groups (at most 256 members simultaneously), and these often target specific topics of discussion (as defined in the group name or description) [Caetano et al., 2019; Resende et al., 2019b], characterizing the curiosity of such small user populations over time may offer useful insights into social behavior (much more than in more open and unconstrained spaces like Twitter and Facebook). Moreover, inspired by prior analyses of WhatsApp [Caetano et al., 2021; Nobre et al., 2022], we also propose metrics to explicitly capture the *indirect effect* of strong influencers that may emerge in such constrained spaces.

In sum, compared to the existing literature, we here propose to quantify social influence from a novel perspective – curiosity stimulation – offering a more general approach that is inspired by arguments available in the Psychology literature [Berlyne, 1960; Silvia, 2006], and that addresses particular challenges of a currently very popular communication platform, namely WhatsApp groups.

Table 2.2: Overview of prior work and our present effort to estimate social influence.

Features	Goyal *et al.* [Goyal et al., 2010]	Luceri *et al.* [Luceri et al., 2019]	Li & Xiong [Li and Xiong, 2017]	Zhang *et al.* [Zhang et al., 2016]	Bonchi *et al.* [Bonchi et al., 2018]	Tang *et al.* [Tang et al., 2009]	Tan *et al.* [Tan et al., 2010]	Kumar *et al.* [Kumar et al., 2016]	Steeg *et al.* [Ver Steeg and Galstyan, 2012, 2013]	**This work**
Social network free									✓	✓
Model independent									✓	✓
Non-linear		✓							✓	✓
Temporal dynamics	✓	✓			✓		✓	✓	✓	✓
Indirect influence										✓
Group activity										✓
Platforms of study	Flickr	Plancast, Foursquare	Tencent Weibo	Weibo, CrossFire	Flixster, Twitter, LastFM	Arnetminer, Wikipedia Films	Twitter, Flickr, Arnetminer	Sina Weibo	Twitter	WhatsApp

2.4 Summary

In this chapter, we have discussed prior studies on curiosity modeling and its applications in different areas of Computer Science, notably recommender systems. As discussed in the previous chapters, the existing literature has some important gaps and limitations, some of which we aim at tackling in this dissertation.

Recall that, as argued in the beginning of Section 2.2, the general curiosity appraisal process consists of two key steps, notably, the evaluation of the *stimulation level* based on collative variables and the evaluation of the *curiosity level* based, for instance, on the optimal level of stimulation captured by the Wundt's curve model.

With respect to the first step, the vast majority of computational models exploring psychological theories of curiosity considers only one (or a few) collative variables, thus ignoring various factors that may affect the level of curiosity stimulation on a person as well as the interplay between them. Novelty is by far the most investigated collative variable [Zhao and Lee, 2016; Shrestha et al., 2020; Xu et al., 2021], often captured by metrics related to frequency, recency, dissimilarity and surprise. The latter is an auxiliary variable and, according to Berlyne, does not capture key aspects of novelty [Berlyne, 1960]. A few other studies focused on different collative variables, notably uncertainty [Wu et al., 2017] and (social) conflict [Xu et al., 2021]. In contrast, we here propose metrics to capture all four collative variables proposed by [Berlyne, 1960], namely, novelty, conflict, uncertainty and complexity, to study online information consumption and production in different case studies.

Moreover, despite not covered by the seminal Berlyne's work, recent studies have argued for the role of social curiosity as as a component of curiosity [Renner, 2006; Litman and Pezzo, 2007; Hartung and Renner, 2013; Kashdan et al., 2018, 2020a]. Yet, despite the rich literature on social influence analyses and modeling, briefly reviewed in Section 2.3, prior efforts to operationalize its concept as (part of) a collative variable of curiosity stimulation are quite scarce [Xu et al., 2021; Wu et al., 2016, 2017], often relying on the explicit social network of users (e.g., friendship relations) and indirectly derived from user ratings. In contrast, we here propose metrics, grounded on information theory, to estimate the importance of social influence as a stimulus to user curiosity driving information sharing. Our metrics are derived for a domain, WhatsApp groups, where no explicit user network exists, and capture both direct and indirect influences that may exist among users on a group.

Concerning the second step of the curiosity appraisal process, namely, the evaluation of the curiosity level based on a given stimulus, prior studies proposed to represent the Wundt's curve of each individual by a bell-shape model, be it a Beta [Zhao and

Lee, 2016] or a Normal [Shrestha et al., 2020] distribution, estimating distribution parameters from user behavioral data. Others have modeled the curiosity level as a predictive task, relying on sigmoid functions to model the areas of boredom, anxiety and maximum curiosity. Yet other studies have opted to skip this step and rather apply the curiosity stimuli, as captured by the collative variables, directly as curiosity scores [Wu et al., 2016, 2017; Al-Doulat, 2018; Niu and Al-Doulat, 2021]. By doing so, these studies neglect the possible effects associated with the non-linear relationship between stimulus intensity and curiosity [Zhao and Lee, 2016]. Also, in general terms, the assumption of a single model of user curiosity may be quite restrictive and unrealistic, given that user curiosity, notably in online applications, can be highly heterogeneous, dynamic and in constant change [Loewenstein, 1994; Litman and Pezzo, 2007], even considering the same human being. In other words, a person's curiosity may be stimulated quite differently in different situations and may also exhibit long-term changes, reflecting individual evolution. Thus, a broader investigation of the extent to which the Wundt's curve is an accurate model of human curiosity, specially for online users, is in great need. In this dissertation, we take a step in this direction by investigating the benefits of considering a combination of distributions to capture the heterogeneity of user curiosity stimulation.

Finally, though the main focus on designing recommendation systems has been on suggesting *relevant* items to the user, a number of recent studies have explored user curiosity scores to better guide the recommendation function towards effective personalized suggestions [Zhao and Lee, 2016; Abbas and Niu, 2019; Shrestha et al., 2020; Xu et al., 2021]. Yet, given the limitations of the curiosity computational models adopted by such studies, argued above, these curiosity-driven recommender systems may still have great potential of improvements. In this dissertation, we intend to investigate to which extent the use of multiple collative variables as part of a curiosity stimulation model can offer more effective (relevant and yet personalized to the user's current need) recommendations.

To that end, we start by presenting our problem statement in the next chapter. Specifically, we introduce the two scenarios covered by our case studies, which relate to complementary perspectives of the information dissemination process: information consumption and information production. We also introduce the main assumptions that guided our work and formally define the problem we tackle.

What about curiosity wearing off?

Chapter 3

Problem Statement

As stated in Chapter 1, this dissertation investigates the hypothesis that multi-faceted models of human curiosity, capturing multiple components of curiosity stimulation, can uncover relevant user behavior profiles for the sake of better understanding online information dissemination as well as designing more effective information services (notably personalized recommendation systems). We investigate this hypothesis in light of two complementary perspectives of online information dissemination, notably information consumption and information production. We are particularly interested in studying curiosity as a driving force behind information spread in social media applications, where users play central roles not only in consuming but also in producing and reproducing (i.e., sharing) information.

As such, Figure 3.1 offers an an overview of our target problem. Historical traces of user interactions with an online information service (i.e., behavioral data) are used as key input to enable the inference of a user's curiosity at different moments in time. Specifically, we assume that curiosity is a main driving force behind each user interaction with the system, be it the consumption of a particular item (e.g., listening to a song, reading an article) or the production of one (e.g., posting a message). As such, for each new interaction of a user at a given point in time, historical data on previous interactions of the user (e.g., items previously consumed/produced) are used to compute different metrics associated with curiosity stimulation (different collative variables). These metrics can then be used to infer the curiosity stimulus on the user at the interaction time, as well as her expected response to such stimulus (i.e., curiosity level). Such inferences, in turn, can be used to build user curiosity profiles, which can be explored in the design of more effective personalized services.

Driven by the overall scenario depicted in Figure 3.1, we propose to analyze human curiosity in two case studies covering the aforementioned two perspectives of

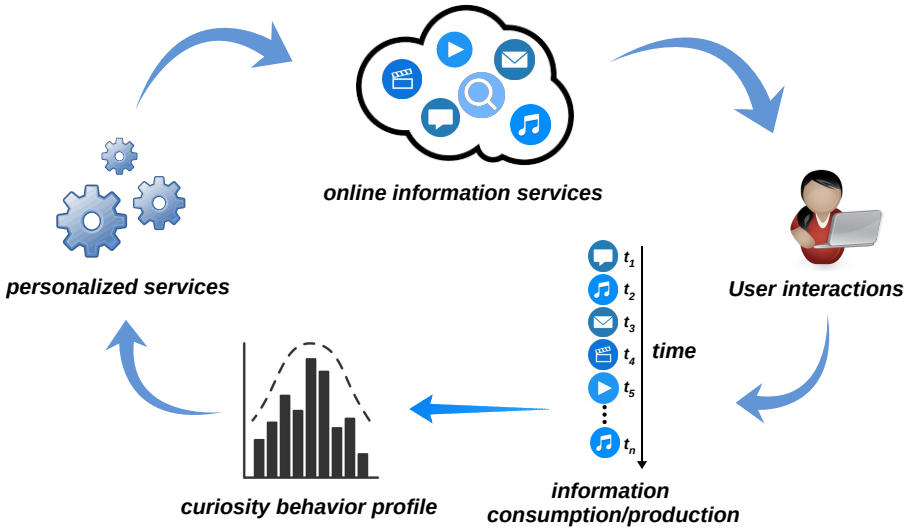


Figure 3.1: Overview of the problem statement.

information dissemination. Moreover, based on the general curiosity appraisal process, described in [Wu and Miao, 2013a], we propose novel metrics capturing different collative variables associated with curiosity stimulation in each such scenario. We also investigate the use of the Wundt's curve as single approach to evaluate human curiosity as a function of the input stimulus.

Before introducing our contributions in the two case studies, which are presented in Chapters 4 and 5, we first discuss some general aspects of the problem we intend to tackle in this chapter. We start by introducing in Section 3.1 the scenarios of evaluation corresponding to the two case studies. Next, in Section 3.2, we present some fundamental assumptions on which our user curiosity modeling approach is based. We formally introduce the problem we aim at addressing as well as some general notation in Section 3.3, and briefly summarize this chapter in Section 3.4.

3.1 Case Studies

This dissertation explores the notion that curiosity may propel users to participate in the information dissemination process by consuming previously created content as well as by producing and sharing (new and existing) content. In that case, it seems reasonable that the role of curiosity as such a driving force may vary across platforms, as different application features (e.g., explicit social links, group communication, etc) and content properties may stimulate one's curiosity quite differently [Litman and Pezzo, 2007; Schneider et al., 2013; Kashdan et al., 2020b].

In general terms, the fundamental elements of the information dissemination

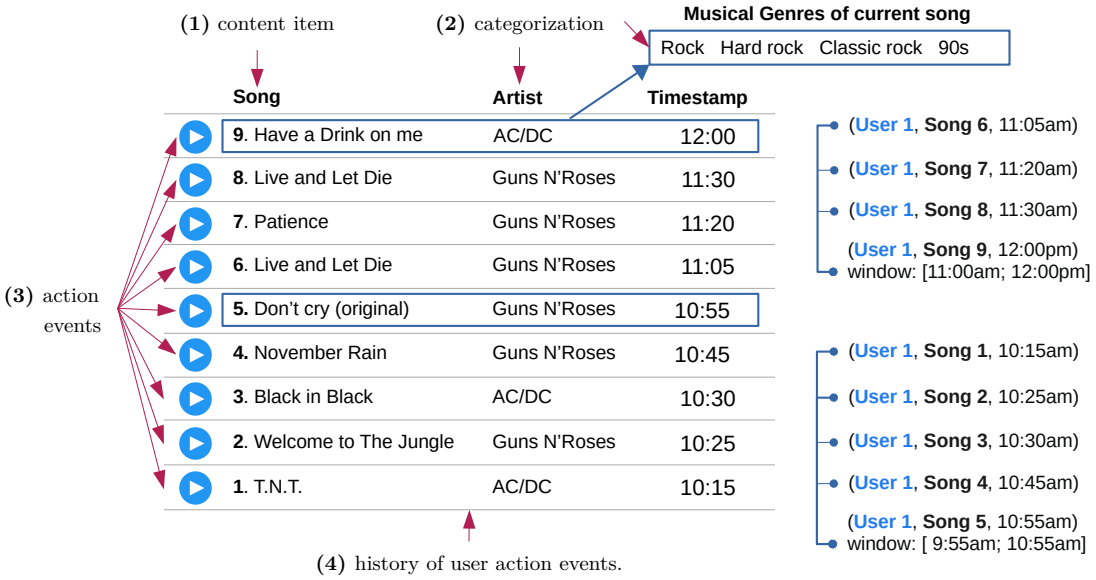
process that we must be concerned with are:

- *Content items* of information that are produced, consumed or shared. Special attention should be paid to particular content properties that may stimulate one's curiosity differently such as the content's topics or any form of content categorization (e.g., musical or movie genre, media type);
- *Users* who interact with the information system by consuming or producing (sharing) content items. Users may act mostly independently or may be influenced by others through implicit or explicit social links;
- *User actions* are the different types of actions a user may have while contributing to information spread (e.g., listening to an audio, watching a video, reading/writing a textual message, sharing an image, etc);
- *Action events* represent specific user actions related to an item's production or consumption. An action event is characterized by a given user producing/consuming a particular content item at a given moment in time.
- *History of user action events*, recorded by the system as logs of user interactions over a given period of time.

Given these elements and the great diversity of possible scenarios, we here choose to focus on two specific case studies, covering platforms with different features, which foster different types of user interactions. Fundamentally the two selected case studies cover different and complementary perspectives of the information dissemination process, namely, information consumption and information production/sharing.

In our first case study, presented in Chapter 4, we investigate the role of user curiosity as a driving force behind users listening online music in one of the largest online music service, namely LastFM. In this case, the focus is on information consumption: items are songs characterized by given artists and categories (e.g., different musical genres). The user actions of interest are listening events and user curiosity is assumed to be stimulated primarily by content and individual characteristics, captured in traces of user listenings on the platform. Figure 3.2 illustrates the main elements of this case study.

In our second case study, presented in Chapter 5, we turn our attention to information production and sharing. In that case, we consider a quite different platform, namely WhatsApp (publicly accessible) groups. In such context, the user action of interest is the sharing of content, which can be in different media types (e.g., textual



(a) Songs listened by User 1 in online music consumption. (b) Windows of interaction for songs 9 and 5 listened by User 1 ($\delta_T = 1$ hour).

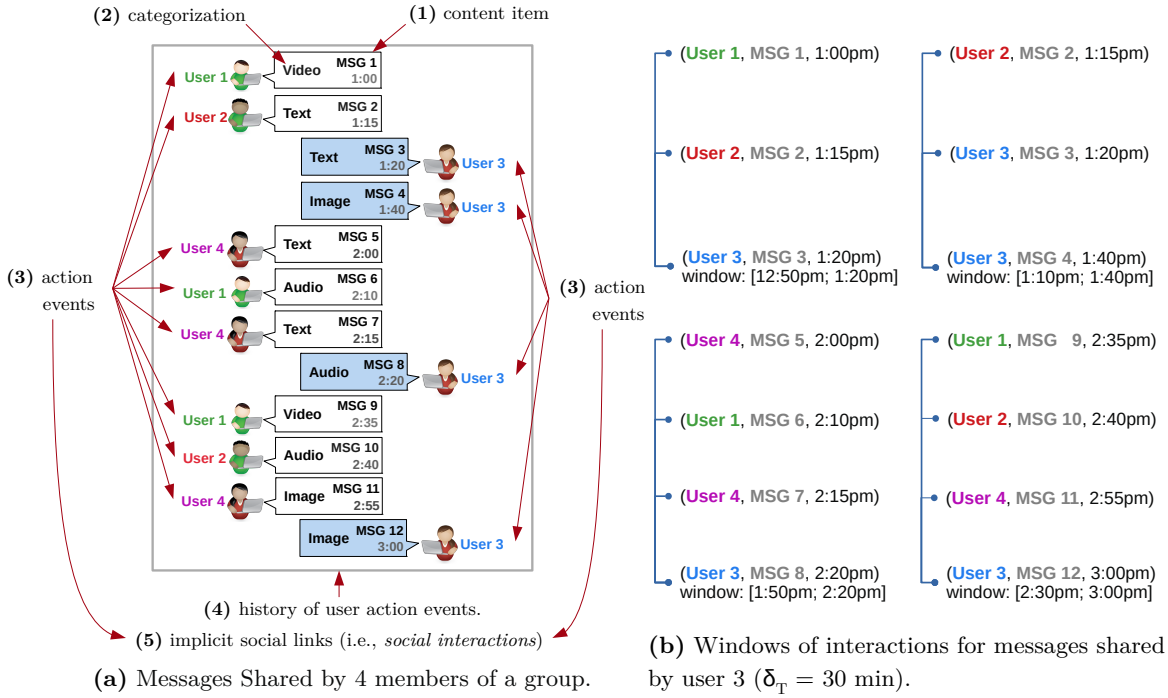
Figure 3.2: Main elements of our first case study: consumption of online music in LastFM.

messages, audio, video or image content). Moreover, unlike in the first case study, user curiosity may be stimulated not only by content and individual characteristics but also by the actions of other group members. In other words, we argue that, in a close environment of group communication (as is the case of a WhatsApp group), actions (message sharings) by group members may stimulate others' curiosity and influence them towards particular actions. However, estimating how one's curiosity is stimulated by others' actions is quite a challenge and, in the absence of explicit links connecting group members, must be inferred from prior user actions. Figure 3.3 illustrate the main elements of our second case study.

Having presented our two case studies, we introduce the main assumptions we make to model user curiosity in each such case next.

3.2 Assumptions

Our approach to model user curiosity relies on the derivation of a number of metrics capturing different collative variables associated with curiosity stimulation. These metrics are subsequently used to build user curiosity profiles, which in turn can be applied directly or by means of a function mapping stimuli to curiosity (e.g., the Wundt's curve) as a component of an information service. Towards deriving the proposed metrics, we



(a) Messages Shared by 4 members of a group.

(b) Windows of interactions for messages shared by user 3 ($\delta_T = 30$ min).

Figure 3.3: Main elements of our second case study: WhatsApp groups.

make three fundamental assumptions that hold for both case studies:

1. The curiosity of a user may be triggered differently depending on various features of the target platform and the types of user actions (e.g., content production, content consumption) under consideration. As such, though grounded on general (information theoretic) guidelines specified by [Berlyne, 1960; Silvia, 2006], the derivation of metrics capturing different collative variables should take the particularities of the target scenario into account.

2. The way that the curiosity of a user reacts to a given stimulus may change over time. In other words, a stimulus that triggers great curiosity in a person may later be mostly innocuous, or fall into other regions of the stimulation process (i.e., boredom, anxiety). Thus, the proposed metrics should be (re-)quantified at each action event of a user. Figure 3.2 depicts 9 action events related to item consumption (i.e., song listenings in this case). Thus, the level of curiosity stimulation the user is subjected to should be computed at the time of each listening. The same holds for the message sharings shown in Figure 3.3: the level of curiosity stimulus imposed on each user should be recomputed each time the user shares a new message.

3. The curiosity driving an action (be it the consumption, production or sharing of

an item) by a given user u at a given time t has a period of activation δ_T, that is, the time interval preceding the user's action event during which previous actions by the user u (and by other users, in case of social influence) may contribute to stimulate u's curiosity. We refer to that period as a *window of interaction*, and it includes all *action events* that may stimulate u's curiosity at time t. User actions that occurred before the window of interaction (i.e., prior to time $t - \delta_T$) are assumed to be too old to have any impact on u's curiosity at time t, being thus disregarded for the sake of computing the curiosity stimulation metrics at that time.

We exemplify the concept of window of interaction in the right side of Figures 3.2 and 3.3. Take the latter for illustration purposes. Assuming $\delta_T = 30$ minutes, the figure shows the window of interaction, notably the messages and users that fall within such window, defined for each message shared by user 3. The messages (and users) in a window will be used as input to compute the metrics of curiosity stimulation for user 3 at each given time. For example, user 3 shares a textual message (*msg 3*) at 1:20pm. The stimulus to user 3's curiosity driving this particular action will be computed based on the messages shared in the interval [12:50pm; 1:20pm]. Similarly, the stimulus to user 3's curiosity when sharing the audio message at time 2:20pm (*msg 8*) will be (re-) computed based on *msgs 5, 6, 7, and possibly 8* shared within the window [1:50pm; 2:20pm].

In addition to the aforementioned general assumptions, we also rely on a few more premises to design and derive metrics to capture social influence as a collative variable of curiosity stimulation. That is, in our second case study, discussed in Chapter 5, which addresses user curiosity in group communication on WhatsApp, we consider the following additional assumptions:

4. The curiosity of a user may be triggered differently depending on the people with whom the user is interacting and the ongoing discussions among them [Renner, 2006; Kashdan et al., 2020b]. In other words, one's curiosity stimulation may vary depending on the particular group the individual is participating in. More yet, if a person participates in multiple groups at the same time, her curiosity stimulation may vary across these groups. Thus our metrics will be computed and our analyses will be performed for different groups separately, considering only the messages shared in each group at a time.

5. The curiosity of a user sharing some content may be stimulated by the other users who shared content in the same group during the window of interaction

via *social influence*. We do acknowledge that the content shared itself as well as characteristics of the target user whose curiosity is under analysis may also stimulate one's curiosity. However, when modeling *social* aspects of curiosity, we disregard such characteristics, focusing rather only on all the other *users* who shared content during the window of interaction. Properties of the content and of the target user are exploited to derive metrics related to other collative variables.

As an example, the right side of Figure 3.3 shows the users considered as stimulating user 3's curiosity each time she shares a message. These are all users included in the defined window of interaction *but user 3*. For example, users 1 and 2 are considered to compute the stimulus at time 1:20pm, when user 3 shares her first message. Note that, although user 3 herself is also included in the window of interaction (due to *msg 3* shared at time 1:20pm), only the other users in the window are considered for the sake of analyzing the *social* curiosity driving user 3's behavior. Message *3*, shared at time 1:20pm, will be taken into consideration to derive metrics related to other collative variables, as further discussed in Chapter 5. Similarly, only user 2 is considered as stimulating user 3's curiosity when she shares *msg 4* at time 1:40pm.

6. The extent to which user's curiosity is stimulated by other users can be estimated by historical patterns. Specifically, the influence of user u_1 on user u_2, which relates to the curiosity stimulus triggered on u_2 by u_1, is proportional to the frequency at which a message shared by u_1 is followed by a message by u_2 within an interval not greater than δ_T (that is, u_1 falls in the window of interaction defined for a message shared by u_2). This frequency is estimated based on the message sharing patterns prior to the current window of interaction (thus, historical patterns).

   For example, in the right side of Figure 3.3, the influence of user 1 on user 3 at the time she shares *msg 12* is based on the frequency at which user 3 shared content after user 1 (within a 30 minute maximum interval), *before* the beginning of the current window of interaction, i.e., before 2.30pm. This implies that we will look at all messages shared by user 1 before 2:30pm and check how often user 3 shared a message after it, with a maximum delay of δ_T. In the particular scenario depicted in the figure, it happened twice: *msg 1* at 1:00pm was followed by *msg 3* at 1:20pm, and *msg 6* at 2:10pm was followed by *msg 8* at 2:20pm.

3.3 Problem Definition

We now formally introduce the problem we aim at investigating:

Problem Definition: *Given a set of users $\mathcal{U}$ of an information service and a set of content items $\mathcal{I}$ available on the service, where each item is characterized according to a set of predefined categories $\mathcal{C}$, we aim at quantifying the stimulus a user $u \in \mathcal{U}$ is exposed to when performing an action*[1] *on a content item $i \in \mathcal{I}$, characterized by (one or more) categories $c \in \mathcal{C}$, which serve as proxy for representing content properties.*

To that end, we we will explore a set of chronologically ordered tuples $(u, i, \mathcal{C}_s, t)$, where each tuple specifies that a user u acted on item i, characterized by the categories $\mathcal{C}_s \subseteq \mathcal{C}$ at time t. In other words, the set of tuples build a *history of users' action events* from which user curiosity (stimulation) will be inferred at each action time t. We assume $|\mathcal{U}| = n_{users}$, $|\mathcal{I}| = n_{items}$ and $|\mathcal{C}| = n_{cat}$.

3.4 Summary

In this chapter, we have discussed general elements of the work developed in this dissertation. We have introduced the two case studies as well as some fundamental assumptions that we use to guide our developments. We then finished the chapter with a formal definition of our target problem. All these elements serve as pillars for the information offered in the next two chapters where we present the studies performed on each scenario of investigation. We start our discussion in Chapter 4 with our first case study, focused on information consumption. Specifically, we present our efforts to model user curiosity as a force behind user's listening to online music.

[1] We note that, since in both case studies, user actions are of a single type, either music listenings (case study 1) or message sharings (case study 2), we do not consider the action type as a dimension of exploration and, as such, do not explicitly consider it in our notation. Yet, future extensions may consider different action types in a single scenario of study. In that case, notation must be adjusted to include them elements of the problem definition.

Chapter 4

Analyzing and Modeling User Curiosity in Online Content Consumption: A LastFM Case Study

This chapter is organized as follows. The next section presents the introduction of our first case study. The Section 4.2 discusses the modeling of user curiosity. Next, we present our results addressing the aforementioned research questions in Sections 4.3 and 4.4, and Section 4.5 offers a discussion of the results and possible directions for future work.

4.1 Introduction

Our personality traits naturally emerge as we perform our daily activities, these traits are manifested also in our online interactions. There has been a recent surge of studies on psychological aspects of online behavior thanks to the big data [Singer, 2016]. Such as the Big Five personality test to identify personality traits, which has been the focus of several efforts [Youyou et al., 2015; Santos, 2015]. By identifying such traits it is possible to tune and enhance services and technologies to better meet one's individual personality, ultimately improving user satisfaction [Zhao and Lee, 2016].

Several studies have shown that some personality traits are strongly associated with *curiosity* [Greasley et al., 2013; Youyou et al., 2015]. Indeed, curiosity has been recognized as a critical factor that influences human behavior in positive and negative

ways at all stages of life.

As discussed by prior efforts, an individual's curiosity is essentially driven by external stimuli [Berlyne, 1960; Loewenstein, 1994] which are captured by collative variables. These variables refer to different characteristics of a stimulus, i.e., external factors that govern how one's curiosity is stimulated [Berlyne, 1960]. Novelty, complexity, uncertainty and conflict are collative variables that capture aspects related to curiosity stimulation [Loewenstein, 1994].

Starting from the point that stimulus is as a combination of collative variables, the Wundt's curve has been proposed to model one's curiosity. The curve captures curiosity as a function of the stimulus intensity [Wu and Miao, 2013a]. A Wundt's follows a bell-shape, indicating that too-little or too-much stimuli will respectively lead an individual to relaxation and anxiety. Moderate stimulation leads to curious behavior.

Curiosity models based on the Wundt's curve have already been applied in the design of artificial creativity systems, to control the behavior of non-player characters in digital games and in reconfigurable robots [Merrick and Maher, 2009]. In the particular domain of online information consumption, Zhao *et al* [Zhao and Lee, 2016] recently used it in the design of a personalized recommendation system. Yet, the authors used a single collative variable, namely novelty, as stimulus to curiosity. The use of other collative variables, strongly related to curiosity, has not been investigated yet. The authors applied the model without assessing the extent to which it actually captures user curiosity in their target domain.

Our goal here is to model and analyze user curiosity as a driving force behind user consumption of online content. Building upon existing models [Zhao and Lee, 2016], we consider all collative variables as sources of stimulus to curiosity. Also, we investigate whether a single Wundt's curve is a reasonably model of user curiosity. As case study, we focus on the consumption of online music from LastFM, as musical tastes are related to personality characteristics [Greasley et al., 2013]. To that end, we analyze a large dataset consisting of over 243M listening events, covering more than 84K users and 9M songs. Specifically, we address two research questions:

- **Research Question 1 (RQ1)**: Are there distinct user behavior profiles in terms of curiosity stimuli, as captured by the collative variables?
- **Research Question 3 (RQ3)**: To which extent user curiosity can be accurately modeled by a Wundt's curve?

Given our focus on online music, we start by proposing stimulus metrics related

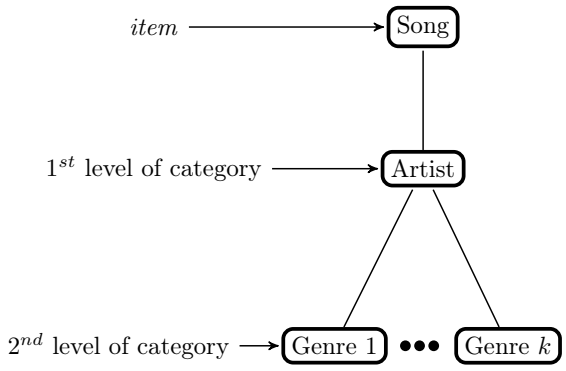


Figure 4.1: Hierarchy of content categorization for song (content items).

to songs, artists and their musical genres capturing the collative variables. Next, we find that, though the single Wundt's curve is a reasonable model, for more than 40% of them the curiosity curve is multi-modal, reflecting a mixed behavior. So, it can be well modelled by the combination of two or three Wundt's curves.

In sum, this chapter presents a thorough effort of modeling user curiosity as a driving force of online music consumption. Compared to prior studies, our contributions are: (1) the proposition and analysis of a set of metrics that capture different perspectives of the stimuli offered to users as they are exposed to online content; and (2) new models of curiosity that more accurately account for the diversity of user behavior.

4.2 Modeling User Curiosity

In this section, we present our notation used to define our metrics and, then, we present metrics for each one of collative variables of *novelty*, *uncertainty*, *complexity* and *conflict* in the context of LastFM.

4.2.1 Notation

Our context of analysis is composed by sequences of songs consumed by users in LastFM dataset, a platform of online music consumption. Therefore, a given song belongs a certain artist, which is considered as a *content categorization*, whose yet has a *categorization* of given musical genres. That categorization of song with respect to artist and musical genres is defined as *hierarchy of categorization* and can be visualized in Figure 4.1. In that picture, one can to see a possible structure of these categorization hierarchy in form of a tree, where: (a) the root node is the song; its respective child node is (b) the artist (i.e., the first level of content category); and, next below, there

Table 4.1: Basic notation used to derive the curiosity stimulation metrics.

Notation	Description
$\mathcal{U}$	set of all users (individual user indicated by u)
$\mathcal{S}$	set of all songs (one group indicated by s)
$\mathcal{A}$	set of all artists (one group indicated by a)
$\mathcal{C}$	set of all musical genres (i.e., categories) used as classification of artists or band
$\mathcal{C}_a$	set of musical genres associated with a given artist of listened song (one category is indicated by c)
$t_{i\|u}$	timestamp of the i^{th} song consumed by user u
δ_T	duration of window of interaction
u	user whose curiosity we are analyzing
$\vec{}$	indicates some quantity computed for current window of interaction
$\mathcal{S}_t^{\vec{}}$	set of (distinct) songs consumed during window$[t-\delta_T;t]$
$\mathcal{A}_t^{\vec{}}$	set of (distinct) artists from songs consumed during window$[t-\delta_T;t]$
$\mathcal{C}_t^{\vec{}}$	set of (distinct) musical genres of artists from songs consumed during window $[t-\delta_T;t]$
S	random variable associated with songs
A	random variable associated with artists
C	random variable associated with musical genres
$n_{s\|t}^{\vec{}}$	number of times that song s was consumed during window $[t-\delta_T;t]$
$n_{a\|t}^{\vec{}}$	number of times that artist a occurs for consumed songs during window $[t-\delta_T;t]$
$n_{c\|t}^{\vec{}}$	number of times that musical genre c occurs for consumed songs during window $[t-\delta_T;t]$
$P_t^{\vec{}}(X)$	Probability of random variable X computed based on window of interaction $[t-\delta_T;t]$

are (c) k leaves, child nodes of artist, which denote the musical genres, it denotes the second level of content category. We emphasize that an artist may have one or more musical genres.

Our present goal is to analyze and model user curiosity in the particular context of online music consumption. To that end, we consider the temporal sequence of listening events of a user u. Each listening event of user u is associated with a timestamp t and a unique song (track) s by an artist (singer/band) a. We use the existing musical genres of LastFM to categorizing the artists, then we define $\mathcal{C} = \{$"rnb", "rap", "eletronic", "rock", "new age", "classical", "reggae", "blues", "country", "world", "folk", "easy listening", "jazz", "vocal", "children's", "punk", "alternative", "spoken word", "pop", "heavy metal"$\}$.

The set $\mathcal{C}_a$ includes all musical genres used to classify the artist/band of a song. Thereby, each artist a has one or more tracks and is characterized by k musical genres with $1 \leq k \leq n_{cat}$, wherein $|\mathcal{C}| = n_{cat}$, and, yet, $\mathcal{C}_a \subseteq \mathcal{C}$. In turn, the set $\mathcal{U}$ denotes the set of users whom consume songs in LastFM platform and $u \in \mathcal{U}$. Thus, each listening event is defined by a tuple $(u, s, a, \mathcal{C}_a, t)$.

Also, the previous definitions demonstrate the basis of our notation. Variables are represented as small letters (e.g., user u and song s), random variables as capital letters (e.g., S and A to song and artist, respectively), and sets are represented as calligraphic

letters (e.g., $\mathcal{S}$ and $\mathcal{A}$, song and artist, respectively). We employ subscripts to determine subsets (e.g., songs of specific user u as S_u). Finally, we use the Bayesian notation to define constraints/dependencies (e.g., $t_{i|u}$ is the time of the i-th listening event from a given user u). The complete set of notation, including those used above as others introduced in the following sections, is listed in Table 4.1.

We deal with the final metrics in a hierarchy to show how they relate to one another: songs, artist and genres. This hierarchical approach is crucial to understand exactly which concepts drive user curiosity. That is, some users may be curious towards more specific concepts, such as songs, while others focus on general concepts such as genres. For genres, the metrics complement one another in capturing different aspects of the same concept.

4.2.2 Computing Stimuli

We begin by stating that one important factor when understanding user behavior online is time. In fact, temporal aspects such as recency (e.g., last time an item was accessed) will naturally correlate with our variables. In this sense, all metrics are computed for a given user u and specific listening event i, using as input all previous listening events in the window $[t_{i|u} - \delta_T, t_{i|u}]$. Thus, the curiosity of the user for access i depends on every previous access inside the window.

Based on such temporal aspect from assumption 3, it is important to consider that users' curiosity will likely dwindle after long periods. Thus, in our analysis, we consider a $\delta_T = 24$ hours. That is, the curiosity of the user depends **only** on the events from the last 24 hours, that is, during the window of interaction (thus the use of notation $^{\rightarrow}$ in this section). We leave the evaluation of other size windows as future work.

All of our metrics will be defined in the unit of *bits*. To do so, we shall take the $\log_2$ of probabilities and compute the surprisal (in bits) [MacKay, 2005]. This was a design choice to facilitate the aggregation of metrics. With s capturing the current song (track), a the current artist and $\mathcal{C}_a$ the current musical genres, we define $songNovelty(u, t = t_{i|u}, s)$, $artistNovelty(u, t = t_{i|u}, a)$ and $genreNovelty(u, t = t_{i|u}, \mathcal{C}_a)$ as the novelty of tracks, artists and musical genres to given artist. The novelty of given song is defined as:

$$songNovelty(u, t = t_{i|u}, s) = \begin{cases} -\log_2\left(P_t^{\rightarrow}(S = s)\right), & \text{if } P_t^{\rightarrow}(S = s) > 0 \\ -\log_2\left(1/|\mathcal{S}_t^{\rightarrow}|\right), & \text{otherwise} \end{cases} \quad (4.1)$$

where $P_t^{\rightarrow}(S = s)$ is the probability of song s being consumed during the window of interaction, which is defined as $n^{\rightarrow}_{s|t,u} / \sum_{s \in \mathcal{S}_u} n^{\rightarrow}_{s|t,u}$, where $n^{\rightarrow}_{s|t,u}$ denotes the number of times which the song s appeared into window of interaction of user u at time t. Note that $P_t^{\rightarrow}(S = s) = 0$ corresponds to maximal surprisal associated with song s. In such case, we set the surprisal to its maximum value possible, which corresponds to a uniform is the set of distinct songs, i.e., $-\log_2(1/|\mathcal{S}_t^{\rightarrow}|)$, where $\mathcal{S}_t^{\rightarrow}$ is the set of distinct songs that occur during the window of interaction $[t - \delta_T; t]$.

Similarly, artist novelty is defined by looking at artists instead of songs. Both metrics captures will be higher for less accessed artists or songs. So, it is given by

$$artistNovelty(u, t = t_{i|u}, a) = \begin{cases} -\log_2 \left(P_t^{\rightarrow}(A = a)\right), & \text{if } P_t^{\rightarrow}(A = a) > 0 \\ -\log_2 \left(1/|\mathcal{A}_t^{\rightarrow}|\right), & \text{otherwise} \end{cases} \tag{4.2}$$

where $P_t^{\rightarrow}(A = a)$ is the probability of artist a to appear for given song consumed into window of interaction. These probability is defined as $n^{\rightarrow}_{a|t,u} / \sum_{a \in \mathcal{A}_u} n^{\rightarrow}_{s|t,u}$, where $n^{\rightarrow}_{a|t,u}$ denotes the number of times which the artist a appeared into window of interaction of user u at time t. $\mathcal{A}_t^{\rightarrow}$ is the set of distinct artist that occur during the window of interaction $[t - \delta_T; t]$.

Given a particular artist of a consumed song at time t, let C be the random variable associated with musical genre, c be a musical genre associated with the given artist from a song (i.e., $c \in \mathcal{C}_a$) and $\mathcal{C}_t^{\rightarrow}$ be the *set* of musical genres associated with all messages shared within the window of interaction $[t - \delta_T; t]$. We recall that $\mathcal{C}$ contains all possible existing musical genres into LastFM platform, in total, there are 20 musical genres (vide Section 4.2.1). In the same way, as defined for previous metrics, we states the novelty metric related to musical genres, grounded on the surprisal associated with that variable:

$$genreNovelty(\mathcal{C}_a, t = t_{i|u}, a) = \begin{cases} -\log_2 \left(\bar{P}_t^{\rightarrow}(\mathcal{C}_a)\right), & \text{if } \bar{P}_t^{\rightarrow}(\mathcal{C}_a) > 0 \\ -\log_2 \left(1/|\mathcal{C}_t^{\rightarrow}|\right), & \text{otherwise} \end{cases} \tag{4.3}$$

where $\bar{P}_t^{\rightarrow}(\mathcal{C}_a)$ is the average probability of musical genres associated with the artist (musical genres in $\mathcal{C}_a$), and it is given by

$$\bar{P}_t^{\rightarrow}(\mathcal{C}_a) = \frac{1}{|\mathcal{C}_a|} \sum_{c \in \mathcal{C}_a} P_t^{\rightarrow}(C = c),$$

yet, the probability of a specific musical genre is estimated as $P_t^{\rightarrow}(C = c) = n^{\rightarrow}_{c|t,a} / \sum_{k \in \mathcal{C}_t^{\rightarrow}} n^{\rightarrow}_{k|t,a}$, wherein $n^{\rightarrow}_{c|t,a}$ is the number of musical genres equal to c of an artist

a from a song consumed in the window of interaction $[t - \delta_T, t]$. That metric captures how surprising it is that the song consumed at time $t = t_{i|u}$ has musical genre c.

We propose to capture uncertainty and conflict associated with given listening event based on the probabilities of the user listening to different musical genres, denoted by $P_t^{\rightarrow}(C = c)$. The probability of a class (i.e., musical genres) denotes the strength of the corresponding response, in which case the uncertainty of a stimulus relates to the Shannon entropy [MacKay, 2005; Cover and Thomas, 2006] of the classes. Specifically, uncertainty is quantified by the entropy of musical genres. Let $P_t^{\rightarrow}(C = c)$ as previously defined, it is computed for all genres listened by the user in the window of interaction (i.e., genres in set $\mathcal{C}_t^{\rightarrow}$). Uncertainty at current listening event is defined as:

$$genreUncertainty(t = t_{i|u}, a) = \sum_{c \in \mathcal{C}_t^{\rightarrow}} P_t^{\rightarrow}(C = c) \log_2 \left(P_t^{\rightarrow}(C = c) \right). \tag{4.4}$$

In the same way, conflict is estimated by the average probability of genre, computed over all genres listened by the user in the window of interaction:

$$genreConflict(t = t_{i|u}, a) = -\log_2 \left(\frac{1}{|\mathcal{C}_t^{\rightarrow}|} \sum_{c \in \mathcal{C}_t^{\rightarrow}} P_t^{\rightarrow}(C = c) \right). \tag{4.5}$$

Finally, we define the two complexity metrics, instantaneous and overall complexity, denoted by $instGenComplex$ and $overallGenComplex$ respectively, as:

$$instGenComplex(t = t_{i|u}, a) = -\log_2 \left(\frac{|\mathcal{C}_a|}{|\mathcal{C}|} \right) \tag{4.6}$$

and

$$overallGenComplex(t = t_{i|u}, a) = -\log_2 \left(\frac{|\mathcal{C}_t^{\rightarrow}|}{|\mathcal{C}|} \right), \tag{4.7}$$

where $|\mathcal{C}|$ is the total number of distinct genres in the complete LastFM platform. The instantaneous complexity is simply the fraction of the number of genres in the current song, $|\mathcal{C}_a|$, divided by the number of genres consumed so far (including this song), $|\mathcal{C}_t^{\rightarrow}|$. The overall complexity is measured as fraction of the total number of distinct genres listened so far, i.e. $|\mathcal{C}_t^{\rightarrow}|$, that exist inside window of interaction divided by the $|\mathcal{C}|$. The instantaneous complexity is related to the current song. It answers question's such as: how complex is this song when compared to the other songs consumed in the window of interaction. In contrast, the overall complexity is related to the window of interaction and answers questions as: How complex is the behavior of the user in the last 24 hours.

Yet, one can note that differently uncertainty and conflict, which captures the

Table 4.2: Metrics of curiosity stimulation for music consumption in LastFM.

Collative Variable	**Metric**	**Definition**
Novelty	$songNovelty(u, t, s)$	Novelty associated with user u whose consumed a given song s at time t (Equation 4.1)
Novelty	$artistNovelty(u, t, a)$	Novelty associated with user u whose consumed a given song of an artist a at time t (Equation 4.2)
Novelty	$genreNovelty(\mathcal{C}_m, t, a)$	Novelty associated with set $\mathcal{C}_a$ of musical genres of an artist a with respect to the song consumed by user u at time t (Equation 4.3)
Uncertainty	$genreUncertainty(t, a)$	Uncertainty associated with musical genres from all consumed songs, measured at time t (Equation 4.4)
Conflict	$genreConflict(t, a)$	Conflict associated with musical genres from a consumed song of artist a, measured at time t (Equation 4.5)
Complexity	$instGenComplex(t, a)$	Instantaneous complexity associated with categories of messages shared in group g, measured at time t (Equation 4.6)
Complexity	$overallGenComplex(t, a)$	Overall complexity associated with musical genres from consumed song of artist a, measured at time t (Equation4.7)

diversity of song with respect the artist's musical genres considering only those included in the window of interaction, in turn, complexity captures a somewhat different notion of diversity that takes into account (i) all possible musical genres in the current song and (ii) all possible genres listened in the window of interaction regarding to whole genres from LastFM.

In short, the Table 4.2 summarizes the metrics of curiosity stimulation for online music consumption in LastFM platform.

4.2.3 The LastFM Dataset

As a case study, we employ the LFM-1B Dataset of [Schedl, 2016]. This dataset captures the listening behavior of users on the LastFM website. The data covers the period of January 2013 to August 2014. Each event recorded contains the following features: user, artist, album, track and timestamp of listening. To capture complexity, we incorporate into the dataset the genre of artists (e.g., The Beatles belongs to the rock genre). In particular, we make use of the genres defined in [Schedl and Ferwerda, 2017]. It is comprised of: *rnb, rap, eletronic, rock, new age, classical, reggae, blues, country, world, folk, easy listening, jazz, vocal, children's, punk, alternative, spoken word, pop* and *heavy metal*. Table 4.3 summarizes our dataset.

Table 4.3: Statistics of LFM-1B Dataset.

Item	Number
Users	84,466
Artists	219,055
Albums	4,351,218
Tracks	9,319,305
Listening events	243,555,074
Users with at least 1K listening	43,220

Recall that we compute the metrics over time windows of 24 hours. In order to avoid data sparsity issues, we consider in our analysis only users with at least 1,000 listening events in the whole period. We also only consider events with at least 30 previous listening events in each 24-hour window between January 2013 to August 2014. After such filtering, we were left with 43,220 users with 149,949,511 stimulus events.

4.3 Understanding Metrics

Before combining the proposed metrics to build a user curiosity model we first perform a careful exploration of each individual metric. In particular, we identify complementary and redundant metrics, before computing our final stimulus. Moreover, we also present a characterization of user profiles based on the subset of metrics that we consider.

4.3.1 Variable Selection

In order to understand how metrics relate to one another, we measured the Spearman correlation coefficient ρ_s between all possible pairs of metrics. Recall that this coefficient ranges from $[-1, 1]$, with each extreme indicating perfect correlations. The Spearman coefficient was measured by creating a vector for each metric considering every access i of a user. Thus, if a user has accessed 10K songs, we have 7 vectors of size 10K to correlate.

Our end goal is to filter out redundant metrics. Thus, with the correlation coefficient we aim to identify metrics where for the majority of users the Spearman correlation exceeds a threshold of ± 0.5. In particular, we determine that two metrics are *redundant* when: over 90% of users have the value ρ_s *outside* of the range: $\rho_s \notin [-0.5, +0.5]$. When correlations are in this range for 90% of the users, we define the pair of metrics as complementary.

We define the operation $A \longleftrightarrow B$ as the relationship of correlation between two variables, where A and B are two variables from stimulus metric. It serves as initial evidence of redundant metrics. After inspecting the figure, we computed our redundancy threshold (above). We found evidence redundant relationships for:

- $genreConflict \longleftrightarrow genreUncertainty$;
- $genreConflict \longleftrightarrow overallGenComplex$;
- $overallGenComplex \longleftrightarrow genreUncertainty$.

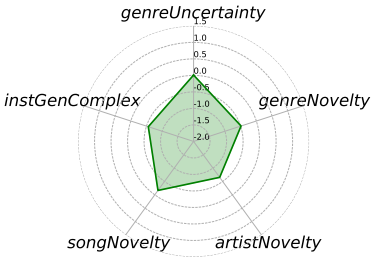
To understand these findings, recall that $overallGenComplex$ is related to the number of genres inside window's period of time. $genreConflict$ is the probability of genre on the window. In the end, both metrics end up capturing correlated stimuli. Also, $genreUncertainty$ is the entropy of genres inside window which, by definition, is derived from $genreConflict$.

Although there are some cases where above correlation are somewhat expected, there are interesting cases of metrics capturing semantically related concepts (e.g., the novelty of artists, songs, and genres) that were not correlated. By filtering out redundant metrics, we can better understand the curiosity of users. Based on these findings, we keep five metrics: $songNovelty$, $artistNovelty$, $genreNovelty$, $instGenComplex$ and $genreUncertainty$.

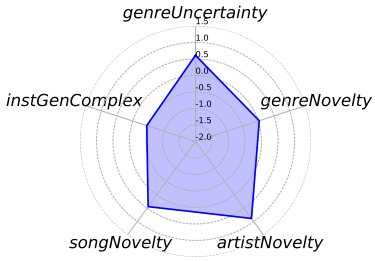
4.3.2 Clustering by Access Profile

We now turn our attention to uncovering the common curiosity patterns in our dataset. According to Section 4.2.3, there are close to 150M listening events when we consider the 43K users with at least 1,000 listening events. In order to understand the common patterns, we employ a clustering approach In particular, measure each one of the five *non-redundant* metrics). In the end, we have a 150M by 5 matrix that we use to cluster *listening events*.

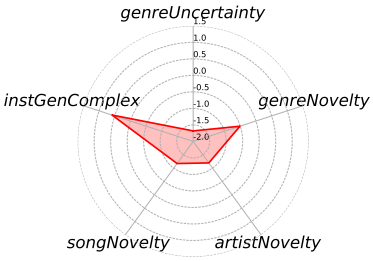
In our clustering, we make use of the Mini-Batch K-Means that that is suitable for large Web datasets [Sculley, 2010]. To define the number of clusters k, we made use the β_{CV} and Average Silhouette score. The former, Average Silhouette score, is an aggregate measure of how similar an element is to its own cluster compared another clusters. The later, β_{CV}, is defined as the ratio of coefficient of variation (CV) of intra-clusters distances and CV of inter-cluster distances. The smallest value of k where β_{CV} remains roughly stable is selected [Menasce and Almeida, 2000].



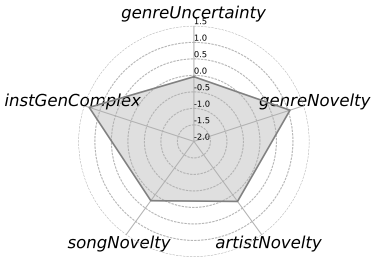
(a) C_0: Underground.



(b) C_1: Eclectic.



(c) C_2: Niche Radical.



(d) C_3: Explorer.

Figure 4.2: Radar-chart with normalized stimulus metrics per cluster.

Before clustering, we re-normalized the metrics using a Z-scale. It is well known that K-Means will lead to poor clustering when values are not normalized. Even though we omit detailed figures due to space constraints, we point out that both strategies (β_{CV} and Average Silhouette score) pointed to $k = 4$ as the more adequate choice.

To interpret our clusters, in Figure 4.2 we depict the centroids of each cluster. Recall that a centroid is simply the average of each of the 5 metrics. To provide further evidence that our clustering is adequate, we computed the confidence interval of 95% for each metric from the centroids. Based on these intervals, we verified that each cluster is different from the others (there is no overlap in the intervals for each metric).

To better understand our clusters, we adopt the labels used by [Ramos et al., 2013]. In particular, it is important to understand that each cluster summarizes the type of listening events, not users. These are labelled as follows:

- C_0 (*Underground*): The access is towards an item with similar stimuli metrics. However, the values are low indicating a tendency to stay in the relaxation zone for every metric. There are $48,627,376$ events in this cluster and it is exhibited in the Figure 4.2a and represents 32.43% of listening events;

- C_1 (*Eclectic*): Notice that the accesses in this cluster is regularly spread out across each metric. Different from the previous case, the metrics are higher (i.e. higher stimulus). $52,829,351$ events and it is in the Figure 4.2b with 35.23% of listening events;

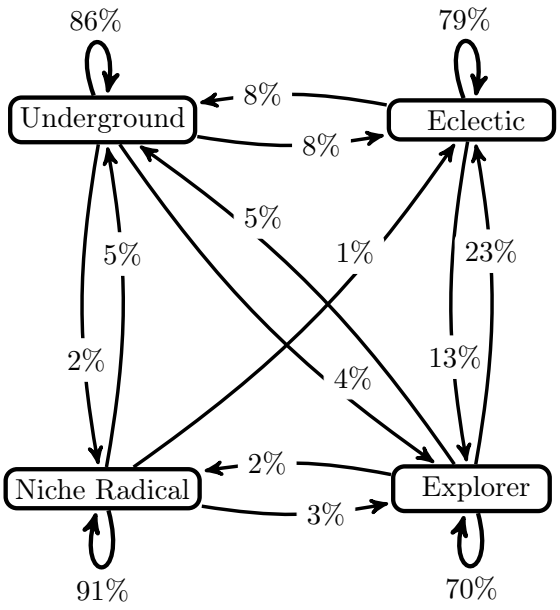


Figure 4.3: Typical CBMG from overall user behavior.

- C_2 (*Niche Radical*): the Figure 4.2c exhibits the radar-chart for this cluster, a set of access which the user listening to repeated songs (smaller novelty). Also the songs are more complex. In total, $18,645,990$ events and denotes 12.44% of events;

- C_3 (*Explorer*): the Figure 4.2d shows the *Explorer*. The event is very similar to *Eclectic* with $29,846,794$ events. However, its main difference is the accesses pursuing more complex music (i.e. songs with many genres) representing 19.90%.

To better understand how users transition across different clusters, we built a Customer Behavior Model Graph (CBMG) [Menasce and Almeida, 2000]. This is shown in Figure 4.3. The CBMG is a Markovian model. On it, nodes represent the event profile (i.e. *Underground*, *Eclectic*, *Niche Radical* and *Explorer*) and the arcs denote the transitions of specific access's profile to another. The weight of arcs denotes probabilities of transitions pattern occurring. The sum of all outgoing probabilities for each state is 1. We weight each edge as being proportional to the number of transitions across the clusters for every user.

At a first glance, one can clearly see a tendency towards repeated behavior. This is somewhat expected as repeated consumption of songs/artists is common online [Benson et al., 2016]. What is more interesting is how users tend to have higher transitions across the *Eclectic* and *Explorer* behavior. While the self-loop is high in these cases,

there is a tendency (above 13%) to migrate. When the users are in the *Underground* and *Niche Radical* phases, this tendency to migrate decreases.

So far, our results point out that user behavior in terms of curiosity is quite complex. Each cluster represents a different pattern of behavior (metric values). Also, users tend to transition across these patterns. These results motivate the need of Wundt models that are able to capture such complexities. These models are discussed in the next section.

4.4 Curiosity Models

We now tackle our second research question: To which extent user curiosity can be accurately modelled by a Wundt's curve? Following the approach of to Zhao *et al* [Zhao and Lee, 2016], we can estimate a *single* Wundt's curve as a Beta distribution. The Beta is a continuous probability distribution with shape parameters $\alpha > 1$ and $\beta > 1$. The probability density function (PDF) of Beta distribution is:

$$b_{\alpha,\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \quad \text{where} \quad B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

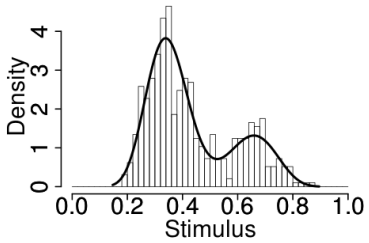
and Γ is the Gamma function, α and β are two positive shape parameters. The Beta is a flexible distribution that is able to capture different shapes related to the Wundt curve. In our analysis, the accesses of each user will be used to learn multiple Wundt curves. That is, we shall estimate Beta distributions for each user using as input the access metrics.

To fit the Beta models to our data, we need to map the final value of curiosity to the $[0, 1]$ range (the domain of the Beta). Thus, we initially take the mean of the 5 metrics (notice that they are all in bits) for each listening event. Thus, we are capturing the average number of bits of the access. We then map the mean to $[0, 1]$ range using a min-max normalization.

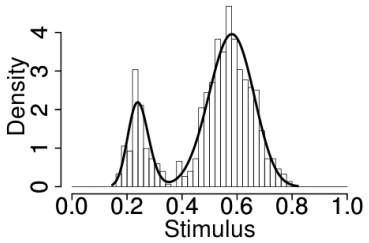
One of the fallacies of Zhao *et al* [Zhao and Lee, 2016] was to assume that each user may be modelled by a single Wundt curve. Here, we shall show that this is not the case. To capture more complex behavior, we employ a mixture of Beta distributions, i.e.:

$$f(x) = \pi_1 b_{\alpha_1,\beta_1}(x) + \ldots + \pi_M b_{\alpha_M,\beta_M}(x) = \sum_{i=1}^{M} \pi_i b_{\alpha_i,\beta_i}(x),$$

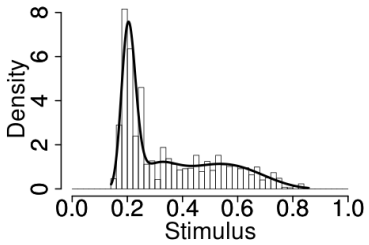
where M is the number of mixture components and π_i, α_i and β_i are respectively the i-th mixture component and two shape parameters from Beta mixture distribution. We



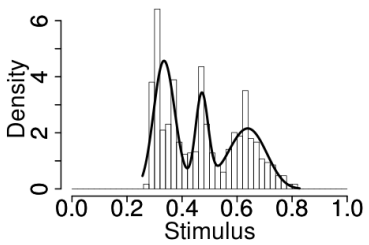
(a) 2 Beta: User ID 2349062.



(b) 2 Beta: User ID 3058490.



(c) 3 Beta: User ID 3058795.



(d) 3 Beta: User ID 4432965.

Figure 4.4: Mixed Wundt curve: examples of 4 users.

estimate mixture of betas considering $M \in \{1, 2, 3\}$. For the estimation of parameters, we used the Iterated Method of Moments [Schröder and Rahmann, 2017].

For each user in our dataset, we fit Beta mixtures considering $M \in \{1, 2, 3\}$. To validate each mixture, we employed the Kolmogorov-Smirnov (KS) Test. Notice that we are fitting thousands of users, and thus are performing thousands of hypothesis (one for each user) tests. Given that this is a multiple comparison setting, to correct our p-values, we employ the Benjamini-Hochberg (BH) method. After p-value correction, we employ a *corrected* significance level of $\alpha^+ = 0.01$. We state the each user is modelled by the smallest mixture of Betas, smallest M, that is validated by the test.

So, when we consider all of the users, for $M \in \{1, 2, 3\}$ mixture of Betas 57.85% are accurately modelled by a single Beta, 36.46% of users require a two Beta mixture, and 3.96% of users are modelled by three Betas. In the end, only 1.76% were not validated by at least three Betas. These users were discarded. To exemplify our results, in Figure 4.4 we show four examples of users that were modelled with two and three mixtures. The Figures 4.4a, 4.4b and 4.4d clearly show the multi-modality of curves.

We also correlated the mixture of Betas models with the types of access (clusters). To do so, Table 4.4 exhibits the fraction per access profile inside of each mixture model. This fraction is computed over all of the accesses for every user captured with 1, 2 or 3 mixtures (each column sums to 100%). From the table, we can see that *Underground*

Table 4.4: Statistics of Access Profile values per user.

Access Profile	1 Beta	2 Beta mix	3 Beta mix
Underground	35.54%	35.09%	36.11%
Eclectic	34.39%	32.39%	20.49%
Niche Radical	10.70%	15.38%	27.47%
Explorer	20.37%	17.14%	15.93%
Total:	100.00%	100.00%	100.00%

has almost constant presence in each different model (rows). In contrast, *Eclectic* and *Explorer* show a decrease in presence as the number of mixture increases. This indicates that users captured with a single beta will mostly be stimulated by *Underground* and *Eclectic* accesses. More complex users, captured more (2 or 3) Betas, show a higher variation in such behavior. Overall, these findings serve as evidence that users that transition across multiple clusters will require more mixtures.

4.5 Summary

In this work we provide an in-depth analysis on how to model curiosity in online content consumption. Motivated by advances both in data mining as well as psychology, we model curiosity as a mixture of Beta distribution. Each distribution capturing a Wundt's curve. In particular, we answer two research questions:

- **RQ1**: *Are there distinct user behavior profiles in terms of curiosity stimuli, as captured by the collative variables?*
- **RQ3**: *To which extent user curiosity can be accurately modeled by a Wundt's curve?*

Our findings are important as it unveils some of the flaws of prior endeavors. In particular, we carefully inspect which collative variables should be used to model curiosity stimuli. Then we show that user curiosity is more complex than it seems, with the same user possibly being curious towards different profiles (clusters) of stimuli (RQ1). Moreover, for a large fraction of users we require more than one Wundt curve to model curiosity (RQ3).

Thus, this work serve as a basis for designing user-centric applications that consider curiosity. For instance, by identifying whether the user is in anxiety or boredom zone it is possible to stimulate the user to move until the curiosity zone recommending the right items. So, the user will achieve higher levels of curiosity score, hence leading to the user to involve more deeply into system. As future work, we can extend the

findings in this chapter to provide better personalized search results and recommendations. Finally, our results can be naturally extended to other domains besides music since that there is some way to classify the set of items on categories. Thus, it will be possible to compute the stimulus metrics according to collative variables and to fit the user's curiosity.

Chapter 5

Metrics of Social Curiosity: The WhatsApp Case

This chapter is organized as follows. Section 5.2 presents initial considerations about the work of this chapter. Next, Section 5.3 describes the WhatsApp communication platform. Our novel metrics of curiosity are introduced in Section 5.4. Our characterization results, including the dataset used, are discussed in Section 5.5. We discuss some limitations of our study and their implications in Section 5.6. Finally, summary and future work are offered in Section 5.7.

5.1 Introduction

Curiosity has been characterized as an important trait of our personality, influencing one's behavior in positive and negative ways at all stages of life [Kidd and Hayden, 2015]. According to [Berlyne, 1960; Loewenstein, 1994], an individual's curiosity is essentially driven by external stimuli: an individual repeatedly receives stimuli from the environment and selectively responds to those that induce pleasure. This selection process has been explained by means of *collative variables*, notably novelty, complexity, uncertainty and conflict, which capture different external factors that govern how one's curiosity is stimulated [Berlyne, 1960]. In essence, each stimulus is indeed a combination of various collative variables, instantiated by different metrics, and the response to a stimulus depends on the particular individual upon whom it is imposed. A number of studies have explored models of human curiosity, relating an input stimulus to the curiosity induced in an individual, in the design of personalized and enhanced systems [Wu et al., 2014; Twomey and Westermann, 2018; Gordon et al., 2015; Macedo and Cardoso, 2005; Arif et al., 2020].

In the particular domain of online information dissemination, there have been recent efforts to formally introduce curiosity models into the design and evaluation of recommendation systems [Zhao and Lee, 2016; Chen et al., 2019; Xu et al., 2019; Wang et al., 2020; Shandhilya and Srivastava, 2020]. However, most of them have used metrics related to a single collative variable, and have neglected one important factor that may impact one's curiosity: social influence. Social influence has been widely studied as a key component in various behavioral phenomena, from information dissemination to opinion adoption (e.g., [Bakshy et al., 2012; Kloumann et al., 2015]). Yet, the focus has been mostly on those who have more influence on the community [Adamic and Adar, 2003; Zhu et al., 2020] and their impact on the phenomenon under study [Huang et al., 2020; Choi et al., 2020], and do not explicitly analyze the impact of such influence on individual behavior. Other studies have explored principles derived from social influence to design recommendation systems (e.g., [Wu et al., 2016, 2017; Shokeen and Rana, 2020]), but we are aware of only one recent effort to explicitly apply metrics related to social influence as part of a curiosity model, which in turn is used in the design of a recommendation method [Xu et al., 2019].

In contrast, some other studies have already discussed the concept of *social curiosity*, a facet of curiosity, which has been defined as the general interest in gaining new social information motivating exploratory behaviors [Hartung and Renner, 2013; Renner, 2006; Litman and Pezzo, 2007]. Specifically, social curiosity entails two different aspects: a general interest in obtaining new information about how others think, behave or act as well as an interest in interpersonal information that is obtained through exploratory behavior. As such, social influence can be seen as yet another collative variable that stimulates one's curiosity and ultimately drives one's behavior.

In this context, our present effort is driven by three research questions:

- **Research Question 2.1 (RQ2.1):** *How to quantify social influence as a stimulus to one's curiosity driving the information dissemination process?*
- **Research Question 2.2 (RQ2.2):** *How does social influence relate to other collative variables priorly associated with curiosity stimulation?*
- **Research Question 2.3 (RQ2.3):** *How are users characterized in terms of social stimulation to curiosity?*

We tackle these questions using WhatsApp as case study. WhatsApp is a free messaging app that has surpassed the mark of 2 billion users worldwide in 2020[1]. It

[1] `https://www.whatsapp.com/about/`

footnote after punctuation.

connects users in end-to-end as well as group conversations, and the latter has been shown to be an effective vehicle for information dissemination at large [Resende et al., 2019b]. Thus, we here propose to *model the social stimulus to the curiosity that drives users to share content (thus communicating with each other) in WhatsApp groups.*

To that end, we start by proposing a set of novel metrics to quantify such stimulus (RQ2.1). Towards deriving these metrics, we follow the seminal work by Berlyne [Berlyne, 1960], which proposes a methodology to quantify collative variables related to curiosity stimulation based on information theoretical metrics. Berlyne applied these concepts to the derivation of metrics related to more traditional collative variables (e.g., novelty, conflict). We here follow his arguments, as well as similar ones by other authors [Silvia, 2006], and apply his methodology to derive metrics capturing the effects of social influence as a component of curiosity stimulation.

Moreover, following the same methodology and building on our prior experience [Sousa et al., 2019], we also adapt metrics capturing other collative variables related to curiosity stimulation, notably novelty, complexity, conflict and uncertainty to the particular domain of WhatsApp groups. These variables have been studied before in other domains [Zhao and Lee, 2016; Xu et al., 2019; Wu and Miao, 2013a; Sousa et al., 2019] but not in a platform of group communication, such as the one provided by WhatsApp.

Our study is carried out in a dataset consisting of over 2 million messages shared by more than 7.5 thousand users in 335 publicly accessible groups in Brazil during a one-month period. By correlating the two sets of metrics – metrics related to social curiosity and metrics related to the other collative variables – we show that our social curiosity metrics do indeed capture novel aspects of one's curiosity stimulation which are not represented by the other (priorly studied) collative variables (RQ2.2).

Finally, we use our metrics to characterize the social curiosity stimulation of users when sharing content in WhatsApp groups (RQ2.3). Our characterization is performed at three levels of aggregations: we quantify social curiosity stimulation driving the sharing of individual messages, the overall behavior of individual users as well as WhatsApp groups.

Our main contributions can be summarized as follows:

- We propose four new metrics to capture social influence as a component of curiosity stimulation at the user level, as well as one metric to capture the same aspect at the group level. We instantiate the proposed metrics for the specific context of content sharing in WhatsApp groups, but they are general enough to be applied to other platforms of online group communication. Additionally,

we also propose seven metrics that capture other traditionally studied collative variables, namely novelty, complexity, conflict and uncertainty, to the same context. To our knowledge, this is the first effort to propose metrics to quantify components of curiosity stimulation in such environment.

- We offer an extensive evaluation of social influence as a component of the curiosity stimulation driving content sharing in WhatsApp groups. Our study reveals that social influence, as captured by our proposed metrics, is complementary to the other traditionally analyzed collative variables, thus reflecting a novel and important component of the curiosity stimulation process. Also, we found great diversity and dynamics in social curiosity stimulation in WhatsApp, at the message, individual and group levels. Specifically, we found three profiles of social curiosity stimulation driving user behavior when sharing individual messages, and used them to uncover five different profiles describing how the social curiosity of a user in a given group evolves over time. We also found evidence that the social curiosity stimulation of a particular user may be quite different depending on the group she participates in, which hints at a role the group has on its members. Yet, different groups may exhibit quite different and very dynamic group-level social curiosity stimulation, as such dynamics results from the aggregation of the behavior of those users that are more actively sharing content at each time.

5.2 Initial Considerations

In this chapter, in addition to the four aforementioned collative variables in the Section 2.1, social influence should also be considered as part of the stimulus to one's curiosity, especially in online social networks and social media applications, where, to a large extent, one's behavior is driven by connections, common interests with others and social observation. This argument is aligned with the concept of *social curiosity*, recently introduced in [Kashdan et al., 2018, 2020a], as a key component of a five dimensional curiosity model. However, that studies were based on surveys with real people.

Here, we are interested in proposing quantifiable measures of social curiosity as a driving force behind online behavior. Specifically, we propose different metrics to quantify the impact of social influence as a component of curiosity stimulation driving users to share content on WhatsApp. To our knowledge, prior attempts to operationalize the concept of social curiosity by proposing quantitative metrics to estimate it are quite scarce and mostly limited [Zhao and Lee, 2016; Xu et al., 2019; Sousa et al., 2019]. This is probably due to the challenges to capture, quantitatively, the aspects associ-

ated with such a highly subjective concept, as argued by several recent studies in the Psychology and Neuroscience domains [Kashdan et al., 2018, 2020a; Valji et al., 2019; Ahmadlou et al., 2021; Lau et al., 2020; Alicart et al., 2020]. The derivation of our metrics, discussed in Section 5.4, follows the rationale on Berlyne's original arguments and his proposed methodology, here adapted to capture social curiosity. Specifically, we explore metrics from information theory to estimate how such highly subjective concept can compose the curiosity stimulation process in the particular setup of interest, i.e., group communication on WhatsApp. To our knowledge, we are the first to operationalize Berlyne's ideas to estimate social curiosity, especially in the selected domain.

The main research gap we address in this chapter with respect to the studies in Section 2.2.2 is the modeling of a novel collative variable associated with curiosity stimulation, namely social curiosity. Thus, the closest study to ours is that by Xu *et al.* [Xu et al., 2019]. However, unlike this prior work, we do not aim at improving item recommendation. Rather, our primary focus is on modeling user behavior, with particular interest in how users' decisions to share content is driven by social curiosity. Moreover, we explore a completely different setting – content sharing in a group communication application, where curiosity most probably is stimulated differently. We here must capture how the curiosity that drives one to communicate (by sharing content) is impacted by the other members participating in the group. To that end, we propose several new metrics that help uncovering important components of (social) curiosity stimulation. As such, our work is completely orthogonal to [Xu et al., 2019].

5.3 WhatsApp

WhatsApp has become one of the main communication platforms in many countries [Resende et al., 2019b]. It allows for one-to-one and group conversations, both encrypted. Groups are, by default, private spaces limited to 256 simultaneous users, although a user may join and leave a group at any time. Yet, as shown in [Resende et al., 2019b], group administrators can make a group publicly accessible by sharing an invitation link in public websites, since anyone with the link can join the group. By gathering a large number of such publicly available invitation links, researchers were able to join the groups automatically and, once a member, gather data for posterior analysis.

For example, some researchers developed automatic tools to expose, in an anonymized fashion, the content being shared in publicly accessible groups [Melo et al.,

2019; Garimella and Tyson, 2018], whereas others analyzed properties of such content [Moreno et al., 2017] with notable focus on pieces of information that had been previously checked as fake by fact checking agencies [Resende et al., 2019b,a; Maros et al., 2020; Caetano et al., 2019]. Orthogonally, Melo *et al.* analyzed whether limiting message forwarding could mitigate misinformation spread on WhatsApp [Melo et al., 2019]. They found that, though effective in slowing down the process, such approach would not stop misinformation from being widely distributed.

In a complementary directions, a few prior studies focused on exposing the importance of the underlying networks that connect users across different WhatsApp groups to information spread [Resende et al., 2019b; Nobre et al., 2020]. In [Resende et al., 2019b], the authors analyzed the structural properties of the network built from connecting users belonging to the same group, finding that this network has several properties, often observed in other online social networks, that had been associated with content virality. More recently, Nobre *et al.* focused on the media co-sharing network, built by connecting users who shared the same content, revealing the presence of strongly connected user communities that consistently help speeding up information spread [Nobre et al., 2020].

Despite the recent surge of interest in analyzing content properties and user sharing patterns in WhatsApp, we are not aware of any prior attempt to investigate how to model social aspects of curiosity in this platform, as we do here. We believe that understanding the drivers behind users' sharing actions is key to proper modeling and understanding information dissemination on the platform. We take a step in that direction by proposing metrics to quantify one such driver – social curiosity.

5.4 Novel Metrics of User Curiosity

In this section, we present our novel metrics of curiosity stimulation driving user participation in group communication. Although we use WhatsApp as case study, the metrics are derived to be applicable to group communication in general. We consider that group membership is naturally dynamic, as members join or leave the group at their will. However, it is expected that (a subset of) members remain interacting with each other, exchanging opinions and content in general, for some time. We here are interested in quantifying the extent to which the bond created by such interactions can stimulate a member's curiosity towards carrying on the conversation. Specifically, we focus on quantifying stimuli to participating in such group conversations by *sharing content*. In other words, we propose metrics that capture different aspects of the stimuli

one is driven by when choosing to share a piece of content with the group. As mentioned before, our derivation of the proposed metrics follows Berlyne's arguments and his proposed methodology [Berlyne, 1960]. Specifically we make use of measures from information theory to derive metrics that aim to capture different collative variables associated with curiosity stimulation.

In the following, we first present the notation used to define our metrics (Section 5.4.1). We then introduce our novel metrics of social curiosity, which aim to capture social influence as a stimulus to curiosity and constitute a key contribution of this work (Section 5.4.2). Next, we present metrics that instantiate other traditionally studied collative variables to the particular context of WhatsApp groups (Section 5.4.3). The latter are inspired by the metrics introduced in [Sousa et al., 2019], originally developed for the context of online music consumption, and here adapted to a very different domain, notably content sharing in a group communication platform. They are used in this work for comparison purposes, so as to emphasize the complementary role of social influence as a component of curiosity stimulation. We defer to Section 5.6 a discussion on limitations introduced by some of our assumptions and design decisions and their implications to the study.

5.4.1 Notation

We consider a universe of analysis consisting of sequences of messages shared by a number of users in various (independent) groups. A message is composed of pieces of content in any of four different media types, namely text, image, audio or video. Figure 3.3 of Chapter 3 illustrates the case of a sequence of 12 messages shared by 4 user members of a given group in the interval between 1:00pm and 3:00pm. The left part of the figure shows the sequence as seen by user 3, with the messages shared by her highlighted in blue. Note that the figure shows, for each message, the user who shared it, the message identifier, the content's media type and the time of sharing. In the following, as in the next sections, we will use this example to illustrate the main concepts behind our metrics.

We now introduce the notation used to derive them in the following sections. We start by defining the sets of users, groups, and types of media used to compose the sequence of messages under analysis as $\mathcal{U}$, $\mathcal{G}$, $\mathcal{M}$, respectively. We do not assume access to message content, as it may not be available. Instead, we use the different types of media used to compose the message content as a representation of it, that is, we define $\mathcal{M} = \{$"text", "image", "audio", "video", "URL"$\}$[2]. Each message under

[2]Other possible categories include emojis, stickers and gifs. However, these are not available in

analysis is then a tuple $(u, \mathcal{C}_m, g, t)$ indicating that it was shared by user $u \in \mathcal{U}$ in group $g \in \mathcal{G}$ at time t. Set $\mathcal{C}_m \subseteq \mathcal{M}$ includes all media types used to compose the message's content. We refer to such media types as the *categories* of the message. Note that the same message may contain content in different medias (e.g., a text and an image), thus a message may have multiple categories captured in set $\mathcal{C}_m$[3].

We also note that, even though URLs are presented in textual format, we do consider them a separate category because URLs refer to external content (e.g., a post in another platform), which is not immediately visible to the user. Thus, from the perspective of curiosity stimulation, we speculate they may have a different impact compared to the rest of the textual content (which is immediately visible). More broadly, the use of media types as message categories is based on the assumption that different media types require different amounts of effort from the user to see and process the content, which should impact how the user curiosity is stimulated by it. We further elaborate on this point and discuss the implications of our choice of media types as categories in Section 5.6.

We use notation $t_{i|u,g}$ to refer to the timestamp of the i^{th} message shared by user u in group g. We assume time is continuous starting at 0. Thus, no two messages can be shared at exactly the same time in a group. Each group is composed by a set of users $\mathcal{U}_g \subseteq \mathcal{U}$, with size $|\mathcal{U}_g|$. Moreover, let the groups be associated with the random variable G.

These definitions illustrate the basis of our notation. Variables are represented as small letters (e.g., user u and group g), random variables as capital letters (e.g., G), and sets are represented as calligraphic letters (e.g., $\mathcal{G}$ and $\mathcal{U}$). We employ subscripts to determine subsets (e.g., users in a group $\mathcal{U}_g$), and the Bayesian notation to define constraints/dependencies (e.g., $t_{i|u,g}$ is the time of the i^{th} message shared by a given user u in a given group g). The complete set of notation, including those used above as others introduced in the following sections, is listed in Table 5.1.

5.4.2 Social Curiosity

Having presented the general concepts, notation and assumptions, we proceed by zooming in the novel metrics that capture the impact of social influence as a driver to user curiosity (or simply social curiosity [Kashdan et al., 2018]). Recall that these metrics are meant to capture the influence that a set of users have on a particular (distinct) target user, as stimuli to this user's curiosity towards sharing content. Thus, in order to

our dataset, and thus are not explored in this work.

[3]The same media type used multiple times to build a message's content (e.g., multiple images associated with it) is counted only once as part of $\mathcal{C}_m$.

Table 5.1: Main notation used to derive the curiosity stimulation metrics.

Notation	Description
$\mathcal{U}$	set of all users (individual user indicated by u)
$\mathcal{G}$	set of all users (one group indicated by g)
$\mathcal{M}$	set of all media types (i.e., categories) used as content representation ($\mathcal{M}$ = {"text", "image", "audio", "video", "URL"})
$\mathcal{C}_m$	set of categories associated with a given message (one category indicated by c)
$\mathcal{U}_g$	set of user members of group g
$t_{i\|u,g}$	timestamp of the i^{th} message shared by user u in group g
δ_T	duration of window of interaction
d	destination user whose curiosity we are analyzing
o	origin user who may stimulate the curiosity of a destination at a given time
$\rightarrow$	quantity computed for current window of interaction
$\leftarrow$	quantity computed for historical windows
$\mathcal{U}^{\rightarrow}_{t,g}$	set of (distinct) users sharing content in group g during window of interaction $[t-\delta_T; t]$
$\mathcal{O}^{\rightarrow}_{t_{i\|d,g}}$	set of (distinct) origins for destination d in group g at time $t_{i\|d,g}$ (all users, except for d, who shared content during the window $[t_{i\|d,g}-\delta_T; t_{i\|d,g}]$)
$\mathcal{C}^{\rightarrow}_{t,g}$	set of (distinct) message categories shared during window$[t-\delta_T;t]$
$\mathbf{S}^{\leftarrow}_{t,g}$	contingency table with historical patterns computed for group g at time t (see Figure 5.1)
$n^{\rightarrow}_{u\|t,g}$	number of times user u shared content in group g during window $[t-\delta_T; t]$
$n^{\rightarrow}_{c\|t,g}$	number of messages of category c shared in group g during window $[t-\delta_T; t]$
$n^{\leftarrow}_{o,d\|t,g}$	number of times o and d shared content in group g (in that order) within a δ_T time interval *before* the window $[t-\delta_T; t]$
G	random variable associated with groups
O	random variable associated with origin users
D	random variable associated with destination users
C	random variable associated with the categories of messages
$P^{\leftarrow}_{t,g}(X)$	Probability of random variable X computed based on historical patterns within group g before window of interaction $[t-\delta_T; t]$
$PMI^{\leftarrow}_{t,g}(D{=}d, O{=}o)$	Pointwise mutual information between the destination d and origin o in group g, measured at time t (Equation 5.1)
$socInf^{\leftarrow}_{t,g}(D{=}d, O{=}o)$	Social influence from origin o on destination d in group g, measured at time t (Equation 5.2)
$MI^{\leftarrow}_{t,g}(D, O{=}o)$	Mutual information of all destinations conditioned on an particular origin o in group g, measured at time t (Equation 5.5)
$indSocInf^{\leftarrow}_{t,g}(D, O{=}o)$	Indirect social influence from origin o on all destinations (users in D) in group g, measured at time t (Equation 5.6)
$H^{\leftarrow}_{t,g}(D \mid O{=}o)$	Entropy of destinations conditioned on a particular origin o in group g, measured at time t (Equation 5.9)
$groupEntropy(t{=}t_{i\|d,g}, g)$	Group-level entropy of destinations conditioned on all origins in group g, measured at time t (Equation 5.10)
$H^{\leftarrow}_{t,g}(D)$	Entropy of destinations in group g, regardless of any social influence, measured at time t (Equation 5.11)

distinguish between the user whose curiosity we are analyzing, i.e., user u who shared a message in group g at time $t_{i|u,g}$, and the other users who may stimulate u's curiosity through social influence at time $t_{i|u,g}$, we refer to the former as the *destination* and the latter as *origins* (of social influence).

In the following, we use notations d to refer to a particular destination user and o to refer to one of his origins. We also define $\mathcal{O}^{\rightarrow}_{t_{i|d,g}}$ as the *set* of all origins for the destination user d sharing a message at time $t_{i|d,g}$, i.e., the set of all other users who shared content in the window of interactions $[t_{i|u,g} - \delta_T;\ t_{i|u,g}]$ and thus may stimulate d's curiosity at time $t_{i|u,g}$. Note that, by definition, the destination d herself does not belong to $\mathcal{O}^{\rightarrow}_{t_{i|d,g}}$, regardless if she also shared content during that period. In other words, we are interested in looking into the *other* users stimulating d's curiosity. Also, as a set, multiple occurrences of the same origin in the window of interaction count as one.

Given our assumption that influence is estimated based on historical patterns (assumption 5), we must distinguish between the *current* window of interaction and previous/historical windows (always with duration δ_T) which occurred before the current window initiated. Historical windows are used to estimate the influence of each origin o in d's content sharing at time $t_{i|d,g}$. We make such distinction by adopting notation $^{\rightarrow}$ (as in $\mathcal{O}^{\rightarrow}_{t_{i|d,g}}$) to refer to the current window of interaction and $^{\leftarrow}$ to refer to historical windows. Moreover, since we assume social influence may change over time, it must be recomputed using all available historical interactions up to (the beginning of) the current window of interaction.

For each window of interaction defined by a message sharing at time $t = t_{i|d,g}$, historical interactions are captured by a contingency table $\mathbf{S}^{\leftarrow}_{t,g}$. Each cell in the table determines the number of times origin o and destination d shared content in group g (in that order) within a maximum time interval of δ_T *before* the current window of interaction started (i.e., before time $t - \delta_T$), with $o, d \in \mathcal{U}_g$ and $o \neq d$. We use notation $n^{\leftarrow}_{o,d|t,g}$ to refer to this number. Note that, by definition, $n^{\leftarrow}_{d,d|t,g} = 0$, i.e., self-influence (diagonal) is not considered. For the general case of a group with $|U_g|$ users, table $\mathbf{S}^{\leftarrow}_{t,g}$ has size $|U_g| \times |U_g|$.

Figure 5.1 shows the contingency table computed for the example in Figure 3.3 at 3:00pm, when user 3 shared *msg 12*. All messages shared up to the beginning of the window of interaction defined for *msg 12* should be considered. Thus, *msgs 1–8*, shared before 2:30, are taken into account to compute the table. The column representing user 3 as destination is shown in red. For example, we can see that user 3 followed a message by user 1 within $\delta_T = 30$ minutes twice in the past (specifically when user 3 shared *msgs 3* and *8*). Similarly, user 3 followed messages by user 4 within

Origin of influence \ Destination of influence	u_1	u_2	u_3	u_4	Total of rows:
u_1	0	1	2	1	4
u_2	0	0	2	0	2
u_3	1	0	0	1	2
u_4	1	0	2	0	3
Total of columns:	2	1	6	2	Total: 11

Figure 5.1: Contingency table $\mathbf{S}^{\leftarrow}_{t,g}$ computed when user 3 shares message 12 in Figure 3.3 of Chapter 3.

that interval also twice in the past: *msg 8* by user 3 follows *msgs 5* and *7* by user 4. The figure also shows the joint count of the number of messages for each origin (row), each destination (column) and the overall total.

Now, let O and D be random variables associated with origins and destinations, respectively. Given the contingency table computed for a message shared at time t by destination d, we define the following probabilities for a group g, computed based on historical patterns (note the use of $^{\leftarrow}$):

(a) Probability of an origin o on group g:

$$P^{\leftarrow}_{t,g}(O=o) = P^{\leftarrow}(O=o \,|\, t_{i|d,g}=t, G=g) = \frac{\sum_d n^{\leftarrow}_{o,d|t,g}}{\sum_{o,d} n^{\leftarrow}_{o,d|t,g}};$$

(b) Probability of a destination d on group g:

$$P^{\leftarrow}_{t,g}(D=d) = P^{\leftarrow}(D=d \,|\, t_{i|d,g}=t, G=g) = \frac{\sum_o n^{\leftarrow}_{o,d|t,g}}{\sum_{o,d} n^{\leftarrow}_{o,d|t,g}};$$

(c) Probability of a destination given an origin:

$$P^{\leftarrow}_{t,g}(D=d \,|\, O=o) = P^{\leftarrow}(D=d \,|\, O=o, t_{i|d,g}=t, G=g) = \frac{n^{\leftarrow}_{o,d|t,g}}{\sum_d n^{\leftarrow}_{o,d|t,g}}.$$

(d) Joint probability of a destination and an origin:

$$P^{\leftarrow}_{t,g}(D=d, O=o) = P^{\leftarrow}(D=d, O=o, t_{i|d,g}=t, G=g) = \frac{n^{\leftarrow}_{o,d|t,g}}{\sum_{o,d} n^{\leftarrow}_{o,d|t,g}}.$$

Note that all probabilities are conditioned on the group and timestamp of the current message. Thus, to simplify notation we drop these two conditions, using sub-

scripts instead. That is, we define $P^{\leftarrow}_{t,g}(\cdot) = P^{\leftarrow}(\cdot \mid t_{i|d,g} = t, G = g)$ as exemplified above[4]. We make the same simplification in the notation used below to improve readability.

Taking the contingency table shown in Figure 5.1 as input, we can make, for example, the following computations at the time when user 3 shares *msg 12*:

(a) Probability of origin 3: $P^{\leftarrow}_{t,g}(O = u_3) = \frac{2}{11}$;

(b) Probability of destination 3: $P^{\leftarrow}_{t,g}(D = u_3) = \frac{6}{11}$;

(c) Probability of destination 3 given origin 2:
$P^{\leftarrow}_{t,g}(D = u_3 \mid O = u_2) = \frac{2}{2} = 1$;

(d) Probability of destination 3 given origin 4:
$P^{\leftarrow}_{t,g}(D = u_3 \mid O = u_4) = \frac{2}{3}$;

(e) Joint probability of destination 3 and origin 4:
$P^{\leftarrow}_{t,g}(D = u_3, O = u_4) = \frac{2}{11}$.

Having defined these probabilities, we are ready to define our new metrics of social curiosity. In total, we propose four new metrics of the stimulation a user is exposed to by social influence from others in the group. All metrics are based on the concept of *mutual information* [Cover and Thomas, 2006]. In essence, the Mutual Information (MI) of two random variables is a measure of the mutual dependence between them. Specifically, it quantifies the reduction in the *uncertainty* (or entropy [MacKay, 2005]) associated with one variable as we observe the other random variable.

Our first two metrics are based on the concept of Pointwise Mutual Information (PMI), which is computed for a pair of outcomes x and y belonging to discrete random variables X and Y. In our case, we define the pointwise mutual information of a particular destination d and a particular origin o as [Bossomaier et al., 2016]:

$$PMI^{\leftarrow}_{t,g}(D{=}d, O{=}o) = \begin{cases} \log_2\left(\frac{P^{\leftarrow}_{t,g}(D{=}d \mid O{=}o)}{P^{\leftarrow}_{t,g}(D{=}d)}\right), \text{if } P^{\leftarrow}_{t,g}(D{=}d) > 0. \\ 0, \qquad\qquad\qquad\qquad \text{otherwise.} \end{cases} \tag{5.1}$$

The PMI of d and o measures the reduction in the uncertainty of destination d sharing a message due to the knowledge that o also shared a message (within the

[4]In other words, unless otherwise noted, whenever t is used we refer to time $t_{i|d,g}$ when the i^{th} message of the target destination d was shared in group g.

window of interaction)[5]. This reduction comes from the social influence that o has on d, which in turn stimulates d's curiosity. It means that, based on historical patterns (social influence estimate), the behavior of (i.e., sharing of a message by) user d in the group is influenced by recent behavior of user o. Naturally, we set the PMI value to 0 if there is no history of messages shared by d (i.e., $P^{\leftarrow}_{t,g}(D{=}d) = 0$).

Note that, for the sake of estimating the social influence of o over d as a stimulus to d's curiosity, we are interested in *positive values* of PMI. However, there are cases in which Equation 5.1 returns non-positive values. For example, $P^{\leftarrow}_{t,g}(D = d \mid O = o) = P^{\leftarrow}_{t,g}(D = d)$ suggests that the actions of d and o are independent, i.e., o has no influence of d's behavior. In that case, the PMI value is equal to 0. Similarly, if $P^{\leftarrow}_{t,g}(D{=}d \,|\, O{=}o) = 0$, there is no historical evidence of messages shared by o and d within a maximum interval δ_T, and, as defined, there is no evidence of social influence from o on d. In this case, the PMI value is negative. More generally, the PMI value will be negative whenever $P^{\leftarrow}_{t,g}(D{=}d \mid O{=}o) < P^{\leftarrow}_{t,g}(D{=}d)$. This case implies that the particular pair of destination d and origin o occurs less frequently than would be expected under the assumption of independent behavior, which might reflect unreliable statistics. Regardless, in all such cases, there is clearly no evidence of social influence of o on d. Thus, as in prior studies [S and Kaimal, 2012], we choose to clip negative values of PMI at 0, thus defining the following estimate of the social influence of origin o on destination d:

$$socInf^{\leftarrow}_{t,g}(D{=}d,O{=}o){=}\max\Big(PMI^{\leftarrow}_{t,g}(D{=}d,O{=}o),0\Big). \tag{5.2}$$

Given that the current window of interaction may have multiple origins stimulating the curiosity of destination d, we need to aggregate the mutual information computed for each origin o that appears in set $\mathcal{O}^{\rightarrow}_{t_{i|d,g}}$. We do so by computing the average and maximum social influence for all origins o on d. We thus build two metrics capturing the average and the maximum *direct influence* of the origins on a destination d, *based on historical patterns*:

$$avgDirInf(d, t{=}t_{i|d,g}, g) = \frac{\sum_{o\in\mathcal{O}^{\rightarrow}_{t_{i|d,g}}} socInf^{\leftarrow}_{t,g}(D{=}d, O{=}o)}{|\mathcal{O}^{\rightarrow}_{t_{i|d,g}}|} \tag{5.3}$$

[5]Alternative information theory-based metrics, such as transfer entropy [Schreiber, 2000; Ver Steeg and Galstyan, 2013] could also be used. We here chose PMI, instead, as it refers to single events, i.e., a single message shared by a user. Transfer entropy, instead, being based on mutual information, is meant to capture an aggregate relationship between two processes – specifically, the amount of directed transfer of information between them.

$$maxDirInf(d, t{=}t_{i|d,g}, g){=}\max_{o\in\mathcal{O}^{\rightarrow}_{t_{i|d,g}}}\Big(socInf^{\leftarrow}_{t,g}(D{=}d, O{=}o)\Big) \tag{5.4}$$

In addition to direct influence, we argue that some users may have a natural ability to influence others. This is not a novel concept, neither in general nor in the particular case of WhatsApp, although we are the first to propose a metric to quantify it as a component of curiosity stimulation. Indeed, Guo *et al.* [Guo et al., 2019] made a distinction between direct and indirect influence, emphasizing that the latter may not have an obvious manifestation and may have gradual consequences on user behavior. Caetano *et al.* [Caetano et al., 2021], in turn, identified the presence of (what they called) activists in public WhatsApp groups, i.e., users who often share content and, in doing so, keep the discussion going by driving others to also share content. This latter observation, in particular, motivated us to include metrics to capture such *indirect* influence: in a constrained (i.e., limited to 256 simultaneous users) and often focused space of communication, such as a WhatsApp group, it is reasonable to expect that these very active users (and activists) may indeed influence others much more often than it would occur in other unconstrained environments. That is, we consider that, even in the absence of prior experiences (i.e., $P^{\leftarrow}_{t,g}(D = d \mid O = o) = 0$), such influencers may still stimulate the curiosity of a user (e.g., a newcomer) in the group.

We identify these *indirect* influencers by searching for origins o that tend to have *strong* influence towards some destinations d. Given this rationale, we propose two other metrics of social curiosity to capture the impact of such indirect influencers on destination d at time $t_{i|d,g}$. These metrics are based on the *mutual information of the destinations conditioned on an particular origin*, that is:

$$MI^{\leftarrow}_{t,g}(D,O{=}o){=}\sum_{d'\in\mathcal{U}_g}P^{\leftarrow}_{t,g}(D{=}d',O{=}o)PMI^{\leftarrow}_{t,g}(D{=}d',O{=}o) \tag{5.5}$$

We note that, as done in Equation 5.2, in order to capture *indirect social influence*, we clip non-positive values of mutual information at 0 as a reflection of no social influence from origin o on any destination $d \in D$. We thus define:

$$indSocInf^{\leftarrow}_{t,g}(D, O{=}o){=}\max\Big(MI^{\leftarrow}_{t,g}(D, O{=}o), 0\Big) \tag{5.6}$$

The random variable of interest in Equations 5.5 and 5.6 is D, conditioned on a particular origin o. Smaller values of $indSocInf^{\leftarrow}_{t,g}(D, O{=}o)$ suggest weaker and no clear influence of o over any user $d \in D$. Take, for example, the case when $P^{\leftarrow}_{t,g}(D{=}d' \,|\, O{=}o)$

is roughly uniform for all destinations d'. Based on the values of $P^{\leftarrow}_{t,g}(D = d' \mid O = o)$, origin o does not strongly influence any particular destination d. The mutual information is this case is minimum. In contrast, larger values of $indSocInf^{\leftarrow}_{t,g}(D, O{=}o)$ imply that the influence of origin o tends to be concentrated and stronger on fewer destinations. These are the cases we are searching for. That is, the higher the value of $indSocInf^{\leftarrow}_{t,g}(D, O{=}o)$, computed based on historical patterns, the higher the *indirect* influence that o may have over the target destination d under analysis, i.e., the user whose curiosity stimulation is being quantified[6].

Again, we aggregate the above metric for all origins in the window of interaction via the average and maximum functions, referring to them as average and maximum *indirect* influence:

$$avgIndInf(d, t{=}t_{i|d,g}, g) = \frac{\sum_{o \in \mathcal{O}^{\rightarrow}_{t_{i|d,g}}} MI^{\leftarrow}_{t,g}(D, O{=}o)}{|\mathcal{O}^{\rightarrow}_{t_{i|d,g}}|}, \tag{5.7}$$

$$maxIndInf(d, t{=}t_{i|d,g}, g) = \max_{o \in \mathcal{O}^{\rightarrow}_{t_{i|d,g}}} \Big(MI^{\leftarrow}_{t,g}(D, O{=}o) \Big). \tag{5.8}$$

As mentioned, the indirect social influence metrics are particularly useful to capture the influence that an origin may exhibit on a destination she has not interacted with yet. If this is a recurring pattern, the direct influence metrics, computed based on $P^{\leftarrow}_{t,g}(D = d \mid O = o)$, will rise as time passes.

Going back to the example in Figure 3.3, let's estimate the metrics of social curiosity applied to user 3 at the time t = 3:00pm, when *msg 12* was shared. To that end, we will use the contingency table shown in Figure 5.1. The computation of all metrics are shown in Table 5.2. Note that users 1, 2 and 4 are origins of influence for user 3 at this time (i.e., they shared content during the window of interaction). Using Equations 5.1-5.4, we compute the average and maximum direct social influence of these origins on user 3 as $avgDirInf(u_3, t{=}3{:}00, g) = 0.39$ and $maxDirInf(u_3, t{=}3{:}00, g) = 0.87$, respectively. Also, using Equations 5.5–5.8, we compute average and maximum indirect social influence of the origins on user 3 as $avgIndInf(d, t = 3{:}00, g){=}0.59$ and $maxIndInf(d, t = 3{:}00, g){=}0.87$, respectively.

Before moving forward, we note that mutual information can also be used to quantify how social influence drives curiosity of a group of users. As such, it may offer

[6]We note that it may seem at first that Equations 5.5 and 5.6 lack this target destination d. Yet, recall that, as the metrics in Equations 5.1 – 5.4, it is conditioned on the time $t = t_{i|d,g}$ that the target destination d shared her i^{th} message in group g. This condition was omitted from the equations simply to improve readability.

Table 5.2: Computing the metrics of social curiosity for destination user 3 at time t=3pm (reference: Figure 5.1).

Metric	Value
$PMI^{\leftarrow}_{3:00,g}(D=u_3, O=u_1)$	−0.13
$PMI^{\leftarrow}_{3:00,g}(D=u_3, O=u_2)$	0.87
$PMI^{\leftarrow}_{3:00,g}(D=u_3, O=u_4)$	0.29
$socInf^{\leftarrow}_{3:00,g}(D=u_3, O=u_1)$	0.00
$socInf^{\leftarrow}_{3:00,g}(D=u_3, O=u_2)$	0.87
$socInf^{\leftarrow}_{3:00,g}(D=u_3, O=u_4)$	0.29
$avgDirInf(d, t=t_{i\|d,g}, g)$	0.39
$maxDirInf(d, t=t_{i\|d,g}, g)$	0.87
$MI^{\leftarrow}_{t,g}(D, O=u_1)$	0.42
$MI^{\leftarrow}_{t,g}(D, O=u_2)$	0.87
$MI^{\leftarrow}_{t,g}(D, O=u_4)$	0.48
$indSocInf^{\leftarrow}_{t,g}(D, O=u_1)$	0.42
$indSocInf^{\leftarrow}_{t,g}(D, O=u_2)$	0.87
$indSocInf^{\leftarrow}_{t,g}(D, O=u_4)$	0.48
$avgIndInf(d, t=t_{i\|d,g}, g)$	0.59
$maxIndInf(d, t=t_{i\|d,g}, g)$	0.87

an aggregate view of (social) curiosity stimulation in the ecosystem of a particular group g up to time $t_{i|d,g}$ when a member d shares a message. Towards defining such a metric, we start by introducing the entropy of destinations conditioned on a particular origin o [MacKay, 2005], given by:

$$H^{\leftarrow}_{t,g}(D\,|O{=}\,o) = -\sum_{d' \in \mathcal{U}_g} P^{\leftarrow}_{t,g}(D=d'|O=o) \log_2\Big(P^{\leftarrow}_{t,g}(D=d'|O=o)\Big). \tag{5.9}$$

We then aggregate the conditional probability in Equation 5.9 over all origins in the group to build a group-level metric :

$$groupEntropy(t=t_{i|d,g}, g) = \sum_{o \in \mathcal{U}_g} P^{\leftarrow}_{t,g}(O{=}o)\, H^{\leftarrow}_{t,g}(D\,|\,O=o). \tag{5.10}$$

The *groupEntropy* metric should be analyzed in light of the entropy of destinations D *regardless of social influence*, defined as follows:

$$H^{\leftarrow}_{t,g}(D) = -\sum_{d \in \mathcal{U}_g} P^{\leftarrow}_{t,g}(D=d) \log_2\big(P^{\leftarrow}_{t,g}(D=d)\big)\,. \tag{5.11}$$

Based on Equations 5.10 and 5.11, we introduce the group-level mutual information between destinations and origins to capture the reduction in uncertainty associated with *destinations* due to the knowledge of the social influence from other users (origins)

in the group. Specifically, we define:

$$groupMutInf(t{=}t_{i|d,g},g){=}H^{\leftarrow}_{t,g}(D){-}groupEntropy(t{=}t_{i|d,g},g). \tag{5.12}$$

The higher the mutual information, the greater the influence of origins over the destinations in group g. This may happen because either there are some highly influential origins or group curiosity is often stimulated by many group members.

Going back once again to our example, we also estimate the group-level social curiosity at the time user 3 shared *msg 12* as follows. We first compute the entropy of destinations $H^{\leftarrow}_{3{:}00,g}(D) = 1.91$ and the group-level entropy of destinations conditioned on origins $groupEntropy(t{=}t_{i|d,g}, g) = 1.39$. Based on these two values, we compute $groupMutInf(t{=}3{:}00, g) = 1.91 - 1.39 = 0.52$ which corresponds to only 27% of the original entropy of destinations in the group. Thus the knowledge of the social influence from the origins causes a reduction of only 27% $\left(\frac{0.52}{1.91}\right)$ on the entropy of destinations. We could say that, up to this time, the group is not very strongly stimulated by social influence.

A summary of all user/individual and group level metrics of social curiosity is presented in the top part of Table 5.3. Unless otherwise noted, all metrics should be interpreted as: *the larger the value, the greater the curiosity stimulation.*

5.4.3 Other Collative Variables

Having introduced our metrics of social curiosity, we now present 7 metrics that instantiate other collative variables, notably novelty, uncertainty, conflict and complexity, priorly associated with curiosity stimulation. We build on metrics originally proposed to the context of online music consumption [Sousa et al., 2019], adapting them here to the domain of content sharing in group communication. Based on assumption 3, we assume once again that user curiosity is triggered based only on what happens (*who* shares *what*) during the window of interaction[7] (thus the use of notation $^{\rightarrow}$ throughout this section). Therefore, the metrics are computed based solely on properties of the *user* who is sharing the content (i.e., the destination user) and of the content being shared, as captured by the media types (i.e., categories) associated with it. A summary of these metrics is presented in the bottom part of Table 5.3.

We here propose two metrics that instantiate the *novelty* component of the stimulus imposed on destination d at time $t_{i|d,g}$. The first one – $userNovelty$ – captures

[7]This was also a guideline for deriving the social curiosity metrics. However, for those metrics in particular, we used historical patterns to estimate the social influence of origins and destination identified in the window of interaction.

Table 5.3: Metrics of curiosity stimulation for content sharing in WhatsApp groups.

Colative Variable	**Metric**	**Definition**
Social curiosity (individual)	$avgDirInf(d,t,g)$	Average direct influence of all origins on destination d at time t in group g (Equation 5.3)
Social curiosity (individual)	$maxDirInf(d,t,g)$	Maximum direct influence of any origin on destination d at time t in group g (Equation 5.4)
Social curiosity (individual)	$avgIndInf(d,t,g)$	Average indirect influence of all origins on destination d at time t in group g (Equation 5.7)
Social curiosity (individual)	$maxIndInf(d,t,g)$	Maximum indirect influence of any origin on destination d at time t in group g (Equation 5.8)
Social curiosity (group)	$groupMutInf(t,g)$	Mutual information between destinations and origins in group g at time t (Equation 5.12: large values imply many strong curiosity stimulators in group)
Novelty	$userNovelty(d,t,g)$	Novelty associated with destination d sharing message in group g at time t (Equation 5.13)
Novelty	$catNovelty(\mathcal{C}_m,t,g)$	Novelty associated with set $\mathcal{C}_m$ of categories of the message shared by destination d in group g at time t (Equation 5.14)
Uncertainty	$userUncertainty(t,g)$	Uncertainty associated with users in group g, measured at time t (Equation 5.15)
Uncertainty	$catUncertainty(t,g)$	Uncertainty associated with categories of messages in group g, measured at time t (Equation 5.16)
Conflict	$userConflict(t,g)$	Conflict associated with users sharing content in group g, measured at time t (Equation 5.17)
Conflict	$catConflict(t,g)$	Conflict associated with categories of messages shared in group g, measured at time t (Equation 5.18)
Complexity	$catComplex(t,g)$	Complexity associated with categories of messages in group g, measured at time t (Equation 5.19)

the novelty related to the user sharing the content, i.e., user d. The idea is that the experience of sharing content is less novel to users who share more often, which ultimately affects the curiosity driving the action. This claim is inspired by arguments by Loewenstein [Loewenstein, 1994], who states that novelty is determinant of exploratory behavior in the individual's curiosity, and by Kashdan [Kashdan et al., 2018, 2020a], who argues that social curiosity may also be stimulated by novelty. Thus, we propose user novelty as a means to capture possible effects that a novel experience may have as a factor driving a user to participate in a group by sharing content. To our knowledge, we are the first to propose a novelty metric related to users, but we believe that in the specific domain of investigation – group communication – this variable may be a relevant component of curiosity stimulation. Indeed, as we will show in Section 5.5,

the proposed metric is indeed complementary to the others in many cases, which is an indication that, to some degree, the metric does capture new information.

The *userNovelty* metric is defined as the *surprisal*[8] associated with destination d:

$$userNovelty(d,t{=}t_{i|d,g},g){=}\begin{cases}-\log_2\big(P^{\rightarrow}_{t,g}(D{=}d)\big), \text{if } P^{\rightarrow}_{t,g}(D{=}d){>}0\\ -\log_2\big(1/|\mathcal{U}^{\rightarrow}_{t,g}|\big), \quad \text{otherwise}\end{cases} \tag{5.13}$$

where $P^{\rightarrow}_{t,g}(D{=}d)$ is the probability of d sharing content in group g during the current window of interaction, being thus defined as $n^{\rightarrow}_{d|t,g}/\sum_{u\in\mathcal{U}_g} n^{\rightarrow}_{u|t,g}$. Note that $P^{\rightarrow}_{t,g}(D{=}d) = 0$ corresponds to the maximum surprisal associated with destination d. In such case, we set the surprisal to its maximum value possible, which corresponds to a uniform distribution of destinations, i.e., $-\log_2\left(1/|\mathcal{U}^{\rightarrow}_{t,g}|\right)$, where $\mathcal{U}^{\rightarrow}_{t,g}$ is the *set* of distinct users sharing content in group g during the current window of interaction $[t-\delta_T; t]$.

Recall that we assume users may share messages with content in the following categories, defined based on media type: $\mathcal{M} = \{$"text", "image", "audio", "video", "URL"$\}$. Given a particular message shared at time t, let C be the random variable associated with category, c be a category associated with the given message (i.e., $c \in \mathcal{C}_m$) and $\mathcal{C}^{\rightarrow}_{t,g}$ be the *set* of media types associated with all messages shared within the window of interaction $[t-\delta_T; t]$. Similarly to what was done for users, we define a novelty metric related to the message categories, based on the surprisal associated with that variable:

$$catNovelty(\mathcal{C}_m,t{=}t_{i|d,g},g){=}\begin{cases}-\log_2\Big(\bar{P}^{\rightarrow}_{t,g}(\mathcal{C}_m)\Big), \text{if } \bar{P}^{\rightarrow}_{t,g}(\mathcal{C}_m){>}0\\ -\log_2\big(1/|\mathcal{C}^{\rightarrow}_{t,g}|\big), \quad \text{otherwise}\end{cases} \tag{5.14}$$

where $\bar{P}^{\rightarrow}_{t,g}(\mathcal{C}_m)$ is the average probability of categories associated with the message (categories in $\mathcal{C}_m$), defined as:

$$\bar{P}^{\rightarrow}_{t,g}(\mathcal{C}_m) = \frac{1}{|\mathcal{C}_m|}\sum_{c\in\mathcal{C}_m} P^{\rightarrow}_{t,g}(C{=}c).$$

The probability of a specific category is estimated as $P^{\rightarrow}_{t,g}(C{=}c){=}n^{\rightarrow}_{c|t,g}/\sum_{k\in\mathcal{C}} n^{\rightarrow}_{k|t,g}$, where $n^{\rightarrow}_{c|t,g}$ is the number of messages with category c shared in group g in the window of interaction $[t-\delta_T; t]$[9]. This metric captures how surprising it is that the message

[8] *Surprisal* is also called of *Shannon information content* [MacKay, 2005].

[9] Note that since messages may have multiple categories, the same message may be counted multiple times in the numerator and in the denominator of $P^{\rightarrow}_{t,g}(C{=}c)$.

shared at time $t = t_{i|d,g}$ has category c.

Taking the example of user 3 sharing *msg 12* at 3:00pm, illustrated in Figure 3.3, we find that a total of four messages were shared during the current window of interaction, defined as [2:30,3:00]. The destination, user 3, shared only once. Thus, the novelty associated with the destination is computed as: $P^{\rightarrow}_{3:00,g}(D = u_3) = 1/4$, hence $userNovelty(u_3, t{=}3{:}00, g) = 2$. Similarly, during this window, two images, one audio and one video were shared. In particular, *msg 12* contains only an image (i.e., $|\mathcal{C}_m| = 1$) . Thus, the novelty associated with the category of this message can be computed as: $\bar{P}^{\rightarrow}_{3:00,g}(\mathcal{C}_m = \{\text{“image”}\}) = 2/4$, which results in $catNovelty(\{\text{“image"}\}, t{=}3{:}00,\ g) = 1$. Thus, the curiosity driving user 3 to share *msg 12* is more stimulated by the novelty the user herself experiences with this action than by the novelty of (the category of) the content shared.

Note that both novelty related metrics focus on a single element: the destination user or the (categories of the) message shared. The aggregation of either metric for all elements in the window of interaction, using entropy, captures the *uncertainty*. Thus, we define metrics of uncertainty related to user and categories as:

$$userUncertainty(t{=}t_{i|d,g},g){=}-\sum_{d\in\mathcal{U}_g} P^{\rightarrow}_{t,g}(D{=}d)\log_2\Big(P^{\rightarrow}_{t,g}(D{=}d)\Big), \tag{5.15}$$

$$catUncertainty(t{=}t_{i|d,g},g){=}-\sum_{c\in\mathcal{C}^{\rightarrow}_{t,g}} P^{\rightarrow}_{t,g}(C{=}c)\log_2\Big(P^{\rightarrow}_{t,g}(C{=}c)\Big). \tag{5.16}$$

The idea behind these metrics is that the curiosity of the destination may be more/less stimulated by the greater/less diversity in the users sharing messages (and in the categories of these messages) during the window of interaction.

For the sake of illustration, let's compute the stimulus related to user uncertainty experienced by user 3 when sharing *msg 12*. We first calculate the probability of each user sharing a message in the current window of interaction, which is equal to 1/4 for all four users in the window. The user uncertainty is then computed as $userUncertainty(t{=}3{:}00, g){=}2$. Turning to the stimulus related to the uncertainty associated with message categories, we compute probabilities $P^{\rightarrow}_{3:00,g}(C{=}\text{“video"}) = 1/4$, $P^{\rightarrow}_{3:00,g}(C = \text{“audio"}) = 1/4$ and $P^{\rightarrow}_{3:00,g}(C = \text{“image"}) = 2/4$, which leads to an uncertainty equal to $catUncertainty(t{=}\ 3{:}00, g) = 1.5$. Thus, at time 3:00 pm, the diversity (uncertainty) in users sharing content is greater than the diversity in content category. Thus, user 3's curiosity is more stimulated by the former.

According to Berlyne [Berlyne, 1960], *conflict* occurs when the same stimulus triggers multiple incompatible responses, being positively related to the strengths of

the competing responses. To operationalize the concept of conflict, we follow the same approach in [Wu and Miao, 2013a]. The basic idea is that the different elements (categories/users) that appear in the window of interaction represent the potentially incompatible responses stimulating the curiosity of the destination user, and the strength of each response is captured by the probability of occurrence of each element (category/user). We instantiate this variable by first computing the average probability of users (categories) over all users (message categories) in the window of interaction, and then taking the surprisal of the result. This procedure leads to two new metrics, one related to users and the other to categories:

$$userConflict(t{=}t_{i|d,g},g){=}-\log_2\left(\frac{1}{|\mathcal{U}^{\rightarrow}_{t,g}|}\sum_{d\in\mathcal{U}_g}P^{\rightarrow}_{t,g}(D{=}d)\right), \tag{5.17}$$

$$catConflict(t{=}t_{i|d,g},g){=}-\log_2\left(\frac{1}{|\mathcal{C}^{\rightarrow}_{t,g}|}\sum_{c\in\mathcal{C}^{\rightarrow}_{t,g}}P^{\rightarrow}_{t,g}(C{=}c)\right). \tag{5.18}$$

Again, we use the example in Figure 3.3 to exemplify how the metrics of conflict are computed for user 3 sharing *msg 12*. We first take the average probabilities across all users and categories which are 1/4 and 1/3, respectively. We then compute $userConflict(t = 3{:}00,\ g){=}2$ and $catConflict(t = 3{:}00,\ g){=}1.59$.

Finally, an alternative form of capturing the diversity of message categories in the current window of interaction is by exploiting the unique occurrences of the media types. This metric, also based on surprisal, captures the *complexity* associated with the categories of messages shared during the current window of interaction, which is also a collative variable associated with curiosity stimulation. It is defined as:

$$catComplex(t{=}t_{i|d,g},\,g) = -\log_2\left(\frac{|\mathcal{C}^{\rightarrow}_{t,g}|}{|\mathcal{M}|}\right). \tag{5.19}$$

Note that, unlike uncertainty and conflict, which captures the diversity of message categories considering only those included in the window of interaction, complexity captures a somewhat different notion of diversity that takes into account all possible categories (included in set $\mathcal{M}$). Once again, we compute the complexity of message categories at time 3:00 when user 3 shared *msg 12* as follows. The current window of interaction includes 3 types of media (notably, "video", "audio" and "image") out of a total of 5 different types possible. Thus, the complexity associated with message categories is computed as $catComplex(t{=}3{:}00, g) = -\log_2\frac{3}{5} = 0.74$.

As a final remark, we note that the metrics defined in Equations 5.13–5.19 are always non-negative and should be interpreted as: the greater the value the higher the stimulus to the curiosity of the destination user. Having introduced the novel metrics of curiosity stimulation, we applied them to study curiosity stimulation behind content sharing in Whatsapp groups. We discuss the main results from our investigation next.

5.5 Curiosity Stimulation in WhatsApp Groups

In this section, we use the metrics introduced in the previous section to characterize user curiosity in WhatsApp groups. Our goal is to both show the applicability of the metrics and uncover behavioral traits in a widely used communication platform.

To that end, we rely on a dataset gathered by the WhatsApp Monitor [10] [Melo et al., 2019], a system for collecting shared messages in publicly accessible WhatsApp groups. The dataset consists of a sequence of messages posted in a number of publicly accessible groups in Brazil over a period of great social and political turmoil in the country (April 1st 2018 to April 30th 2019) which includes the 2018 general elections in the country. The monitored groups are themed around political topics.

Each entry in the dataset consists of posting time, anonymized user identifier, group identifier and message categories (media types)[11]. We note that even though WhatsApp currently allows users to reply to a particular message, thus creating an explicit link between messages, this feature was not available at the time our dataset was collected[12]. As such, any possible link must be inferred from the available data, notably from the time when the messages were shared, since message content is not accessible.

To avoid issues with data sparsity, for each group, we only consider users who shared at least 30 messages during the whole period, and we only compute the metrics for sharing events with at least 10 messages in the window of interaction[13]. These lower bounds were imposed so as to be able to compute reliable values for the metrics,

[10]WhatsApp Monitor is available online at `http://www.monitor-de-whatsapp.dcc.ufmg.br/`.

[11]All user identifiers are indeed anonymized phone numbers as we are not able to identify multiple phone numbers belonging to the same person. Moreover message content, though collected by WhatsApp Monitor, cannot be made publicly available and, as such, was disregarded.

[12]Yet, our metrics might still be used (with some adjustments) when replies are available. In that case, we argue that, even though the original message was the primary source of curiosity stimulation on the user sharing the reply, other messages shared recently could also compose the stimulus driving the user behavior, perhaps with a lower intensity. Thus, a weighting scheme could be employed to combine all signals, favoring the message being replied. Exploring such scheme is an interesting subject for future work.

[13]For the sake of research reproducibility, we make our anonymized data as well as our code publicly available at `https://zenodo.org/record/5790153#.Yb0Osb_MJE4`.

following the arguments in prior work [Voorhis and Morgan, 2007] that state that that samples with fewer than 10 elements often lead to estimations with low statistical power. We also ran some preliminary experiments with other bounds and based on observations from these experiments as well as our own prior experience with curiosity stimulation on LastFM [Sousa et al., 2019], we found that fewer data points often lead to unreliable measures. After applying these filters, the resulting dataset, which is used in our study, is composed on 2,054,302 messages posted by 7,584 distinct users in 335 groups.

5.5.1 Relationships among Collative Variables

We start by quantifying the relationships among different collative variables, namely social influence, novelty, uncertainty, conflict and complexity, as captured by the proposed metrics. We aim at assessing the extent to which different metrics, notably the four novel metrics related to (individual-level) social influence (Equations 5.3, 5.4, 5.7 and 5.8) are able to capture aspects related to curiosity stimulation which are not covered by the other (more traditional) variables (Equations 5.13–5.19). To that end, we classify the metrics into *redundant* or *complementary* as to whether their effect on curiosity stimulation is mostly captured by the others (thus being redundant) or not.

Specifically, we first compute, for each user u in a group g, all metrics (11 in total) for each message shared by u in g. We then use these values to compute the Spearman correlation coefficient between each pair of metrics, for each (u, g) pair. Next, as in [Sousa et al., 2019], we employ a heuristic to identify redundant metrics: two metrics are considered redundant for a given (u, g) pair if their correlation falls in the $(-1, -0.5)$ or $(+0.5, +1)$ ranges, which can be considered moderate-to-strong correlations. Otherwise, the correlation is deemed weak and the metrics are taken as *complementary*.

Figure 5.2 shows a matrix with the fractions of cases, i.e., (user, group) pairs, for which each pair of metrics was considered redundant. One key result is that the four social influence metrics, shown in the last four rows (and columns), are *complementary* when compared to the metrics related to the other collative variables for most (more than 77%) of the cases. Thus, these metrics indeed capture relevant aspects of curiosity stimulation which are not covered by traditionally studied metrics. We also observe that direct and indirect social influence do often capture distinct effects, as their related metrics cannot be considered redundant for a large fraction of the cases (67% for the metrics of maximum influence). Yet, given the same general effect (i.e., direct or indirect social influence), taking either the average or maximum across all origins of

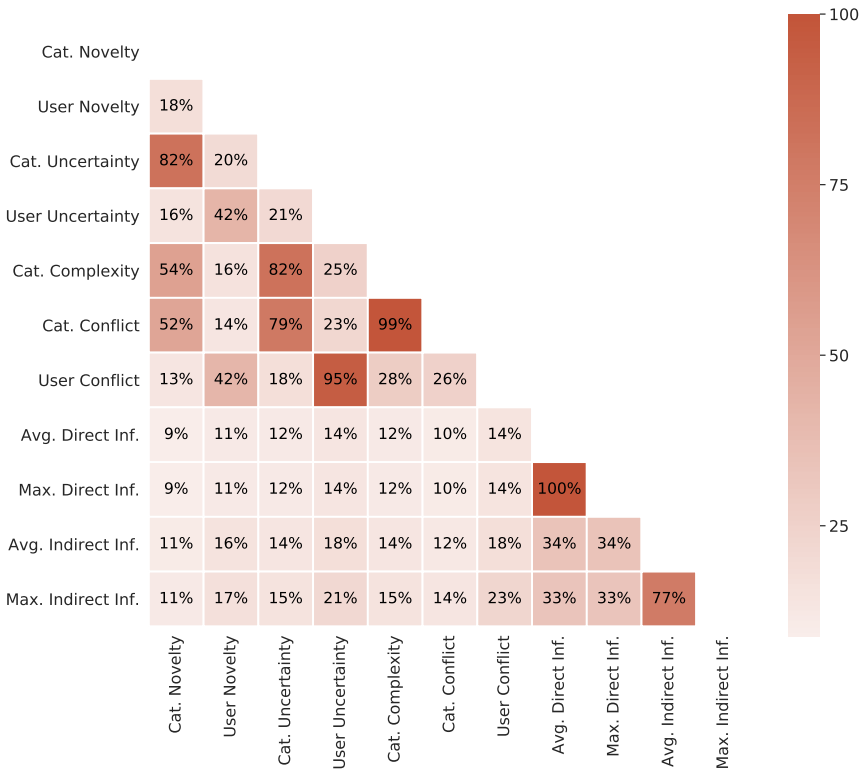


Figure 5.2: Fraction of redundant (user, group) cases for each pair of metrics.

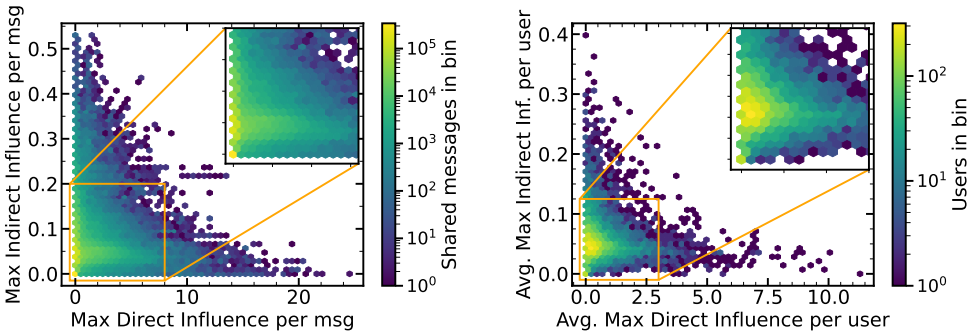
influence is quite similar, as the corresponding metrics are redundant with respect to each other in most cases.

We can also note great redundancy between several traditional metrics, such as between uncertainty and novelty and between uncertainty, conflict and complexity, all related to message categories, as well as between uncertainty and conflict related to users. These redundancies are consistent with those observed in [Sousa et al., 2019] for curiosity in online music consumption, and reflect the similar effects captured by those related metrics.

In the following, given our goal to study social influence as a component of curiosity stimulation, we focus our analyses on the two social influence metrics – maximum direct influence and maximum indirect influence – identified as complementary to each other in most cases.

5.5.2 Diversity and Dynamics of Social Curiosity

As a first analysis, we focus on how diverse user curiosity stimulation is in terms of the two complementary metrics of social curiosity. Recall that both metrics are computed



(a) Stimuli associated with messages.

(b) Average stimuli for users.

Figure 5.3: Diversity of social curiosity stimuli across message sharings and users.

for each *message sharing* event, performed by a user d in a group g at time $t_{i|d,g}$. Figure 5.3 shows results of the two metrics for individual messages (Figure 5.3a) as well as averages across all messages shared by the same user d (Figure 5.3b), both computed for specific groups (i.e., the same user in different groups appear separately). Each graph in the figure shows the number of elements (messages or users) with specific values of *maximum direct influence* (x-axis) and *maximum indirect influence* (y-axis). For both graphs, we zoomed in a specific region of the graph and inset it to the graph. We highlight that the regions zoomed in correspond to 97% and 92% of all messages and users, respectively. The graphs are meant to contrast the values of both social curiosity metrics for each message and user. As such, they also show the ranges of the values observed for both metrics considering individual messages as well as aggregated across all messages shared by the same user.

The graphs reveal great diversity in values of both metrics across messages and users. On one hand, there is a major concentration in smaller values for both metrics, highlighted in the insets. Smaller values suggest that, based on the historical patterns, the behavior of the destination user (in terms of content sharing actions) is only weakly influenced (if influenced at all) by how the others in the current window of interaction (origins) behave. That is, the small amount of information (in terms of bits) about the interactions between these users suggest weaker social influence. In contrast, a fraction of the messages as well as a fraction of the users exhibit very large values of both metrics (two to four times larger than the majority), suggesting that, in those cases, the behavior of the destination user is strongly influenced by past behaviors of the origins. As a consequence, this produces more information (in terms of bits) about the interactions between those users, offering stronger evidence of social influence between them. Also, as the axes of both graphs show, direct influence is clearly spread over

Table 5.4: Message-level social curiosity stimulation profiles.

Cluster	$\%msgs$	**Max. Dir. Influence** (average ±95% C.I.)	**Max. Ind. Influence** (average ±95% C.I.)
Independent	72.6%	0.26 ± 0.000750	0.04 ± 0.000044
Indirect	14.4%	0.43 ± 0.002320	0.15 ± 0.000210
Dependent	13.0%	3.52 ± 0.006623	0.05 ± 0.000099

larger values, as a result of how the metric is computed (i.e., based on pointwise mutual information).

Moreover, by comparing the spreading of values (especially in the x-axis) on both graphs, we can infer that the diversity is great even if we look at the same user over time, i.e., over all the messages the user shared. In other words, the (social) curiosity of a user is stimulated quite differently for different users and even for the same user, considering different messages shared by her (potentially in different groups), which is consistent with some of our key assumptions presented in Section 3.2 of Chapter 3.

Given these observations, we delve deeper into the analysis of the diversity and dynamics of social curiosity by considering three levels of aggregation. First, we focus on individual messages shared by users in the groups (Section 5.5.2.1). Next, we consider all messages shared by the same user in a given group (Section 5.5.2.2). Finally, we look at social curiosity at the group level (Section 5.5.2.3). We discuss our results next.

5.5.2.1 Social Curiosity at the Message Level

Aiming at identifying common patterns of social stimulation driving the sharing of a message, we clustered the message sharing events based on the values of both metrics – $maxDirInf$ and $maxIndInf$. To that end, we employed the Mini-Batch K-means algorithm [Sculley, 2010], a widely used parallel version of K-means adequate for large datasets, to cluster the values of curiosity stimulus (computed by the two metrics) associated with the messages shared by all users (over 2 million messages in total). We varied the number of clusters k from 2 to 18 and, for each value, we repeated each experiment 10 times and computed the Silhouette index [Amorim and Hennig, 2015] of the clustered results. This is a measure of how similar an element is to its own cluster (cohesion) compared to other clusters (separation). Larger values indicate better matching of elements into clusters. Using the results of Silhouette index, we set $k = 3$ as the best number of clusters, as it led to the highest Silhouette index among all values analyzed (equal to 0.55, which suggests that a reasonable clustering structure was obtained [Kaufman and Rousseeuw, 2005]).

Table 5.4 shows a description of the three identified clusters in terms of average values (and associated 95% confidence intervals) of both metrics as well as fraction of messages in each cluster. Note that the confidence intervals are quite tight, implying that the average values of both metrics are very representative of the elements in each cluster. By comparing these values for each metric[14], we label the clusters as:

- (Socially) Independent: both metrics tend to assume the smallest values among the three clusters, implying less frequent co-occurrence of destination and origins, and suggesting that social influence plays a less important role in curiosity stimulation. This occurs in 73% of the messages;

- Indirect (Social Influence): compared to the other clusters, the maximum direct social influence tends to assume moderate values, whereas the maximum indirect influence assumes the largest values of the three clusters. Thus, in comparison to the other two clusters, indirect social influence seems to have a particularly important role on the curiosity stimulation in this case. This occurs in 14% of the messages;

- (Socially) Dependent: this cluster exhibits, by far, the largest values of direct social influence, suggesting more frequent co-occurrences of destination and origins, and thus much stronger social stimuli as captured by this metric. This happens in 13% of the messages.

5.5.2.2 Social Curiosity at the User Level

Having identified patterns of social stimulation associated with individual messages, we now use them to cluster users in terms of their social stimulation profiles. Given our assumption, supported by the results in Figure 5.3b, that social curiosity stimulation may vary across time for the same user, our goal is to build a user profile representation that reflects such dynamics using the patterns of social stimulation at specific messages as building blocks. We do so by first representing each user u by the sequence of social stimulation patterns (independent, indirect or dependent) associated with the sequence of messages shared by u. We then model such sequence using a User Behavior Model Graph (UBMG) [Menasce and Almeida, 2000]. A UBMG is a graph where each node represents a pattern of social stimulation (state) and edges denote the transitions between patterns from a message to the next one, by the same user. Edge weights are

[14]Note that we do not compare values of different metrics as they cover very different intervals. Rather, we analyze differences across the three clusters by looking at the values of each metric at a time.

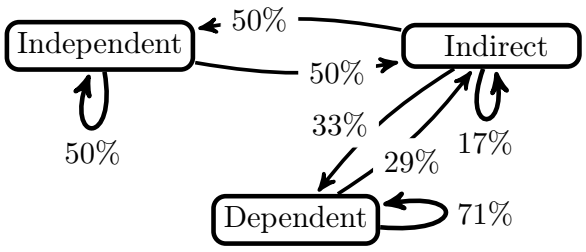
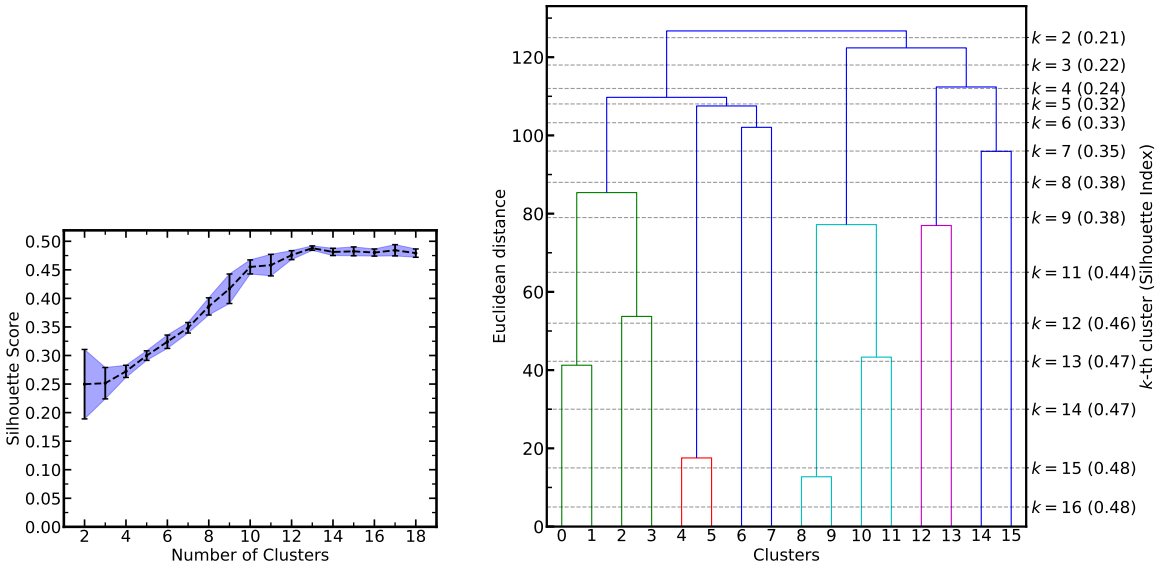


Figure 5.4: Example UBMG for a user u with a sequence of message-level stimulation pattern equal to `PPIDDDIPPIPIIDDDDIPP` in a group g (`P` for independent, `I` for indirect and `D` for dependent).

the probabilities of such transitions. One UBMG is used to model a user in *a given group*. Thus, the same user in different groups are represented by multiple (possibly distinct) UBMGs, one for each (user, group) pair.

We illustrate the UBMG representation by considering a fictitious user u who shared 20 messages in a group g. Given the social stimulation pattern associated with each message shared by u in g – `P` for independent, `I` for indirect and `D` for dependent – we first build a representation of (u, g) as a temporally-ordered sequence of stimulation patterns. Let's say such representation is `PPIDDDIPPIPIIDDDDIPP`. That is, u's curiosity stimulation in g is associated with the independent (`P`) pattern in the first two messages, then changes to the indirect (`I`) pattern in the third message, changing once again to the dependent (`D`) pattern, and so on. The corresponding UBMG is shown in Figure 5.4. Note that once in the *independent* state, user u has 50% of chance of remaining in this state in the next message and 50% of chance of changing to *indirect*. Once in the *indirect* state, u either goes back to the *independent* state, with 50% of chance, changes to *dependent*, with 33% of chance, or remains in the same state (17% of chance). Having built the UBMGs for all (user, group) pairs, we then clustered these UBMGs to uncover patterns of user curiosity stimulation. To that end, we employed once again the Mini-Batch K-Means algorithm, using the Silhouette index to select the best number k of clusters. Figure 5.5a shows the average Silhouette index (computed for 10 runs) along with 95% confidence intervals as a function of k. Note that the average Silhouette increases greatly with k until reaching a rough plateau for $k \geq 13$. Also, the wider confidence intervals for small values k indicate higher variability of the results. Yet, the intervals become quite tight for larger values of k.

Based solely on the Silhouette index results, one would choose a value of k around 13 (or just above this mark) as the number of clusters. However, such large number of clusters makes it difficult to identify distinguishing characteristics. Indeed, by manually analyzing the centroids of these clusters, we found great similarities among many



(a) Silhouette index. (b) Dendrogram

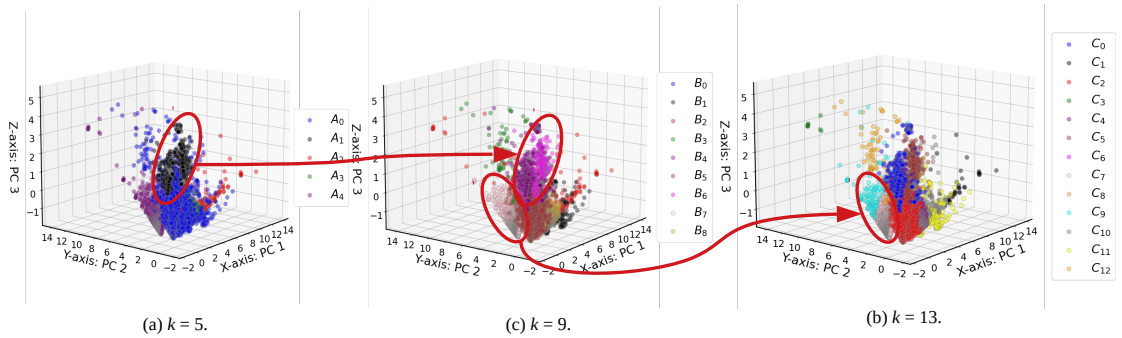
Figure 5.5: Determining the number of UBMG clusters.

of them. Thus, in the interest of obtaining a more interpretable set of UBMG profiles, we applied the Agglomerative Hierarchical clustering algorithm [Ah-Pine, 2018], which seeks to build a hierarchy of clusters. The method starts with n clusters and sequentially combines pairs of similar clusters (the most similar ones first) until a single cluster is obtained. We applied the algorithm starting with $n = 16$ clusters[15]. Figure 5.5b exhibits the dendrogram for these clusters where the sixteen labels are shown along the x-axis. Pairs of similar clusters are linked, and the height of the line connecting two clusters represents the Euclidean distance between them, shown in the y-axis. The right side of the figure shows decreasing values of k (from 16 down to 2), as the most similar clusters are merged together, along with the corresponding Silhouette index (within parentheses).

Based on these results, along with a manual analysis of the centroids of all 16 initial clusters, we chose to set k equal to 5 clusters, as it delivers the best trade-off between clustering quality (i.e., Silhouette index) and interpretability of the properties characterizing each identified cluster[16]. On one side, the Silhouette index for $k = 5$ is 0.32, which, despite lower, is still indicative of an existing clustering structure [Kaufman and Rousseeuw, 2005]. On the other side, we observed that larger values of k led to mostly variants of the same clusters identified with $k = 5$.

[15]This starting point was selected as it is one of the values of k in the plateau of maximum Silhouette index shown in Figure 5.5a.

[16]We found that restricting to 5 clusters facilitates interpretation of the results as the identified



(a) $k = 5$. (c) $k = 9$. (b) $k = 13$.

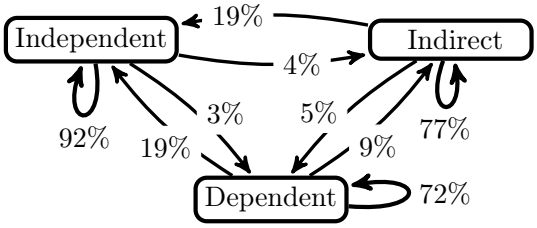
Figure 5.6: Results from Principal Component Analysis applied to $k = 5$, 9 and 13 clusters.

We illustrate the latter point by applying Principal Component Analysis (PCA) to identify the three most important components for three different values of k, notably $k = 5$, $k = 9$ and $k = 13$. Results are shown in Figure 5.6. By comparing Figures 5.6a and 5.6b we can see that increasing the number of clusters from $k = 5$ to $k = 9$ simply led to splitting some of the clusters into multiple ones, but those are quite similar to each other. Note for example cluster A_1, identified in black in Figure 5.6a. It is basically split into two closely related variants, clusters B_4 and B_6 shown in pink and purple, respectively, in Figure 5.6b. Similarly, by comparing results for $k = 9$ and $k = 13$, we note that cluster B_7 (in light pink in Figure 5.6b) is mostly split into two variants, clusters C_7 and C_9 (cyan and light pink, respectively) in Figure 5.6c.

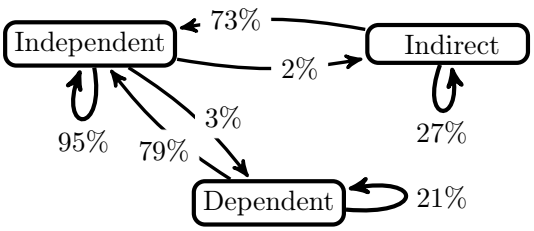
Figure 5.7 shows the five user-level profiles, referred to as $U_0 ... U_4$, identified by the centroids of the five UBMG clusters obtained. For each profile, the figure also shows the number of (user, group) pairs in the cluster, and the average number of messages per such pair. To help interpreting the UBMGs, Table 5.5 shows, for each profile, the average fraction of all messages sent by the user in the given group characterized by each message-level curiosity stimulation profile.

As shown in Figure 5.7a, profile U_0, including almost 30% of all (user, group) pairs in our dataset, is characterized by a general trend of repeatedly staying in the same state of social stimulation (strong self-transitions). In other words, users in this profile tend to experience a rather stable curiosity stimulation process, being consistently driven by one of the three patterns over time, with occasional changes to a different pattern. In particular, as shown in Table 5.5, this user profile is dominated by the *independent* state, which characterizes over 65% of the messages, on average. In other words, users with this profile, who are the most active ones in terms of the average number of messages shared, tend to be mostly insensitive to social influence around

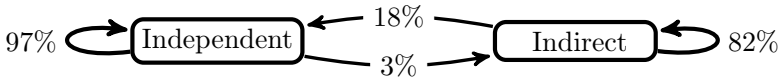
clusters are more clearly distinct and refer to different patterns.



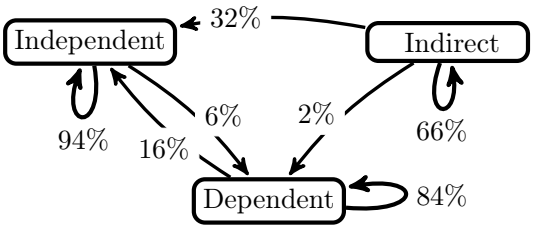
(a) U_0: 2,476 elements (383 msgs/user).



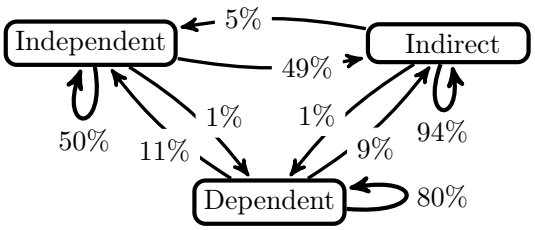
(b) U_1: 889 elements (90 msgs/user).



(c) U_2: 2,550 elements (156 msgs/user).



(d) U_3: 2,156 elements (283 msgs/user).



(e) U_4: 235 elements (67 msgs/user).

Figure 5.7: User-level social curiosity stimulation profiles: state transition diagrams (UBMGs), each element is a (user, group) pair.

two thirds of the time. Yet, social influence does take a more important role as a component of curiosity stimulation, either directly or indirectly, in roughly 35% of the messages shared by those users.

Table 5.5: User-level social curiosity stimulation profiles: percentage of messages in each UBMG state.

UBMG	**User-level profiles**				
state	U_0	U_1	U_2	U_3	U_4
Independent	65.73%	87.72%	80.17%	70.25%	8.66%
Indirect	20.88%	6.23%	19.80%	0.20%	84.49%
Dependent	13.39%	6.06%	–	29.54%	6.85%

Users in profile U_1, shown in Figure 5.7b, may also experience the three states, but changes towards the independent state are much more frequent, which, in turn, tends to dominate the overall profile. In other words, compared to U_0, users in U_1 tend to have less diversity in terms of social curiosity stimulation patterns, suffering little impact from social influence most of the time (87% of the messages). They also tend to be much less active than users in U_0. This cluster includes 11% of all (user, group) pairs.

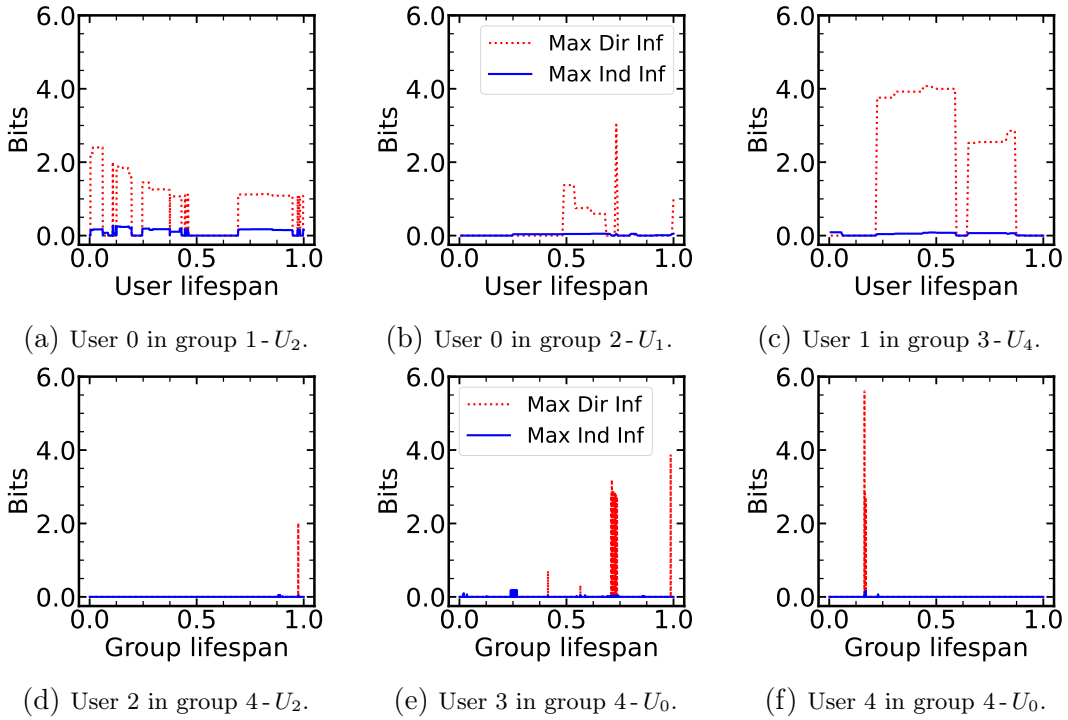
Profile U_2, shown in Figure 5.7c, is characterized by a mixture of independent and indirect states, though being also greatly dominated by the former (80% of the messages, as shown in Table 5.5). Unlike users in the the other clusters, users in U_2 do not experience the dependent state. In other words, the curiosity stimulation of these users is mostly insensitive to *direct* social influence. Moreover, the strong self-transitions imply that users often remain repeatedly in the same state of curiosity stimulation. Around 31% of the (user, group) pairs in our dataset fall into this cluster.

Profile U_3, shown in Figure 5.7d, exhibits a distinguishing characteristic compared to the others. There are no transitions into the indirect state from the other states (only the self-transition). Users in this cluster either start in the indirect state and possibly move towards the other states or never experience it. Compared to users in U_0, U_1 and U_2, users in U_3 tend to fall in the dependent state more often (in roughly 30% of the messages), being thus more sensitive to direct social influence as a driver to curiosity stimulation. These users tend to be very active, falling behind only users in U_0. This cluster includes almost one third of all (user, group) pairs (26%).

Finally profile U_4, shown in Figure 5.7e, shows the distinguishing characteristic of much stronger transitions (including self-transition) to the indirect state, leading to profiles that are dominated by this state, which characterizes 85% of the messages, on average. The independent state, in turn, is much less prevalent compared to the other profiles, as can be noted by the much lower self-transition probability. Indeed, as shown in Table 5.5, the independent state happens in only 8% of the messages shared by users with this profile, on average. These are users whose curiosity stimulation tends to be often influenced by social factors but by *indirect* means, possibly because the lack of frequent activity prevents the formation of strong social ties that could influence user curiosity. Indeed users in U_4, which account for only 3% of all (user, group) pairs in our dataset, are the least active ones in terms of message sharing, on average.

The results in Figure 5.7 show great diversity in social curiosity stimulation profiles across different users and, for many users, very dynamic profiles. To further illustrate this point, we show in Figure 5.8 the time series of both social curiosity metrics – $maxDirInf$ e $maxIndInf$ – for six selected (user, group) pairs. These curves illustrate how the social curiosity of particular users in specific groups evolve over time, as captured by both metrics. To facilitate visualization, the x-axis of Figures 5.8a–5.8f represent the normalized *lifespan* of the particular user in the given group, that is, the time interval between the first and last message shared by this user in the group, normalized to fall between 0 and 1.

Overall, we can see some important fluctuations over time, notably in terms of



(a) User 0 in group 1-U_2. (b) User 0 in group 2-U_1. (c) User 1 in group 3-U_4.

(d) User 2 in group 4-U_2. (e) User 3 in group 4-U_0. (f) User 4 in group 4-U_0.

Figure 5.8: Time series of both metrics of social curiosity for different (user,group) pairs.

the maximum *direct* influence[17]. Such fluctuations follow the natural dynamics of user participation in the group. Figures 5.8a and 5.8b show the time series of the social curiosity of the same user (identified as user 0) in two different groups she participates in (identified as groups 1 and 2). We note very different patterns, temporally and on average. Indeed, the dynamics captured by these two sets of curves fall into two very different profiles, identified by UBMGs 2 and 1, respectively. This observation hints at the validity of one of our key assumptions, that is, that user curiosity stimulation varies depending on the group. Figure 5.8c shows the time series of another user in a different group. As illustrated in the figure, this user's social curiosity fluctuates greatly over time, with alternating behavior of greater and lower stimulation.

Figures 5.8d–5.8f shows the time series of three different users, members of the same group. For that reason, the x-axes of these three graphs represent the normalized lifespan of the given *group* (i.e., identified as group 4), that is the time interval between the first and the last message shared by any user in the group, normalized to fall between 0 and 1. This group was selected as one among those with the largest number of distinct users during the monitoring period. Once again, we see that user social

[17]As already shown in Figure 5.3b, the values of $maxDirInf$ tend to be much larger than those of $maxIndInf$.

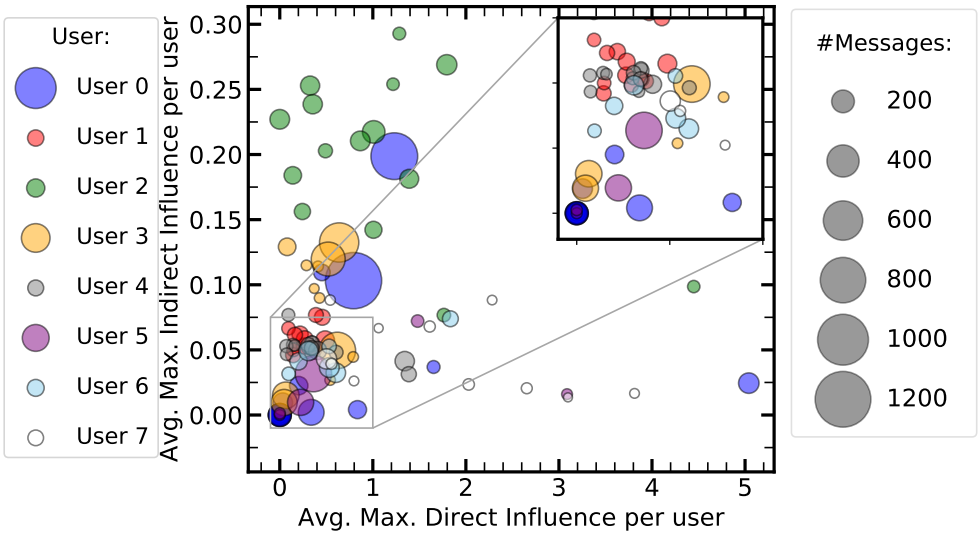


Figure 5.9: Diversity of social stimulation of selected users in different groups: the same user in different groups is represented by the same color.

curiosity fluctuates greatly over time (notably user 3, shown in Figure 5.8e). Most importantly, we see very different patterns across the three users, even though they are members of the same group. This observation hints at the idea that the impact of the group on the curiosity of individual members may be different for distinct users. We further elaborate on this point in the following section, as we present our key results on group-level social curiosity stimulation.

5.5.2.3 Social Curiosity at the Group Level

To analyze social curiosity at the group level, we focus on two aspects, notably: (1) the role of the group on the curiosity stimulation of its members and (2) the overall social curiosity of each group. For the latter, we use the group-level social curiosity metrics defined in Section 5.4.

To investigate the role of the group on the curiosity of its members, we selected the 8 users who shared the largest number of messages in multiple groups, and computed the *average* values of both social influence metrics for each (user, group) pair. Figure 5.9 shows the results as a scatter plot where each circle is a (user, group) pair. The same user (in different groups) is shown in the same color and the circle's diameter represents the number of messages shared by the user (in the given group). Visually, there are great distinctions on both metrics (in terms of average values) for the same user depending on the group she is participating in.

Such distinctions were confirmed by pairwise statistical tests of difference of aver-

Table 5.6: Test of statistical difference of social stimulation metrics of the same user on different groups (users 0–7 refer to selected users in Figure 5.3b).

User	Total # Msgs	# Groups	# Pairwise Comparisons	# (%) Pairs of Groups with Different Metrics	
				Max. Dir. Influence	Max. Ind. Influence
0	4,361	17	136	71 (52.21%)	70 (51.47%)
1	1,210	13	78	15 (19.23%)	11 (14.10%)
2	1,736	15	105	75 (71.43%)	75 (71.43%)
3	2,459	12	66	13 (19.70%)	25 (37.88%)
4	1,168	15	105	32 (30.48%)	31 (29.52%)
5	976	6	15	4 (26.67%)	5 (33.33%)
6	803	7	21	5 (23.81%)	11 (52.38%)
7	558	11	55	14 (25.45%)	3 (5.45%)

ages for each of the two metrics. Specifically, for each user u, we used a Kruskall-Wallis Honestly Significant Difference (HSD) test [Ott and Longnecker, 2015] (with α=0.05) to test for the difference between the averages of maximum direct influence for pairs of groups u participates in. We did the same for the maximum indirect influence. The results are reported in Table 5.6, which shows, for each selected user u, the total number of messages shared by u, the number of groups u participates in, the total number of pairs tested and the number (and percentage) of pairs for which there is a statistical difference in each metric (two rightmost columns). As shown, the differences are statistically significant for a large fraction of the cases for several users. Thus indeed the social stimulation of the same user may be quite distinct depending on the group she participating in, as already suggested by the results in Figures 5.8a and 5.8b. Indeed, we note that user 0 in Table 5.6 refers to the same user 0 shown in those two figures.

As a final analysis, we use the group-level social curiosity stimulation metric, defined in Equation 5.12, to characterize the groups in our dataset. Recall that this metric captures a reduction in uncertainty associated with destinations, defined in Equation 5.11, due to the knowledge of the social influence from other users (origins) in the group. Thus, to facilitate analysis, Figure 5.10 shows a scatter plot with the values of these two metrics – group mutual information ($groupMutInf$) and destination entropy – computed for each of the 335 groups at the time of the last message sharing in each group.

The figure shows great diversity in social curiosity also at the group level. Many groups have a value of $groupMutInf$ (x-axis) very close to the destination entropy $H^{\leftarrow}_{t,g}(D)$ (y-axis), suggesting that a large fraction of the total uncertainty associated with the destinations in those groups can be captured by social influence. In other words, social influence is a key component of curiosity stimulation in the group. This

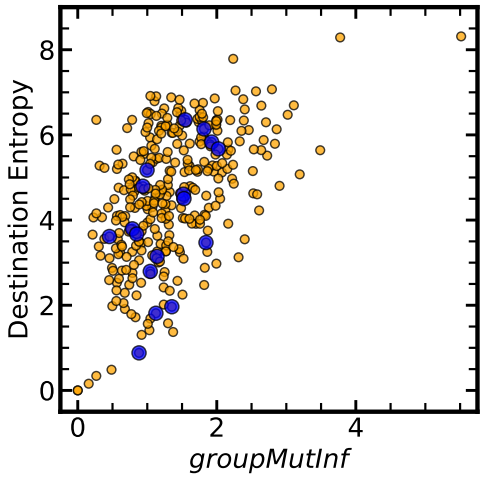
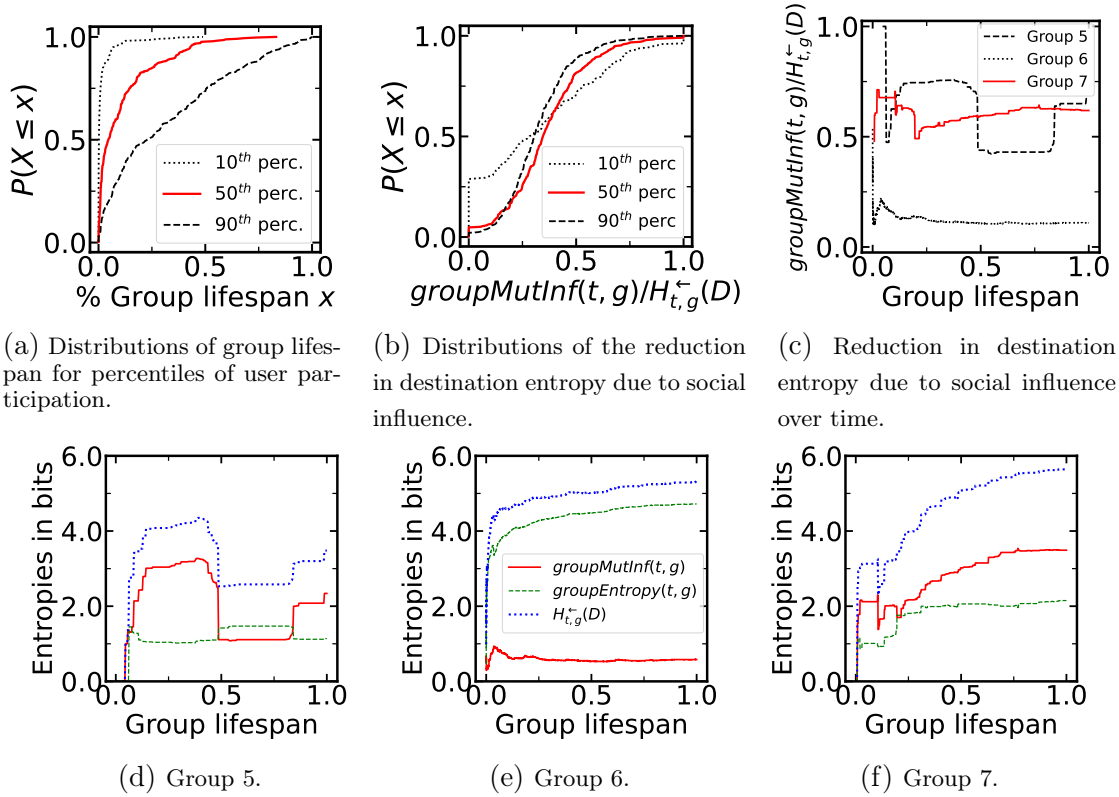


Figure 5.10: Social curiosity stimulation across groups: group mutual information versus destination entropy, measured at the the time of the last message in the group (groups of user 0 in blue).

happens for groups with distinct profiles in terms of number of users and level of activity of these users, which ultimately lead to groups with distinct values of destination entropy. On the other hand, a number of groups have values of $groupMutInf$ much smaller than the total uncertainty associated with the destinations. In those cases, social influence plays a less important role on curiosity stimulation at the group level.

For illustration purposes, Figure 5.10 also highlights (in blue) the groups in which one selected user – user 0 in Figure 5.9 – participates. As shown, the diversity in user 0's social stimulation shown in Figure 5.9 follows the diversity present also in his groups. We measured the correlation between group-level social curiosity and individual-level social curiosity (as captured by both metrics considered) for the selected 8 users in a few selected time windows but found not clear pattern: some cases of positive correlation, several cases of negative correlation and some cases of no-correlation at all. Thus, even though user curiosity stimulation does indeed vary depending on the group, we found no consistent indication that a more socially stimulated group always translates into the same behavior on all its members. Indeed, as shown in Figures 5.8d–5.8f, different members of the same group may exhibit quite distinct patterns of social curiosity. More broadly, it might be the case that different users dominate the overall group-level curiosity stimulation pattern at different time windows.

Recall that Figure 5.10 shows results for a specific time window, that is, the time of the last message sharing in each group. We finish our discussion by showing the dynamics of the group-level social curiosity metric over time. Naturally, one key component of social curiosity at the group level is user participation in the group.



(a) Distributions of group lifespan for percentiles of user participation.

(b) Distributions of the reduction in destination entropy due to social influence.

(c) Reduction in destination entropy due to social influence over time.

(d) Group 5.

(e) Group 6.

(f) Group 7.

Figure 5.11: Group-Level Social Curiosity Stimulation: CDFs of (a) group lifespan and (b) reduction in destination entropy due to social influence, measured when group reaches 10^{th}, 50^{th} and 90^{th} percentiles of user participation; (c) Time series of reduction in destination entropy due to social influence of 3 example groups; (d)–(f) Time series of group-level metrics for the same 3 groups.

Thus, we analyze social curiosity in each group in different points in time defined based on user participation. Specifically, for each group g, we looked at the total number of *distinct users* who shared some message in g *since the first message until the last message shared during the monitoring period*, and considered the time when g reached the i^{th} percentile of user participation[18]. We then looked at the lifespan of each group, and measured the time when each group reached the i^{th} percentile of user participation, as a fraction of its total lifespan.

Figure 5.11a shows the Cumulative Distribution Functions (CDFs) of the fractions of group lifespans associated with the 10^{th}, 50^{th} and 90^{th} percentiles of user participation in each group. Whereas some groups reach all three percentiles quickly, others take much longer to attract greater diversity in user participation, reaching the 90^{th} and even the 50^{th} percentiles very later on during monitoring. For instance, roughly

[18]Obviously, group activity prior to the monitoring period is disregarded. Thus, the analysis reflects group activity during this period only.

40% of the groups take more than 50% of their lifespans to reach the 90^{th} percentile of user participation. In light of this dynamics and heterogeneity, Figure 5.11b shows the CDFs of the reduction in destination entropy due to social influence (defined as the ratio $groupMutInf(t,g)/H^{\leftarrow}_{t,g}(D)$), measured at the same points in time (same three percentiles). Recall that the greater the reduction in destination entropy the more important the role of social influence on curiosity stimulation at the group level. Once again, there is great diversity across groups, here shown over time. Roughly 30% of the groups have reductions of at least 50% at the time they reach the 10^{th} percentile of user participation. In contrast, around 70% of the groups have the same amount of reduction (or less) by the time of the 50^{th} percentile.

Moreover, groups may experience quite dynamic social curiosity stimulation, evolving over time according to different patterns for distinct groups, as an aggregation of the individual patterns of the current group members (which, as discussed in the previous section, is also quite dynamic and heterogeneous). This is illustrated in Figure 5.11c, which shows the time series of the reduction in destination entropy due to social influence for three selected groups. Such patterns illustrate the very dynamic role of social influence as a driving force behind curiosity stimulation in the group: in some cases, social influence starts as an important component of group-level curiosity and continuously decreases over time (e.g., group 6 in the figure); while in other cases, the importance of social influence fluctuates over time (e.g., group 5), possibly reaching rough stability (e.g., group 7).

In order to explain such varied patterns, we plot in Figures 5.11d–5.11f the three metrics associated with group social curiosity, namely the $groupMutInf$ (red line), $groupEntropy$ (green line) and destination entropy, or $H^{\leftarrow}_{t,g}(D)$ (blue line) for the same three selected groups. For the groups shown in Figures 5.11d and 5.11f, social influence plays an important role on the overall group curiosity over large periods of time, as the $groupMutInf$ curves represents a large fraction of the total destination entropy (over 60%, on average). For both groups, such importance decreases mostly when the $groupEntropy$ increases, that is, when the probabilities of the same *destination* conditioned on the same *origins* decrease. In that case, the chance of the same users co-occurring in the same window decreases. Thus, social influence loses relevance as a curiosity driver for the group. When this happens at the same time as an overall drop in destination entropy, possibly due to an increase in the number of distinct users sharing messages, the role of social influence on curiosity at the group level becomes even less important (e.g., group 5 at around 50% of its lifespan). In contrast, $groupEntropy$ is very high, dominating the total destination entropy for group 6 throughout its lifespan, as shown in Figure 5.11e. Clearly, the curiosity of this group, in general, is very weakly

Table 5.7: Three example messages shared by the most active user in group "Science, Religion and Politics".

Msg Id	Curiosity Profile	Max. Dir Influence	Max. Ind. Influence	Window of Interaction # messages	Window of Interaction # users
1	Dependent	6.78	0.03	28	6
2	Indirect	0.50	0.10	61	7
3	Independent	1.50	0.08	90	8

impacted by social influence, as shown in Figure 5.11c.

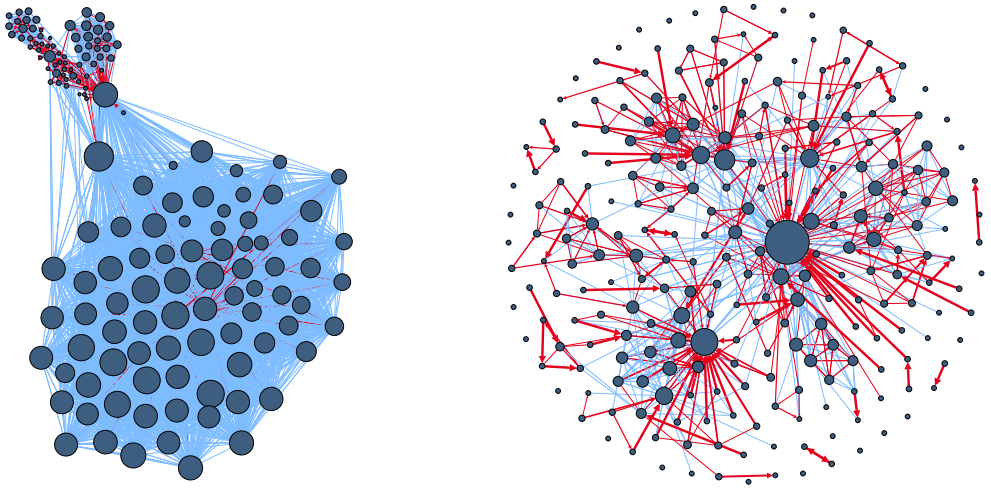
5.5.2.4 Some Illustrative Examples

In this section we discuss a few examples to further illustrate the use of our metrics. We start by looking at the participation of a particular user, referred to as user u, in a group entitled "Science, Religion and Politics" (translated to English), which is one of the groups with the largest number of distinct members (353 users during the complete monitoring period) and messages shared (48,389). User u is the most active user in this group, having shared a total of 3,787 messages. Also, u's curiosity stimulation follows profile U_0 which, as shown in Table 5.5, experiences all three states of message-level curiosity stimulation. Thus, we selected three messages shared by u, referred to by identifiers 1, 2 and 3, associated with each profile – dependent, indirect and independent, respectively.

Table 5.7 presents information associated with each selected message, notably its curiosity stimulation profile, the values of two social curiosity metrics (maximum direct and maximum indirect influences) as well as statistics of the window of interaction defined when the given message was shared. The latter includes the number of distinct users (i.e., origins of influence) and number of messages shared during the period.

We note that the number of potential sources of social influence on u when she shared message 1, i.e., the number of origins of influence, is the smallest among the three messages (6). The same holds for the total number of messages shared during the window of interaction (28). Yet, this message has the largest value of maximum direct influence (6.78), much larger than the other messages, and the smallest value of maximum indirect influence (0.03). Thus, despite in smaller number, the past behavior of these origins and of the destination u offer strong evidence of *direct* social influence driving u's content sharing decisions. This message is characterized by the *dependent* profile.

In contrast, when sharing message 2, user u's curiosity is not strongly stimulated



(a) Group: "We are all HADDAD!". (b) Group: "Lula the father of prisoners".

Figure 5.12: Graph representation of direct social influence among members of two selected groups (node diameters are proportional to out-degrees, red/blue edges refer to strong/weak social influence, thus strong/weak social curiosity stimulation.

by those sharing content during the window of interaction, at least not with respect to the stimulus triggered by direct influence estimated by prior interactions. In other words, the maximum direct influence is very small (only 0.5). Yet, compared to message 1, we do find more evidence of indirect influence from the origins stimulating u's curiosity at this time: the maximum indirect influence has the largest value of the three messages, and the message is characterized by the *indirect* profile. Indeed, some of these origins are among the top 10 most active users in the group. Also, we tracked these origins over successive windows, observing an increase in the *direct* influence from some of them on u, which suggests that, as mentioned in Section 5.4.2, sources of indirect influence may become stronger sources of direct influence as time progresses.

Finally, when sharing message 3, user u's curiosity is not strongly stimulated by neither direct nor indirect influence, despite the presence of large numbers of messages and users in the corresponding window of interaction (the largest ones out of the three messages). The message is characterized by an *independent* profile. Clearly the more intense user activity during the window does not translate into stronger curiosity stimulation on u: both social curiosity metrics have intermediate values.

To illustrate social curiosity at the group level, we build a graph representation of social influence based on the contingency table, which contains historical patterns (see definition in Section 5.4.2). In this graph representation, each node represents a group member and a direct edge from user u to user v is added to represent that u

(origin) has influence on v (destination). Edge weights are given by the conditioned probability of a destination given an origin, i.e., $P^{\leftarrow}_{t,g}(D{=}d \,|\, O{=}o)$, computed based on values in the contingency table.

Figure 5.12 shows the graphs built for two specific groups considering the complete history of user interactions in each group, that is, by the time the last message is shared in the group. In each graph, node diameters are proportional to out-degrees and edge colors denote the strength of the conditioned probabilities used as weights (and thus the strength of the social influence computed from such probabilities). Specifically, red edges are those whose weights are above 0.1, thus representing stronger social curiosity stimulation, and blue edges have weights below 0.1, suggesting weaker social curiosity stimulation.

Figure 5.12a represents a group entitled "We are all HADDAD!", which had 147 distinct members who shared 3,012 messages during the monitoring period. Fernando Haddad was a left-wing presidential candidate in the 2018 general elections in Brazil. Approximately 95% of the edges in the graph are blue, suggesting that social influence does not play an important role in the group. Indeed, the overall reduction of destination entropy due to social influence – $groupMutInf(t,g)/H^{\leftarrow}_{t,g}(D)$ – is below 6%. This happens despite the great number of edges connecting users, that is, there is some evidence of social influence, but in the vast majority of cases this evidence is very weak. Indeed, the authors are aware that most members of this group often shared news about the candidate's campaign, which did not generate much response or engagement from most members. The occasional user interactions resulting from such messages offer only weak evidence of influence (if any at all). However, there are a few cases of more responsive and intense discussions on specific political themes, often with the participation of particular users. These cases show up in the few red edges in the top part of the graph. Note that many of them are linked to a particular node, possibly a user who is frequently involved in the discussions.

Figure 5.12b, in turn, shows the graph representation for the group entitled "Lula the father of prisoners", which had 221 distinct members and 935 messages shared during the monitoring period. Lula is a former president of Brazil and a charismatic representative of the left wing who was arrested during the period of the 2018 general elections. Unlike in Figure 5.12a, this graph is very balanced in terms of the strength of social influence among users: 51% (49%) of the edges are blue (red). Overall, social influence is a much stronger component of curiosity stimulation, with a reduction of destination entropy due to social influence of 62%. This group is mostly driven by very passionate pro-right wing discussions, thus the group's title is a sort sarcasm or irony. As reflected in the graph, such discussions, often involving the same participants, do

offer strong evidence of curiosity stimulation as a driver behind content sharing. Note also that, compared to the graph in Figure 5.12a, the graph in Figure 5.12b is much more sparse, suggesting that such strong curiosity stimulation occurs between specific pairs of users for whom the bond created by prior interactions offer greater evidence of social influence.

5.6 Limitations of our Study and their Implications

In the process of modeling such a highly subjective and complex concept such as social curiosity, we made a number of assumptions and simplifications so as to be able to derive metrics that are reasonably simple while still useful and informative. In the following, we discuss some of these limitations and their implications.

5.6.1 Lack of Content Representation

Message content is most probably an important factor influencing how one's curiosity is stimulated. Specifically, the contents of messages posted by others may indeed be a component of how *social* curiosity affects one's behavior. However, we chose not to exploit it in the derivation of our *social* curiosity metrics, focusing only on the primary factors, which, we argue, are related to who is sharing content in the group,i.e., users and their (inferred) social links. There are several reasons to do so:

1. Access to message content is not always available. Thus, by not using message content we assure that our metrics of social curiosity can be used even when content is not accessible (as is the case of our dataset).

2. As mentioned, by focusing only on the users sharing content, we aim at addressing a novel aspect of curiosity stimulation that is not captured by the other (traditional) collative variables. Thus, we complement these variables by proposing metrics related to a novel aspect.

3. By not including content in the social curiosity metrics, we are able to assess the extent to which the most primary component of social influence (i.e., people themselves) can drive one's curiosity.

We note however that, even though the content factor is not explicitly included in the proposed metrics, these metrics do exploit prior user interactions, which in turn, may have been influenced by the content shared by those users at those prior moments

in time. Thus any content related effect that might impact social influence is being indirectly captured when we make use of such user interactions.

Having argued for not exploiting content to estimate social curiosity, we note that we did use message categories as a proxy for content properties when deriving metrics related to the traditional collative variables (see Section 5.4.3, acknowledging its importance as a component of curiosity stimulation in general – not only from a social perspective). Obviously, categories, as those used by us, offer only a very coarse approximation of content properties and their use as such is another a limitation of the study, as we further elaborate next.

5.6.2 Media Categorization

Our choice of exploiting message categories in the derivation of the traditional collative variables follow prior work [Zhao and Lee, 2016], including ours [Sousa et al., 2019], which also used pre-existing content categories to derive metrics associated with content novelty, uncertainty, conflict and complexity. In our present context, there is no pre-existing categorization of content (as exists in platforms like LastFM [Sousa et al., 2019]). Thus, we chose to use a coarse categorization (the only one available in our dataset) – media type. It is a simplification. However, by doing so, we make our metrics independent of any specific method to build such categorization. The multitude of strategies that could be applied to do so – e.g. topic models [Kapugama Geeganage, 2018], text embedding strategies [Esmeli et al., 2020; Chang et al., 2020; Yates et al., 2021], exploring a finer grained categorization including memes, emojis, stickers and even more recently deployed WhatsApp features[19] – justifies a study on itself, which should be subject of future work.

Also, as mentioned in Section 5.4.1, a key assumption of ours is that since different media types typically require different levels of engagement of the user to access the content, they should have different effects on curiosity stimulation. For example, textual content is immediately visible, whereas URLs require the user to leave WhatsApp and go to another (Web) application, which is something that the user might feel inclined to defer to another time (or even not do at all). Also, reading a textual message (most often) requires much less effort and much less time than watching a video or listening to an audio. Again, the user may simply choose to defer watching the video or listening the audio to a later time, because she either does not feel like or cannot do it at the time she sees the message. Indeed, this assumption finds some evidence in recent studies of WhatsApp messages which revealed distinct properties

[19] `https://faq.whatsapp.com/general/chats/about-forwarding-limits/?lang=en`

in terms of propagation dynamics depending on the type of media [Resende et al., 2019b,a; Maros et al., 2020, 2021].

To offer further evidence of the potential distinct effects of different media types in curiosity stimulation, we computed, for each category c ($c \in$ {“text", “image",“audio", “video", “URL"}), the time interval between the sharing of a message with category c and the next message (regardless of its category) shared in the same group. This time interval is a (rough) estimate of the reaction time, i.e., the time it took for someone in the group to react (by sharing new content) to the first message. Shorter reaction times might suggest a greater impact on user behavior. Obviously, this should be considered only a coarse approximation of reality, as several other (internal and external) factors may have driven the user to share the new content. Yet, we found very different reaction times for the five categories analyzed. Specifically, average reaction times and corresponding 95% confidence intervals, computed for our entire dataset, are as follows: (a) 4.69±0.02 minutes for text; (b) 7.73±0.07 minutes for images; (c) 8.10±0.18 minutes for audios; (d) 10.90 ± 0.10 minutes for videos; and (e) 11.61 ± 0.09 minutes for URLs. Clearly, messages with textual content tend to have shorter reaction times, whereas URLs and videos tend to trigger much longer reaction times, which corroborates the assumption discussed in the last paragraph.

5.6.3 Focus on Intra-group Curiosity Stimulation

As a first effort to quantify social curiosity in group communication, we chose to focus on each group separately, under the assumption that *intra-group social influence* is a primary factor driving user curiosity. Indeed, our results suggest that the curiosity of a user may be stimulated differently depending on the group (see, e.g., results in Figures 5.8a and 5.8b as well as Figure 5.9). By focusing on *intra-group* curiosity stimulation we are able to isolate (potentially secondary) *inter-group* effects, offering a foundation upon which follow-up studies can be developed. Moreover, we note that our dataset, as all datasets used in other analyses of WhatsApp [Garimella and Tyson, 2018; Moreno et al., 2017; Caetano et al., 2019; Kariuki and Ofusori, 2017], are collected from publicly accessible groups. These are essentially different from (private) family and friend groups where members have external (real-world) connections that may influence their curiosity stimulation. In publicly accessible groups, it is not possible to identify such external links among users.

5.6.4 Modeling of Temporal Dynamics

In the derivation of our metrics, we assume that the curiosity driving the sharing of a message has a period of activation δ_T determined by a fixed-duration time window. In other words, we assume that only messages shared within the specific time window stimulate the user's curiosity. This time window allows us to capture changes in the curiosity stimulation of a user over time. However, it is not possible to determine, from the data, whether indeed the user read all messages shared during a time window. Our assumption is thus a design choice to be able to capture the temporal dynamics of user curiosity.

Such choice, and in particular the window duration of 30 minutes, is based on observations in prior work [Caetano et al., 2021], where authors used the same window duration to study collective attention in publicly accessible WhatsApp group. The authors experimented with several window durations (from 5 minutes to 2 hours), selecting 30 minutes after manual investigation of the topics discussed in the analyzed groups during the defined windows. They argue that 30-minute windows led to the best tradeoff between capturing small variations of attention over time and still approximating the duration of continuous conversations in WhatsApp. Given the close relationship between attention and curiosity, we chose to follow the same approach, especially because our work focuses on Whatsapp groups similar to those analyzed in [Caetano et al., 2021], that is, publicly accessible political-oriented groups in Brazil.

As an alternative to fixed-duration time windows, one could consider windows with variable durations. For example, one might argue that the window duration should be adjusted based on the current level of activity (message sharing) in the group. However, we found very weak correlations between the number of messages shared during a window and the average user and group-level social curiosity, as estimated by the proposed metrics (Spearman correlation coefficients below 0.14). Nevertheless, investigating alternative approaches to model the temporal dynamics of user curiosity stimulation is an avenue worth pursuing in the future.

5.7 Summary

In this chapter, we have investigated curiosity stimulation as a driving force behind content sharing in group communication, specifically in WhatsApp groups, focusing on a novel component of curiosity stimulation, namely social influence. To that end, we proposed novel metrics to quantify such component, as well as metrics that instantiate other traditionally studied collative variables related to curiosity stimulation

(though not to social behavior). We used the proposed metrics to show that social influence is indeed a distinct component of curiosity stimulation, as compared to the other traditional variables, as well as to offer a broad characterization of social curiosity stimulation in a number of publicly accessible political-oriented WhatsApp groups in Brazil.

Recall that, as presented in Section 5.1, we here aimed at addressing three research questions, which we repeat below:

- **RQ2.1:** *How to quantify social influence as a stimulus to one's curiosity driving the information dissemination process?*
- **RQ2.2:** *How does social influence relate to other collative variables priorly associated with curiosity stimulation?*
- **RQ2.3:** *How are users characterized in terms of social stimulation to curiosity?*

Towards answering this RQ2.1, we proposed four new metrics to capture social influence as a component of curiosity stimulation driving individual users to communicate with each other by sharing content in a currently very popular group communication platform – WhatsApp. The derivation of our metrics is founded on the arguments of psychologist Daniel Berlyne on the modeling of human curiosity and its connection with information theory [Berlyne, 1960].

The challenges we faced while deriving these metrics relate to how to instantiate Berlyne's general methodology to quantify collative variables associated with curiosity stimulation to the particular context of social influence in group communication. Specifically, we had to make assumptions so as to be able to deal with the lack of explicit signals of user interactions in our dataset. Also, results from prior studies of WhatsApp [Caetano et al., 2021] motivated us to derive metrics capturing not only the traditionally studied direct influence but also indirect influence reflecting the possible presence of some great influencers in the group. Similarly, we argued for the relevance of proposing metrics for both individual users and groups. Notably assessing group-level social curiosity may offer valuable insights into the dynamics of group members provided that groups are reasonably small and discussions are more focused on specific topics, as is the case of WhatsApp groups. This is in contrast to more open spaces of communication (e.g., Twitter, Facebook), where group membership is unlimited and often much larger, and, as such, curiosity stimulation driving the discussions might be more disperse and perhaps more fragmented.

RQ2.2, in turn, relates to how social influence, as captured by the proposed metrics, captures aspects of curiosity stimulation not covered by novelty, complexity, conflict and uncertainty, which are other collative variables associated with curiosity stimulation that have already been studied in several domains (though not in group communication). Towards answering this question, we first presented seven metrics that capture these collative variables for our target environment. The derivation of these metrics is greatly inspired by previous studies of the same variables in other setups, notably our own effort to model user curiosity in LastFM [Sousa et al., 2019], which, in turn, are rooted, once again, on Berlyne's seminal work.

In particular, as done in prior work, we considered two different elements that may stimulate one's curiosity, namely: content, captured by the categories associated with the several portions of the message, and the user sharing the content itself. We then proposed metrics that capture the role of both elements on curiosity stimulation with respect to the four aforementioned collative variables.

Having defined these metrics, we answered RQ2.2 by evaluating how correlated the metrics of social influence are to the metrics related to the other variables. This analysis, as all others discussed below, were performed on a large dataset containing over 2 million messages shared by over 7.5 thousand users in 335 publicly accessible WhatsApp groups in Brazil. The dataset covers a one-year time interval, which includes a period of large use of WhatsApp for the political debate in the country[20]. As shown in Figure 5.2, our key result is that the four metrics related to social influence are complementary to the other metrics in the vast majority of cases. Thus, social influence, as captured by our proposed metrics, does indeed represent a novel component of curiosity stimulation that is not covered by the other variables.

Having addressed RQ2.2, we then proceeded to tackle RQ2.3 by performing a broad characterization of social curiosity stimulation at three levels of aggregation, namely individual messages, individual members of a group (i.e., users) and all members of a group.

To our knowledge, our findings are novel, with respect to prior efforts to characterize social curiosity as well as prior studies of user behavior in group communication platforms (notably WhatsApp). As such, they offer valuable insights into user behavior (i.e., content sharing) on a platform that has had paramount importance as major source of information (including misinformation) in several countries, with reported impacts on political and social aspects of these societies[21]. Understanding the factors

[20] https://www.niemanlab.org/2018/10/what-to-know-about-whatsapp-in-brazil-ahead-of-sundays-election/

[21] https://www.washingtonpost.com/news/theworldpost/wp/2018/11/01/whatsapp-2/;

driving users to share content in such platform, social curiosity being one of them, is thus valuable not only from the perspective of human behavior analysis but also because such understanding may motivate future developments to build more socially driven spaces for online communication.

Complementary, our metrics could be applied to analyze social curiosity in other platforms of group communication (e.g., Telegram), or adapted to other types of social media applications (e.g., Twitter, Facebook). Moreover, one could further analyze the relationship between user and group-level social curiosity stimulation. Our proposed social curiosity metrics could be employed, jointly with metrics related to other collative variables, to build more sophisticated curiosity models, which in turn could be explored to improve information dissemination in the target platforms (e.g., as a component of recommendation or search systems).

https://www.bbc.com/news/world-asia-india-44856910;
https://www.bbc.com/news/education-55332321

Chapter 6

Summary of Results and Next Steps

In this chapter we present the summary of the results achieved so far (Section 6.1) and we provide a plan of the next steps (Section 6.2) to conclude this dissertation.

6.1 Results So Far

Recall from Chapter 1 that our hypothesis in this dissertation is that by modeling and analyzing human curiosity we are able to uncover relevant knowledge to better understand information dissemination and to design more effective personalized information services. With that hypothesis as guideline, we have defined four research questions:

- **RQ1**: *Are there distinct user behavior profiles in terms of curiosity stimuli, as captured by multiple collative variables?*
- **RQ2**: *How can we capture social influence as a component of human curiosity stimulation driving online information dissemination?*
- **RQ3**: *To which extent, user curiosity driving online information dissemination can be accurately modeled by a Wundt's curve?*
- **RQ4**: *Can the curiosity models be explored to improve the effectiveness of online information services, specifically content recommendation?*

We refer to the aforementioned questions as *major research questions* (RQs). In contrast, we refer to specific questions raised to investigate each case study as *supplementary research questions*. We illustrate the relationship between the major research questions and supplementary research questions in the diagram presented Figure 6.1. Major research questions are shown in purple, whereas supplementary

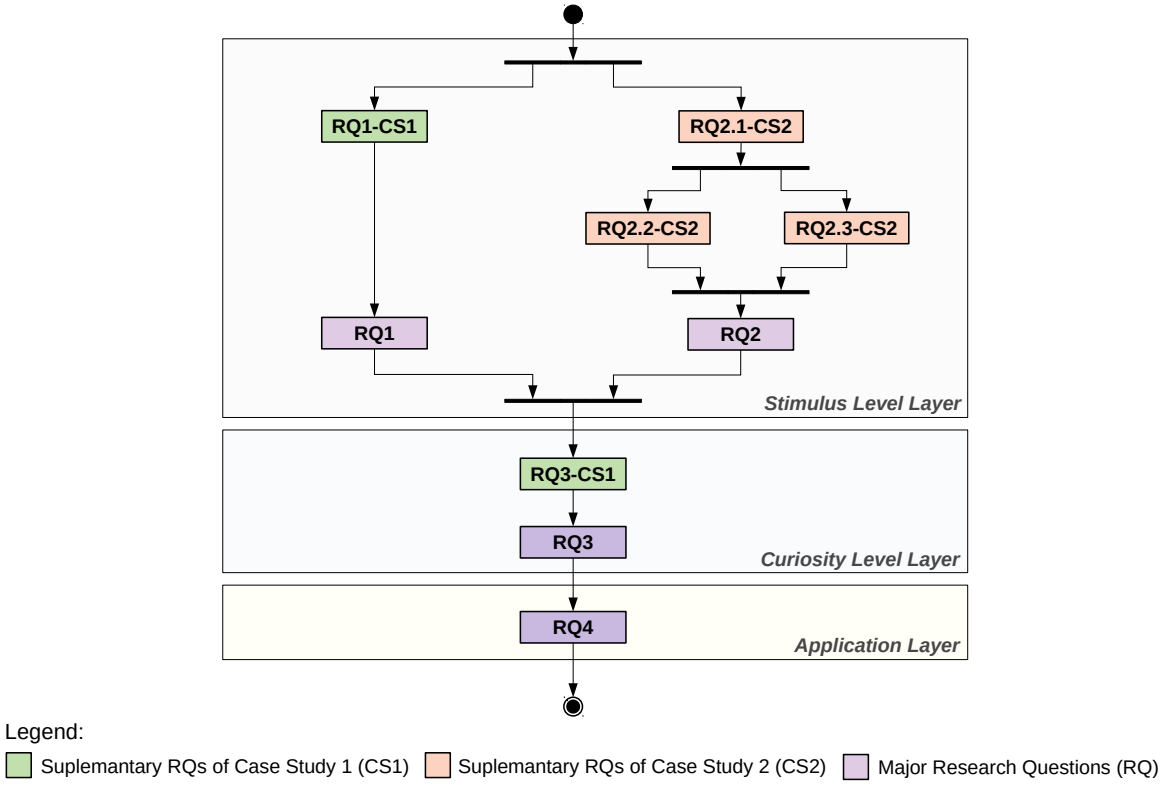


Figure 6.1: Diagram with the major research questions and supplementary research questions associated with each case study.

questions related to case studies 1 (CS1) and 2 (CS2) are shown in green and salmon, respectively. As shown in the figure (major and supplementary) research questions fall into three complementary layers, to know:

(*i*) *Curiosity stimulation layer*: encompasses research questions tackling the first step of the general curiosity appraisal process, that is, the derivation of metrics to enable the quantification of different collative variables associated with curiosity stimulation;

(*ii*) *Curiosity score layer*: encompasses research questions related to the second step of the general appraisal process, which aims to relate estimates of stimuli to a curiosity score driving user behavior. In that particular layer, our focus was on investigating the extent to which a single Wundt's curve (widely used to map curiosity stimuli to curiosity score) is the most adequate model to all users;

(*iii*) *Application layer*: focused on the use of curiosity scores and curiosity stimuli metrics to improve the effectiveness of personalized information systems, specifically recommender systems (RQ4).

Our contributions so far, summarized in the following subsections, are focused on the two first layers, namely curiosity stimulation and curiosity score. Specifically, we have investigated RQ1 and RQ3 as part of case study 1, presented in Chapter 4. We then turned our attention to RQ2, focused on social influence as a component of curiosity, as we developed case study 2[1], as presented in Chapter 5. RQ4 still remains to be tackled, and our plan to address it, making use of our results so far in the application layer, is presented in Section 6.2.

6.1.1 RQ1: Multiple Collative Variables and Discovery of User Curiosity Profiles

The exploration of our first case study (Chapter 4), led to the following main results related to RQ1:

1. We proposed metrics to capture different aspects related to user curiosity stimulation, notably aspects related to the traditionally studied collative variables of uncertainty, novelty, complexity and conflict. We also performed an in depth investigation to identify possible interplay between these variables aiming at identifying complementary as well as redundant metrics. Specifically, we identified that five metrics, namely, novelty (song, artist and musical genre), uncertainty (musical genre) and complexity (musical genre) capture complementary components of curiosity stimulation in most of the cases.

2. Using the five complementary metrics and clustering analysis, we uncovered four different profiles of curiosity stimulation at the access level, namely: (1) *underground*, characterized by low stimulus levels according to all metrics; (2) *niche radical*, characterized by accesses to repeated (low novelty) but complex (multiple musical genres) songs; (3) *eclectic* refers to an opposite pattern in which curiosity stimulation is driven towards new and different songs; and (4) *explorer*, which is similar to *eclectic* but characterized by accesses to more complex songs, i.e., songs with greater diversity of genres.

3. At the user level, by observing the musics a user listens to over time, we were able to uncover dynamic patterns representing how a user's curiosity stimulation changes across different profiles. In particular, we point out two important patterns. Firstly, there is a clear trend towards repeated behavior, i.e., the user tends

[1]We note that even though we did propose metrics associated with various collative variables in case study 2, the analysis and results are mostly focused on *social influence*.

to remain on the curiosity stimulation profile over successive listening events. Secondly, though changes in curiosity stimulation profile do occur occasionally, for all profiles, there are larger chances of a user migrating between Eclectic and Explorer profiles;

4. The most of the curiosity stimulation profiles (*underground*, *eclectic*, and *explorer*) show patterns in which users prefer songs with higher stimulus values (in terms of complexity, uncertainty, and novelty), i.e., they prefer new songs by different artists with different musical genres related to those they have recently listened. This suggests that in the process of online information dissemination, users select content item that is different from what they have consumed in the recent past. This is consistent with the way certain online content goes viral for a short period of time (bursts of popularity) and then fades into oblivion (perhaps from boredom due to repeated consumption – familiarity – or loss of attention).

To our knowledge, our attempt to quantify the curiosity stimulation by multiple collative variables as a driver of online music consumption is a novelty that significantly extends the current state of the art. Our observations show that user curiosity stimulation, as captured by the traditionally studied variables, can be quite complex and to some extent highly dynamic. Building on these results, in our next contributions we extend the set of collative variables to include the effects of social influence on curiosity stimulation and explore how curiosity stimuli, as estimated by the proposed metrics, should map to a curiosity score.

6.1.2 RQ2: Social Influence as Component of Human Curiosity Stimulation

Our key take-home messages regarding RQ2, produced as results of our second case study (Chapter 5), can be summarized as:

1. Curiosity stimulation varies greatly not only across individual messages (consistently with observations made in our first case study) but also across different users and different groups (the latter two taken as average stimuli). By employing clustering analysis, we were able to uncover profiles of social curiosity stimulation at both message and user levels.

2. At the message level, we uncovered three profiles, which indicate that, although social influence does not always play a clear role as a component of curiosity stimulation (profile independent), in almost 30% of the cases, either direct or

indirect social influence, as captured by our metrics, are indeed important components driving one's curiosity to share content (profiles dependent and indirect, respectively).

3. At the user level, we showed that the evolution of user curiosity over time with respect to the three message-level stimulation profiles present great heterogeneity, revealing five different user profiles. While some users do exhibit a rather stable curiosity stimulation process, remaining at the same message-level profile over time, for many other users, the role of (direct/indirect) social influence as a component of curiosity stimulation changes greatly over time, probably in response to the dynamics of interactions within the group.

4. We also showed evidence that curiosity stimulation may change significantly depending on the group the user participates in This observation validates one of our key assumptions (Chapter 3), offering also insights into the role that such (often small) groups have as drivers to user behavior (notably content sharing). In some sense, these results mimic what we see in real life when the same person may behave quite differently (more or less participative/chatty) depending on the group of people she is interacting with.

5. At the group level, we noted that the role of social curiosity is more clearly observed when group members are more engaged in the ongoing discussions, more often sharing content (thus interacting with others), as these actions suggest a greater susceptibility to social curiosity.

6.1.3 RQ3: Investigating the Wundt's Curve as Model of Human Curiosity

In a second step developed in our first case study, we explored the proposed metrics of curiosity stimulation to investigate the suitability of the Wundt's curve as a model for curiosity elicitation. Our main findings revealed that:

1. The traditionally used single bell-shaped Wundt's curve does not adequately model the curiosity elicitation process of a large proportion (almost half) of the users. For these users, we found instead that the curiosity curve has a multimodal shape that may reflect the quite complex and dynamic patterns we uncovered in our analyzes of the curiosity stimulation metrics.

2. For such users, we suggest using a combination of two or three Beta distributions, as these can capture the different distribution modes and thus reduce errors in identifying the areas of greater curiosity stimulation.

Moreover, these observations can be naturally extended to other domains besides music, such as communication platforms (e.g. WhatsApp and Telegram) and content production and sharing services (e.g., Twitter, Instagram and TikTok). Finally, these findings form the basis for designing more effective personalized services. As discussed in Section 6.2, we intend to exploit these findings to propose the incorporation of human curiosity estimates into a recommender system.

6.1.4 Publications

The results we obtained so far, covering each case study, are summarized in the following publications:

- Sousa, A. M.; Almeida, J. M.; Figueiredo, F.. Analyzing and Modeling User Curiosity in Online Content Consumption: A LastFM Case Study. IEEE/ACM The International Conference Series on Advances in Social Network Analysis and Mining (ASONAM), 2019.
- Sousa, A. M.; Almeida, J. M.; Figueiredo, F.. Metrics of Social Curiosity: The WhatsApp Case. Online Social Networks and Media (OSNEM), 2022.

6.2 Next Steps

In this section, we discuss the next steps toward completing this dissertation. So far, our contributions relate specifically to the first three research questions established in Section 1.3. To conclude this dissertation, we focus on applying our model of human curiosity in recommender systems to answer major RQ4. To that end, we intend to begin developing a hybrid recommendation framework that should include: (a) general appraisal process – which consists of a complete set of collative variables (including social curiosity) and our Wundt's curve model for heterogeneous and dynamic behavior of online users; and (b) a machine learning-based recommendation model that incorporates a range of content and user attributes, including elements of curiosity. Secondly, we want to evaluate the performance of our proposed framework against the main state-of-the-art baselines selected from the literature, which are summarized in Table 2.1 (Section 2.2.2). We aim at assessing the extent to which (and possibly in which

scenarios) the addition of the curiosity models improves the quality of recommendations.

As a first step, we want to use *neural networks* to develop a recommendation algorithm. This idea is based on the observations that *neural networks* are generally able to uncover complex and non-linear relationships between users and items [Afsar et al., 2022]. According to [Matsubara et al., 2017; David and Jon, 2010b], online social network data exhibit non-linear relationships that affect the information dissemination. Indeed, recent studies have already used *neural networks* to solve sparsity and cold-start problems in recommendation algorithms [He et al., 2017; Ko et al., 2022], making these models an excellent option for building hybrid recommender systems.

Our intuition is that by introducing the general appraisal process of curiosity into the recommendation machinery we will be able to produce recommendations that better suit the idiosyncrasies of individual users. To accomplish this, we intend to include metrics related to all collative variables considered in the two case studies. Specifically we plan to do the following: (1) adapt the social curiosity metrics to a scenario of information consumption consistent with the domain of recommender systems; (2) explore the set of metrics under consideration that capture complementary aspects of curiosity stimulation; and (3) incorporate heterogeneous multimodal models that map curiosity stimuli to curiosity values. For (1), we intend to explore the user-rating matrix and the explicit social links that may be available, as was done by [Wu et al., 2016, 2017]. For (2), we intend to explore the complementarity/redundancy of different metrics to identify a reduced set of selected metrics. Finally, for (3), we intend to study a Beta-mixture distribution as a model of the Wundt's curve(s) – as in Chapter 4, but we also consider analyzing a Kernel Density Estimation (KDE), which is a *nonparametric* estimate of a probability density function [Theodoridis, 2015]. The advantage of KDE method over the former is that no data scaling is required, unlike in the Beta-mixture distribution (e.g., the range of the Beta distribution is in $[0, 1]$), i.e., the curiosity stimulus values are used in bit units.

We intend to evaluate our framework using the LastFM dataset (the same one used in our first case study). We intend to compare our framework with the main baselines from Table 2.1 to know Zhao and Lee [Zhao and Lee, 2016], Shrestha *et al.* [Shrestha et al., 2020] and Xu *et al.* [Xu et al., 2021].

6.2.1 Planned Schedule

Given our planned next steps, discussed above, the remaining activities to finalize this dissertation are as follows:

1. Develop our hybrid framework of recommender system with human curiosity model incorporated;
2. Implementation of the main baselines from literature;
3. Design and perform a set of experiments to compare and quantify the effectiveness and personalization of proposed framework over existing baselines;
4. Submission of results for publication;
5. Write dissertation;
6. Defense.

The planned scheduled to develop the aforementioned activities is shown below:

Table 6.1: Tentative schedule to finalize this dissertation.

Task	**Months**								
	1	2	3	4	5	6	7	8	9
Develop RS framework	✓	✓	✓						
Implementation of baselines			✓	✓	✓				
Design and perform experiments				✓	✓	✓	✓		
Submission of results								✓	✓
Write Dissertation						✓	✓	✓	
Defense									✓

Bibliography

Abbas, F. and Niu, X. (2019). One size does not fit all: Modeling users' personal curiosity in recommender systems. *arXiv preprint*, arXiv:1907.00119.

Adamic, L. A. and Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, 25(3).

Afsar, M. M., Crump, T., and Far, B. (2022). Reinforcement learning based recommender systems: A survey. *ACM Comput. Surv.* Just Accepted.

Ah-Pine, J. (2018). An efficient and effective generic agglomerative hierarchical clustering approach. *Journal of Machine Learning Research*.

Ahmadlou, M., Houba, J. H. W., van Vierbergen, J. F. M., Giannouli, M., Gimenez, G.-A., van Weeghel, C., Darbanfouladi, M., Shirazi, M. Y., Dziubek, J., Kacem, M., de Winter, F., and Heimel, J. A. (2021). A cell type–specific cortico-subcortical brain circuit for investigatory and novelty-seeking behavior. *Science*, 372(6543).

Akyildiz, I. F. and et al (2016). 5g roadmap: 10 key enabling technologies. *Computer Networks*, 106:17–48.

Al-Doulat, A. (2018). Surprise and curiosity in a recommender system. In *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–2. IEEE.

Alicart, H., Cucurell, D., and Marco-Pallarés, J. (2020). Gossip information increases reward-related oscillatory activity. *NeuroImage*, 210:116520.

Amorim, R. C. and Hennig, C. (2015). Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences*, 324.

Arif, M., Faisal, C. N., Ahmad, H., Habib, M. A., Ahmad, M., and Mehmood, N. (2020). The moderating role of curiosity between interactivity and cognitive motives. In *2020 IEEE 23rd International Multitopic Conference (INMIC)*, pages 1–6. IEEE.

Arnone, M. and Small, R. V. (1995). Arousing and sustaining curiosity: Lessons of arcs model. In *Proceedings of the Annual National Conference of the Association of Educational Communications and Technology (AECT)*, pages 1–17, Anaheim, CA, USA. Association of Educational Communications and Technology (AECT).

Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, page 519–528, New York, NY, USA. Association for Computing Machinery (ACM).

Bechtle, S., Lin, Y., Rai, A., Righetti, L., and Meier, F. (2019). Curious ilqr: Resolving uncertainty in model-based rl. In *Proceedings of the 3rd Conference on Robot Learning*, volume 100, pages 162–171, Osaka, Japan. PMLR.

Benson, A. R., Kumar, R., and Tomkins, A. (2016). Modeling user consumption sequences. In *Proceedings of the 25st international conference on World Wide Web*, WWW'16, pages 519–529, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Berlyne, D. (1960). *Conflict, Arousal and Curiosity*. McGraw-Hill series in psychology. McGraw-Hill, NY, USA.

Bin, S., Chen, C.-C., and Sun, G. (2020). Maximizing social influence in nearly optimal time: Sris model. In *2020 IEEE 2nd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*, pages 201–203. IEEE.

Bonchi, F., Gullo, F., Mishra, B., and Ramazzotti, D. (2018). Probabilistic causal analysis of social influence. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pages 1003–1012, New York, NY, USA. Association for Computing Machinery (ACM).

Bossomaier, T., Barnett, L., Harré, M., and Lizier, J. T. (2016). *Information Theory*. Springer International Publishing, Cham.

Boyle, G. J. (1989). Breadth-depth or state-trait curiosity? a factor analysis of state-trait curiosity and state anxiety scales. *Personality and Individual Differences*, 10(2):175–183.

Caetano, J. A., Almeida, J. M., Gonçalves, M. A., Meira Jr., W., Marques-Neto, H. T., and Almeida, V. A. F. (2021). Analyzing topic attention in online small groups. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social*

Networks Analysis and Mining, ASONAM '21, pages 1–5. Association for Computing Machinery (ACM).

Caetano, J. A., Magno, G., Gonçalves, M. A., Almeida, J. M., Marques-Neto, H. T., and Almeida, V. A. F. (2019). Characterizing attention cascades in whatsapp groups. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 27–36. Association for Computing Machinery (ACM).

Ceha, J., Chhibber, N., Goh, J., McDonald, C., Oudeyer, P.-Y., Kulić, D., and Law, E. (2019). Expression of curiosity in social robots: Design, perception, and effects on behaviour. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, New York, NY, USA. ACM.

Centola, D. (2019). Influential networks. *Nature Human Behaviour*, 3(1):664–665.

Centola, D. and Macy, M. (2007). Complex contagions and the weakness of long ties. *American Journal of Sociology*, 113(3):702–734.

Chang, W.-C., Yu, H.-F., Zhong, K., Yang, Y., and Dhillon, I. S. (2020). Taming pretrained transformers for extreme multi-label text classification. In *Proc. of the 26th ACM SIGKDD*, KDD '20, pages 3163–3171, New York, NY, USA. Association for Computing Machinery.

Chen, B.-L., Zeng, A., and Chen, L. (2015). The effect of heterogeneous dynamics of online users on information filtering. *Physics Letters A*, 379(43):2839–2844.

Chen, L., Yang, Y., Wang, N., Yang, K., and Yuan, Q. (2019). How serendipity improves user satisfaction with recommendations? a large-scale user evaluation. In *The World Wide Web Conference*, WWW' 19, page 240–250, New York, NY, USA. Association for Computing Machinery (ACM).

Choi, M., Aiello, L. M., Varga, K. Z., and Quercia, D. (2020). Ten social dimensions of conversations and relationships. In *Proceedings of The Web Conference 2020*, WWW '20, page 1514–1525, New York, NY, USA. Association for Computing Machinery (ACM).

Collins, R. P., Litman, J. A., and Spielberger, C. D. (2004). The measurement of perceptual curiosity. *Personality and Individual Differences*, 36(5):1127–1141.

Coró, F., D'angelo, G., and Velaj, Y. (2021). Link recommendation for social influence maximization. *ACM Transactions on Knowledge Discovery from Data*, 15(6):94:1–94:23.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience, NY, USA.

David, E. and Jon, K. (2010a). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, USA.

David, E. and Jon, K. (2010b). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, USA. ISBN 0521195330.

Doering, M., Liu, P., Glas, D. F., Kanda, T., Kulić, D., and Ishiguro, H. (2019). Curiosity did not kill the robot: A curiosity-based learning system for a shopkeeper robot. *ACM Transactions Human-Robot Interaction*, 8(3):1–24.

Esmeli, R., Bader-El-Den, M., and Abdullahi, H. (2020). Using word2vec recommendation for improved purchase prediction. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Ferreira, C. H. G., Murai, F., Matos, B. d. S., and Almeida, J. M. (2019). Modeling dynamic ideological behavior in political networks. *The Journal of Web Science*, 7:1–14.

Garimella, K. and Tyson, G. (2018). Whatsapp doc?: A first look at whatsapp public group data. *ICWSM*.

Golman, R. and Loewenstein, G. (2016). Information gaps: A theory of preferences regarding the presence and absence of information. *Decision - American Psychological Association*, 5(3):143–164.

Golman, R. and Loewenstein, G. (2018). The desire for knowledge and wisdom. In Gordon, G., editor, *The New Science of Curiosity*, chapter 1, pages 1–5. Nova Science Publishers.

Gordon, G., Breazeal, C., and Engel, S. (2015). Can children catch curiosity from a social robot? In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 91–98, New York, NY, USA. ACM/IEEE.

Gottlieb, J. and et al (2013). Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences*, 17(11):585– 593.

Goyal, A., Bonchi, F., and Lakshmanan, L. V. (2010). Learning influence probabilities in social networks. In *Proceedings of the Third ACM International Conference on*

Web Search and Data Mining, WSDM '10, pages 241–250, New York, NY, USA. Association for Computing Machinery (ACM).

Grace, K. and Maher, M. L. (2015). Specific curiosity as a cause and consequence of transformational creativity. In Toivonen, H., Colton, S., Cook, M., and Ventura, D., editors, *Proceedings of the Sixth International Conference on Computational Creativity, ICCC 2015*, pages 260–267, Park City, Utah, USA. ICCC.

Greasley, A., Lamont, A., and Sloboda, J. (2013). Exploring musical preferences: An in-depth qualitative study of adults' liking for music in their personal collections. *Qualitative Research in Psychology*, 10(4).

Guilbeault, D., Becker, J., and Centola, D. (2018). *Complex Contagions: A Decade in Review*, chapter Part I: Introduction to Spreading in Social Systems, pages 3–25. Springer International Publishing, Cham.

Guilbeault, D. and Centola, D. (2021). Topological measures for identifying and predicting the spread of complex contagions. *Nature Communications*, 12(1):4430.

Guo, Y., Cao, J., and Lin, W. (2019). Social network influence analysis. In *2019 6th International Conference on Dependable Systems and Their Applications (DSA)*, pages 517–518, Harbin, China. IEEE.

Hartung, F. M. and Renner, B. (2013). Social curiosity and gossip: Related but different drives of social functioning. *PLoS One*, 8(7):e69996.

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., and Chua, T.-S. (2017). Neural collaborative filtering. In *Proceed. of the 26th WWW*, WWW' 17, pages 173–182, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Hsee, C. K. and Ruan, B. (2016). The pandora effect: The power and peril of curiosity. *Psychological science*, 27(5):659–666.

Huang, Z., Wang, Z., Zhang, R., Zhao, Y., and Zheng, F. (2020). Learning bi-directional social influence in information cascades using graph sequence attention networks. In *Proc. WWW*.

Hung, H.-J., Yang, D.-N., and Lee, W.-C. (2016). Social influence-aware reverse nearest neighbor search. *ACM Transactions on Spatial Algorithms and Systems*, 2(3).

Ivanov, S., Theocharidis, K., Terrovitis, M., and Karras, P. (2017). Content recommendation for viral social influence. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 565–574, New York, NY, USA. Association for Computing Machinery (ACM).

John, O. P. and Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of personality: Theory and research*, volume 2, pages 102–138, New York, USA. Guilford Press.

Kapugama Geeganage, D. T. (2018). Concept embedded topic modeling technique. In *Companion Proceedings of the The Web Conference 2018*, page 831–835.

Kariuki, P. and Ofusori, L. O. (2017). Whatsapp-operated stokvels promoting youth entrepreneurship in durban, south africa: Experiences of young entrepreneurs. In *10th ICEGOV*.

Kashdan, T. B., Disabato, D. J., Goodman, F. R., and McKnight, P. E. (2020a). The five-dimensional curiosity scale revised (5dcr): Briefer subscales while separating overt and covert social curiosity. *Personality and Individual Differences*, 157:109836.

Kashdan, T. B., Goodman, F. R., Disabato, D. J., McKnight, P. E., Kelso, K., and Naughton, C. (2020b). Curiosity has comprehensive benefits in the workplace: Developing and validating a multidimensional workplace curiosity scale in united states and german employees. *Personality and Individual Differences*, 155:109717.

Kashdan, T. B., Stiksma, M. C., Disabato, D. J., McKnight, P. E., c, J. B., Kaji, J., and Lazarus, R. (2018). The five-dimensional curiosity scale: Capturing the bandwidth of curiosity and identifying four unique subgroups of curious people. *Journal of Research in Personality*, 73:130–149.

Kaufman, L. and Rousseeuw, P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley, USA.

Kidd, C. and Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, 88(3):449–460.

Kloumann, I., Adamic, L., Kleinberg, J., and Wu, S. (2015). The lifecycles of apps in a social ecosystem. In *Proc. WWW*.

Ko, H., Lee, S., Park, Y., and Choi, A. (2022). A survey of recommendation systems: Recommendation models, techniques, and application fields. *Electronics*, 11(1).

Kosinski, M. S. (2014). *Measurement and Prediction of Individual and Group Differences in The Digital Environment.* PhD dissertation, U. Cambridge, UK.

Kumar, A. and Schrater, P. (2017). Novelty learning via collaborative proximity filtering. In *Proceedings of the 22nd IUI.*

Kumar, N., Guo, R., Aleali, A., and Shakarian, P. (2016). An empirical evaluation of social influence metrics. In *2016 IEEE/ACM ASONAM*, ASONAM' 16, pages 1329–1336, Davis, California, USA. IEEE Press - ASONAM.

Lambiotte, R. and Kosinski, M. (2014). Tracking the digital footprints of personality. *Proceedings of the IEEE*, 102(12):1934–1939.

Lau, J. K. L., Ozono, H., Kuratomi, K., Komiya, A., and Murayama, K. (2020). Shared striatal activity in decisions to satisfy curiosity and hunger at the risk of electric shocks. *Nature Human Behaviour*, 4:531–543.

Law, E., Baghaei Ravari, P., Chhibber, N., Kulic, D., Lin, S., Pantasdo, K. D., Ceha, J., Suh, S., and Dillen, N. (2020). Curiosity notebook: A platform for learning by teaching conversational agents. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA'20, pages 1–9, New York, NY, USA. Association for Computing Machinery.

Law, E. and et al (2016). Curiosity killed the cat, but makes crowdwork better. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI' 16, page 4098–4110, NY, USA. Association for Computing Machinery (ACM).

Li, B., Lu, T., Li, J., Lu, N., Cai, Y., and Wang, S. (2020). Acder: Augmented curiosity-driven experience replay. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4218–4224. IEEE.

Li, C. and Xiong, F. (2017). Social recommendation with multiple influence from direct user interactions. *IEEE Access*, 5:16288–16296.

Li, K., Zhang, L., and Huang, H. (2018). Social influence analysis: Models, methods, and evaluation. *Engineering*, 4(1):40–46. Cybersecurity.

Li, L., Li, A., Hao, B., Guan, Z., and Zhu, T. (2014). Predicting active users' personality based on micro-blogging behaviors. *PLOS ONE*, 9(1):1–11.

Li, Y., Xie, H., Lin, Y., and Lui, J. C. (2019). To be or not to be: Analyzing amp; modeling social recommendation in online social networks. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1180–1185. IEEE.

Litman, J. A. and Pezzo, M. V. (2007). Dimensionality of interpersonal curiosity. *Personality and Individual Differences*, 43(6):1448–1459.

Liu, S., Wang, S., and Zhu, F. (2015). Structured learning from heterogeneous behavior for social identity linkage. *IEEE Transactions on Knowledge and Data Engineering*, 27(7):2005–2019.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, 116(1):75–98.

Loewenstein, G. (2017). Recommender systems and the new new economics of information. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17, page 1, New York, NY, USA. Association for Computing Machinery (ACM).

Logins, A. and Karras, P. (2019). Content-based network influence probabilities: Extraction and application. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 69–72. IEEE.

Luceri, L., Braun, T., and Giordano, S. (2019). Analyzing and inferring human real-life behavior through online social networks with social influence deep learning. *Applied Network Science*, pages 1–25.

Luo, Y., Huang, Z., Zhang, Z., Wang, Z., Li, J., and Yang, Y. (2019). Curiosity-driven reinforcement learning for diverse visual paragraph generation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pages 2341–2350, New York, NY, USA. Association for Computing Machinery (ACM).

Maccatrozzo, V. and et al (2017a). Everybody, more or less, likes serendipity. In *Proceedings 25th ACM UMAP*, UMAP '17, pages 29–34, New York, NY, USA. Association for Computing Machinery (ACM).

Maccatrozzo, V. and et al (2017b). Sirup: Serendipity in recommendations via user perceptions. In *Proceedings of the 22nd ACM IUI*.

Macedo, L. and Cardoso, A. (2005). The role of surprise, curiosity and hunger on exploration of unknown environments populated with entities. In *2005 Portuguese Conference on Artificial Intelligence (EPIA)*, pages 47–53, Covilha, Portugal. IEEE.

MacKay, D. (2005). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, UK.

Maher, M. L., Merrick, K. E., and Saunders, R. (2008). Achieving creative behaviour using curious learning agents. In *Proceedings of AAAI Spring Symposium on Creative Intelligent Systems*, pages 40–46, USA. AAAI.

Maros, A., Almeida, J., Benevenuto, F., and Vasconcelos, M. (2020). Analyzing the use of audio messages in whatsapp groups. In *Proccedings of the ACM WWW*.

Maros, A., Almeida, J. M., and Vasconcelos, M. (2021). A study of misinformation in audio messages shared in whatsapp groups. In Bright, J., Giachanou, A., Spaiser, V., Spezzano, F., George, A., and Pavliuc, A., editors, *Disinformation in Open Online Media*, pages 85–100, Cham. Springer International Publishing.

Marvin, C. B. and Shohamy, D. (2016). Curiosity and reward: Valence predicts choice and information prediction errors enhance learning. *Journal of Experimental Psychology: General*, 145(3):266–272.

Matsubara, Y., Sakurai, Y., Prakash, B. A., Li, L., and Faloutsos, C. (2017). Nonlinear dynamics of information diffusion in social networks. *ACM Transactions on the Web*, 11(2):1–40.

Melo, P. F., Messias, J., Resende, G., Garimella, K., Almeida, J. M., and Benevenuto, F. (2019). Whatsapp monitor: A fact-checking system for whatsapp. In *Proc. ICWSM*.

Menasce, D. A. and Almeida, V. (2000). *Scaling for E Business: Technologies, Models, Performance, and Capacity Planning*. Pren. Hall, USA.

Menon, S. and Soman, D. (2002). Managing the power of curiosity for effective web advertising strategies. *Journal of Advertising*, 31(3):1–14.

Merrick, K. E. and Maher, M. L. (2009). *Motivated Reinforcement Learning: Curious Characteres for Multiuser Gamers*. Springer Berlin, Heidelberg, Germany, 1 edition.

Millan-Cifuentes, J. D. and et al (2014). Curiosity driven search: When is relevance irrelevant? In *Proceedings of the 5th Information Interaction in Context Symposium, IIiX 2014*, IIiX '14, pages 279–282, NY, USA. Association for Computing Machinery (ACM).

Min, B. and San Miguel, M. (2018). Competing contagion processes: Complex contagion triggered by simple contagion. *Scientific Reports*, 8(1):10422.

Mohseni, M., Maher, M. L., Grace, K., Najjar, N., Abbas, F., and Eltayeby, O. (2019). Pique: Recommending a personalized sequence of research papers to engage student curiosity. In *Artificial Intelligence in Education*, volume 11626, pages 201–205, Cham. Springer International Publishing.

Monção, A. C. B. L., Correa, S. L., Viana, A. C., and Cardoso, V. K. (2021). Optimizing Content Caching and Recommendations with Context Information in Multi-Access Edge Computing. Research report, INRIA Saclay - Ile de France (INRIA).

Moreno, A., Garrison, P., and Bhat, K. (2017). Whatsapp for monitoring and response during critical events: Aggie in the ghana 2016 election. In *14th Information Systems for Crisis Response And Management*.

Nassar, H., Benson, A. R., and Gleich, D. F. (2019). Pairwise link prediction. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '19, pages 386–393, New York, NY, USA. Association for Computing Machinery (ACM).

Negi, S. and Chaudhury, S. (2016). Link prediction in heterogeneous social networks. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 609–617, New York, NY, USA. Association for Computing Machinery.

Niehoff, E. and Oosterwijk, S. (2020). To know, to feel, to share? exploring the motives that drive curiosity for negative content. *Current Opinion in Behavioral Sciences*, 35.

Niu, X. and Al-Doulat, A. (2021). Luckyfind: Leveraging surprise to improve user satisfaction and inspire curiosity in a recommender system. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR '21, pages 163–172, New York, NY, USA. Association for Computing Machinery.

Nobre, G. P., Ferreira, C. H., and Almeida, J. M. (2022). A hierarchical network-oriented analysis of user participation in misinformation spread on whatsapp. *Information Processing & Management*, 59(1):102757.

Nobre, G. P., Ferreira, C. H. G., and de Almeida, J. M. (2020). Beyond groups: Uncovering dynamic communities on the whatsapp network of information dissemination. In *12th Social Informatics*.

Odom, W., Wakkary, R., Hol, J., Naus, B., Verburg, P., Amram, T., and Chen, A. Y. S. (2019). Investigating slowness as a frame to design longer-term experiences with personal data: A field study of olly. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI'19, pages 1–16, New York, NY, USA. ACM.

Ott, R. L. and Longnecker, M. T. (2015). *An Introduction to Statistical Methods and Data Analysis*. Cengage Learning, USA, 7th edition.

Oudeyer, P.-Y., Gottlieb, J., and Lopes, M. (2016). *Intrinsic Motivation, Curiosity, and Learning: Theory and Applications in Educational Technologies*, volume 229 of *Progress in Brain Research*, chapter 11, pages 257–284. Elsevier.

Perez, B., Musolesi, M., and Stringhini, G. (2018). You are your metadata: Identification and obfuscation of social media users using metadata information. In *Procedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 1–10, New Orleans, Louisiana, USA. Association for the Advancement of Artificial Intelligence.

Ramos, A. M., Andrade, N., and Marinho, L. B. (2013). Exploring the relation between novelty aspects and preferences in music listening. In *Proceedings of the 14th ISMIR*.

Rani, N., Chu, S. L., Williamson, Y. G., and Wu, S. (2021). *Curiosity-Inspired Learning: Insitu versus Post-Event Approaches to Recall and Reflection*, pages 1–6. CHI EA '21. Association for Computing Machinery (ACM), New York, NY, USA.

Renner, B. (2006). Curiosity about people: The development of a social curiosity measure in adults. *Journal of Personality Assessment*, 87(3):305–316.

Resende, G., Melo, P., C. S. Reis, J., Vasconcelos, M., Almeida, J. M., and Benevenuto, F. (2019a). Analyzing textual (mis)information shared in whatsapp groups. In *10th ACM Conference on Web Science*.

Resende, G., Melo, P., Sousa, H., Messias, J., Vasconcelos, M., Almeida, J., and Benevenuto, F. (2019b). (mis)information dissemination in whatsapp: Gathering, analyzing and countermeasures. In *Proc. WWW*, WWW '19, pages 818–828, New York, NY, USA. Association for Computing Machinery.

Ribeiro, A. C., Azevedo, B., Oliveira e Sá, J., and Baptista, A. A. (2020). How to measure influence in social networks? In Dalpiaz, F., Zdravkovic, J., and Loucopoulos, P., editors, *Research Challenges in Information Science*, pages 38–57, Cham. Springer International Publishing.

S, A. and Kaimal, R. (2012). Document summarization using positive pointwise mutual information. *International Journal of Computer Science & Information Technology (IJCSIT)*.

Samanta, S., Dubey, V. K., and Sarkar, B. (2021). Measure of influences in social networks. *Applied Soft Computing*, 99:106858.

Santos, A. M. d. (2015). A hybrid recommendation system based on human curiosity. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 367–370, New York, NY, USA. Association for Computing Machinery (ACM).

Santos, A. M. d. and Sebastiá, L. (2016). Predicting the human curiosity from users' profiles on facebook. In *Proceedings of the 4th Spanish Conference on Information Retrieval*, CERI '16, pages 1–8, NY, USA. Association for Computing Machinery (ACM).

Santos, A. M. d., Sebastia, L., and Ferreira, R. (2017). Curumim: A serendipitous recommender system based on human curiosity. *Procedia Computer Science*, 112:484–493.

Santos, R. J. D. F. (2017). Inferring the curiosity by using facebook profile data. Master's thesis, Department of Information Systems and Computation, Univ. Polit. València, Spain.

Saunders, R. and Gero, J. S. (2002). How to study artificial creativity. In *Proceedings of the 4th ACM Creativity & Cognition*, C&C '02, pages 80–87, New York, NY, USA. Association for Computing Machinery (ACM).

Schaekermann, M. and et al (2017). Curiously motivated: Profiling curiosity with self-reports and behaviour metrics in the game "destiny". In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '17, page 143–156, New York, NY, USA. Association for Computing Machinery (ACM).

Schedl, M. (2016). The lfm-1b dataset for music retrieval and recommendation. In *Proc. ACM ICMR*.

Schedl, M. (2019). Genre differences of song lyrics and artist wikis: An analysis of popularity, length, repetitiveness, and readability. In *The World Wide Web Conference*, WWW'19, pages 3201–3207, New York, NY, USA. ACM.

Schedl, M. and Ferwerda, B. (2017). Large-scale analysis of group-specific music genre taste from collaborative tags. In *Proc. IEEE ISM*.

Schneider, A., von Krogh, G., and Jäger, P. (2013). "what's coming next?" epistemic curiosity and lurking behavior in online communities. *Computers in Human Behavior*, 29(1):293–303.

Schoenebeck, G. and Tao, B. (2020). Influence maximization on undirected graphs: Toward closing the (1-1/e) gap. *ACM Transactions on Economics and Computation*, 8(4).

Schreiber, T. (2000). Measuring information transfer. *Physical Review Letters*, 85(2):461–464.

Schröder, C. and Rahmann, S. (2017). A hybrid parameter estimation algorithm for beta mixtures and applications to methylation state classification. *Algorithms for Molecular Biology*, 12(1).

Sculley, D. (2010). Web-scale k-means clustering. In *Proccedings of the ACM WWW*.

Shandhilya, T. and Srivastava, N. (2020). Using conceptual incongruity as a basis for making recommendations. In *Fourteenth ACM Conference on Recommender Systems*.

Shankar, A. (2018). Capitalism, curiosity, and specialization. *The Newsletter of the Technical Committee on Cognitive and Development Systems*, 15(1):12.

Shokeen, J. and Rana, C. (2020). A study on features of social recommender systems. *Artificial Intelligence Review*, 53(2).

Shorten, D. P., Spinney, R. E., and Lizier, J. T. (2021). Estimating transfer entropy in continuous time between neural spike trains or other event-based data. *PLOS Computational Biology*, 17(4):1–45.

Shrestha, P., Zhang, M., Liu, Y., and Ma, S. (2020). Curiosity-inspired personalized recommendation. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 33–40. IEEE.

Silvia, P. J. (2006). *Exploring the Psychology of Interest*. Oxford University Press, New York, NY, USA.

Singer, P. (2014). *Modeling Aspects of Human Trails on the Web*. PhD dissertation, Leibniz-Institute for Social Science, Knowledge Technologies Institute, Graz University of Technology, Graz, Austria.

Singer, P. (2016). Modeling aspects of human trails on the web by philipp singer, with prateek jain as coordinator. *ACM SIGWEB Newsletter*, 1(Winter):3:1– 3:2.

Sousa, A. M., Almeida, J. M., and Figueiredo, F. (2019). Analyzing and modeling user curiosity in online content consumption: a lastfm case study. In *Proc. ASONAM*.

Srivastava, A., Chelmis, C., and Prasanna, V. K. (2014). Influence in social networks: A unified model? In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '14, page 451–454. IEEE Press.

Sun, C., Qian, H., and Miao, C. (2022). From psychological curiosity to artificial curiosity: Curiosity-driven learning in artificial intelligence tasks. *arXiv preprint*, arXiv:2201.08300:1–35.

Sun, J. and Tang, J. (2011). *A Survey of Models and Algorithms for Social Influence Analysis*, chapter 7, pages 177–214. Springer US, Boston, MA.

Sun, J. and Tang, J. (2013). Models and algorithms for social influence analysis. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 775–776, New York, NY, USA. Association for Computing Machinery (ACM).

Tan, C., Tang, J., Sun, J., Lin, Q., and Wang, F. (2010). Social action tracking via noise tolerant time-varying factor graphs. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 1049–1058, New York, NY, USA. Association for Computing Machinery (ACM).

Tang, J. and Sun, J. (2014). Computational models for social influence analysis. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 205–206, New York, NY, USA. Association for Computing Machinery (ACM).

Tang, J., Sun, J., Wang, C., and Yang, Z. (2009). Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 807–816, New York, NY, USA. Association for Computing Machinery (ACM).

Theodoridis, S. (2015). *Machine Learning: A Bayesian and Optimization Perspective*. Academic Press, Inc., USA, 1st edition.

Thilakarathna, K., Viana, A. C., Seneviratne, A., and Petander, H. (2017). Design and analysis of an efficient friend-to-friend content dissemination system. *IEEE Transactions on Mobile Computing*, 16(3):702–715.

Timme, N. M. and Lapish, C. (2018). A tutorial for information theory in neuroscience. *eNeuro*, 5(3):1–40.

Tovanich, N., Centellegher, S., Seghouani, N. B., Gladstone, J., Matz, S., and Lepri, B. (2021). Inferring psychological traits from spending categories and dynamic consumption patterns. *EPJ Data Science*.

Twomey, K. E. and Westermann, G. (2018). Curiosity-based learning in infants: A neurocomputational approach. *Developmental Science*, 21(4):e12629.

Valji, A., Priemysheva, A., Hodgetts, C. J., Costigan, A. G., Parker, G. D., Graham, K. S., Lawrence, A. D., and Gruber, M. J. (2019). White matter pathways supporting individual differences in epistemic and perceptual curiosity. *bioRxiv*.

Vega-Oliveros, D. A. and et al (2017). The impact of social curiosity on information spreading on networks. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, page 459–466, NY, USA. IEEE/ACM.

Venneti, L. and Alam, A. (2018). How curiosity can be modeled for a clickbait detector. *ArXiv*, abs/1806.04212.

Ver Steeg, G. and Galstyan, A. (2012). Information transfer in social media. In *Proceedings of the 21st international conference on World Wide Web*, WWW' 12, pages 509–518, New York, NY, USA. Association for Computing Machinery (ACM).

Ver Steeg, G. and Galstyan, A. (2013). Information-theoretic measures of influence based on content dynamics. In *Proceedings of the Sixth ACM international conference on Web Search and Data Mining*, WSDM '13, pages 3–12, New York, NY, USA. Association for Computing Machinery (ACM).

Voorhis, C. R. W. V. and Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology*, 3(2):43–50.

Wang, C., Guan, X., Qin, T., and Yang, T. (2016). Modeling heterogeneous and correlated human dynamics of online activities with double pareto distributions. *Information Sciences*, 330:186–198.

Wang, N., Chen, L., and Yang, Y. (2020). The impacts of item features and user characteristics on users' perceived serendipity of recommendations. In *Proceedings of the 28th ACM User Modeling, Adaptation and Personalization*, UMAP' 20, pages 266–274, New York, NY, USA. Association for Computing Machinery (ACM).

Wojtowicz, Z. and Loewenstein, G. (2020). Curiosity and the economics of attention. *Current Opinion in Behavioral Sciences*, 35:135–140.

Wu, C., Wu, F., An, M., Qi, T., Huang, J., Huang, Y., and Xie, X. (2019a). Neural news recommendation with heterogeneous user behavior. In *Proceedings of the 2019 EMNLP-IJCNLP*, pages 4874–4883, Hong Kong, China. Association for Computational Linguistics.

Wu, Q., Liu, S., and Miao, C. (2017). Modeling uncertainty driven curiosity for social recommendation. In *Proceedings of the International Conference on Web Intelligence*, WI '17, page 790–798, New York, NY, USA. Association for Computing Machinery (ACM).

Wu, Q., Liu, S., Miao, C., Liu, Y., and Leung, C. (2016). A social curiosity inspired recommendation model to improve precision, coverage and diversity. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 240–247. IEEE.

Wu, Q. and Miao, C. (2013a). Curiosity: From psychology to computation. *ACM Computing Surveys*, 46(2):18:1–18:26.

Wu, Q. and Miao, C. (2013b). Modeling curiosity-related emotions for virtual peer learners. *IEEE Computational Intelligence Magazine*, 8(2):50–62.

Wu, Q., Miao, C., and An, B. (2014). Modeling curiosity for virtual learning companions. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS '14, pages 1401–1402, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Wu, Q., Miao, C., and Shen, Z. (2012). A curious learning companion in virtual learning environment. In *IEEE International Conference on Fuzzy Systems*, pages 1–8, Australia. IEEE.

Wu, X., Fu, L., Meng, J., and Wang, X. (2019b). Maximizing influence diffusion over evolving social networks. In *Proceedings of the Fourth International Workshop on Social Sensing*, SocialSense'19, page 6–11, New York, NY, USA. Association for Computing Machinery (ACM).

Wundt, W. M. (1874). *Grundzude Physiologischen Psychologie*, volume 1 of *Duxbury Classic*. Wilhelm Engelmann, Leipzig, Germany.

Xu, K., Mo, J., Cai, Y., and Min, H. (2019). Enhancing recommender systems with a stimulus-evoked curiosity mechanism. *IEEE Trans. on Knowledge and Data Engineering*.

Xu, K., Mo, J., Cai, Y., and Min, H. (2021). Enhancing recommender systems with a stimulus-evoked curiosity mechanism. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2437–2451.

Yates, A., Nogueira, R., and Lin, J. (2021). Pretrained transformers for text ranking: Bert and beyond. In *Proc. of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, pages 1154–1156, New York, NY, USA. Association for Computing Machinery.

Youyou, W., Kosinski, M., and Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *National Academy of Sciences*, 112(4):1036–1040.

Zhang, C., Gao, S., Tang, J., Liu, T. X., Fang, Z., and Cheng, X. (2016). Learning triadic influence in large social networks. In *IEEE/ACM ASONAM*, pages 1380–1381, Davis, California, USA. IEEE/ACM.

Zhang, L., Wang, T., Jin, Z., Su, N., Zhao, C., and He, Y. (2018). The research on social networks public opinion propagation influence models and its controllability. *China Communications*, 15(7):98–110.

Zhao, P. and Lee, D. L. (2016). How much novelty is relevant?: It depends on your curiosity. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 315–324, New York, NY, USA. Association for Computing Machinery (ACM).

Zhu, Y., Tang, J., and Tang, X. (2020). Pricing influential nodes in online social networks. *Proceedings of the VLDB Endowment*, 13(10):1614–1627. ISSN 2150-8097.

Zurn, P. and Bassett, D. S. (2018). On curiosity: A fundamental aspect of personality, a practice of network growth. *Personality Neuroscience*, 1:e13.