# Analyzing and Modeling User Curiosity in Online Information Services

Alexandre Magno Sousa

**Advisors**

Prof. Jussara Almeida

Prof. Flavio Figueiredo

UFMG

UNIVERSIDADE FEDERAL
DE MINAS GERAIS

# Roadmap

Contextualization and Motivation

**Goals**, Hypothesis and Research Questions

Related Work

**Problem Statement**: **Elements** and **Assumptions**

Case Studies and Contributions

Conclusions and Next Steps

# Roadmap

**Contextualization and Motivation**

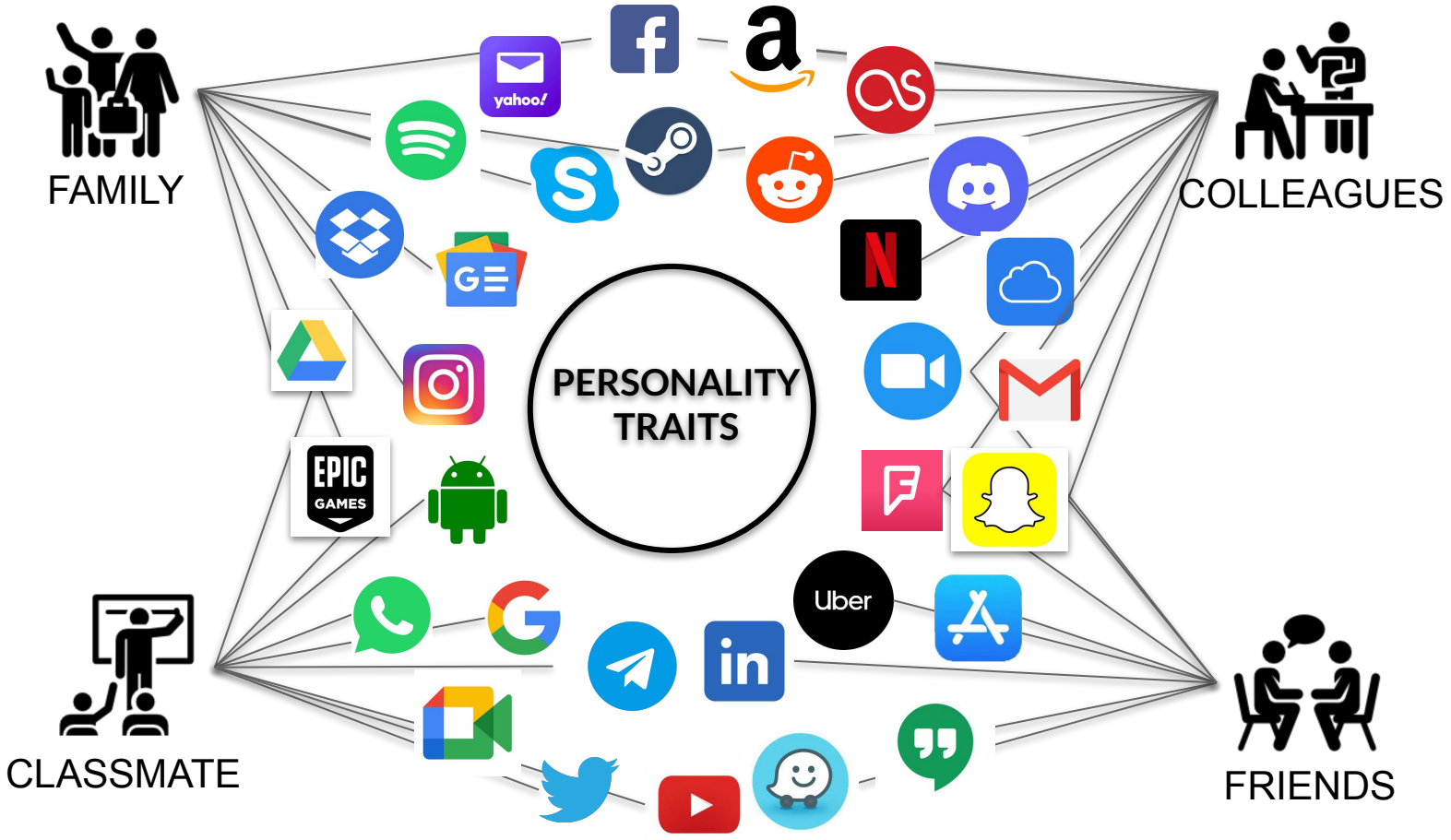
Goals, Hypothesis and Research Questions

Related Work

Problem Statement: Elements and Assumptions

Case Studies and Contributions

Conclusions and Next Steps


FAMILY
COLLEAGUES
PERSONALITY TRAITS
CLASSMATE
FRIENDS
yahoo!
EPIC GAMES
Uber

# Curiosity

Some types of ***personality traits*** are *associated* with ***human curiosity***.

***Critical factor*** that *influences human behavior* in positive and negative ways at all stages of live.

***Modeling*** *and* ***analysis*** *of human curiosity as* ***driven force*** *behind online user behavior deserves further investigation*.

*"**Curiosity** has been recognized as the **desire for information** which facilitates learning, promote new discoveries and enriches life of knowledge"*

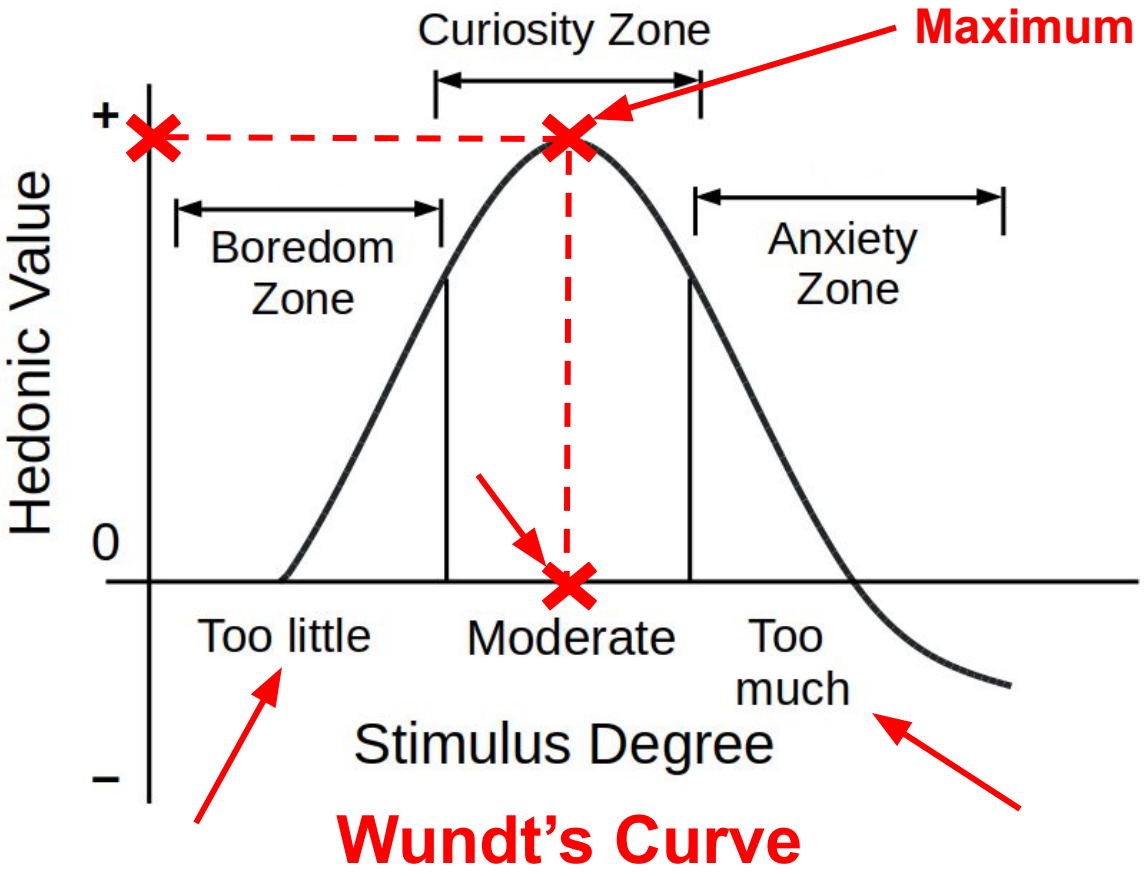
[Oudeyer et al., 2016]

***Our focus**: is on the **curiosity** of human beings and its role in **online information dissemination**.*

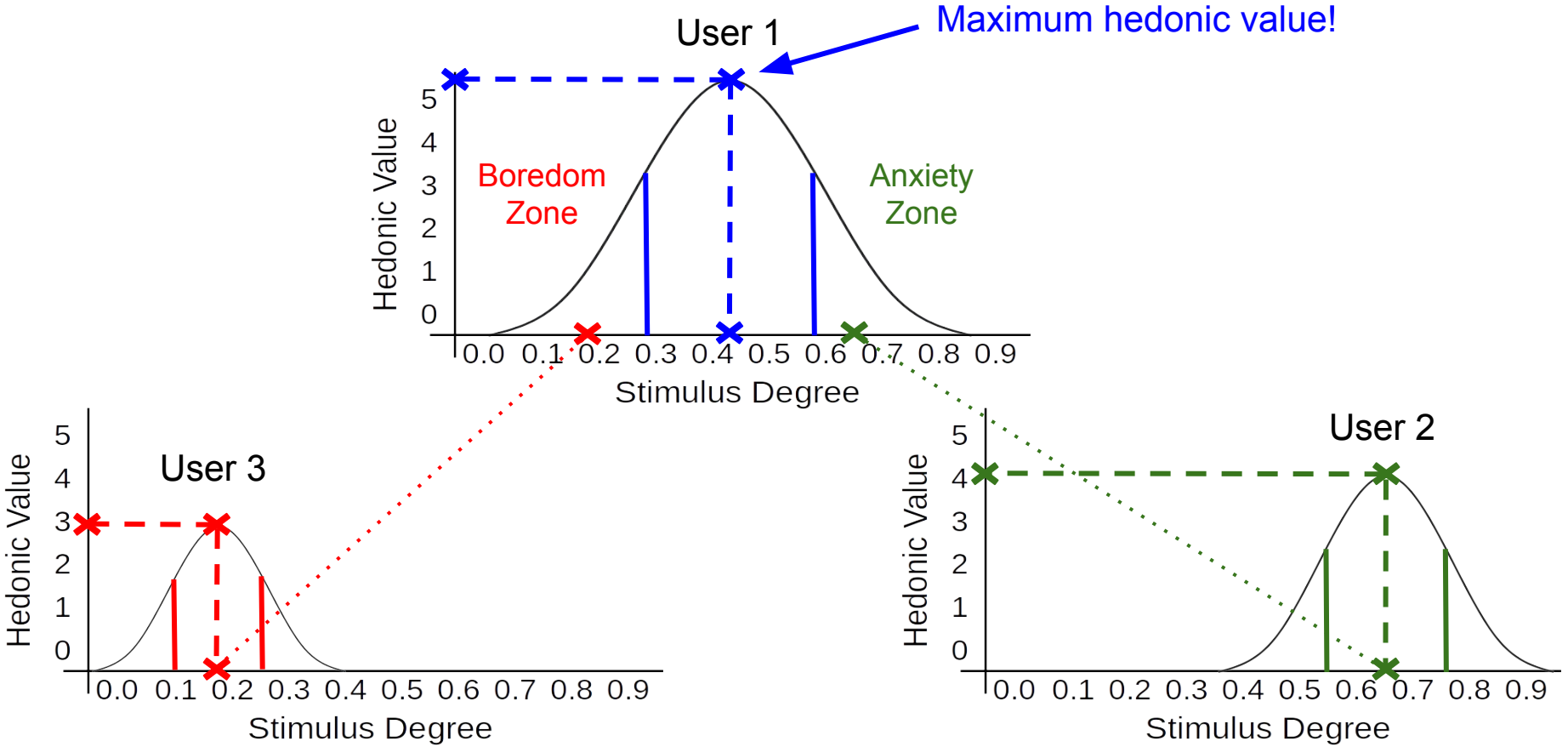
# Guiding Question

*"Can models capturing* ***multiple facets of human curiosity*** *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****?"*

# Optimal Level of Stimulation [Wundt, 1874]



# Optimal Level of Stimulation



# Collative Variables

Different collative variables govern the curiosity stimulation process [Berlyne, 1960]:

1. ***Novelty*** refers ***how new*** the stimulus experienced is.
2. ***Uncertainty*** is related to the ***difficulty*** in deciding ***how to respond*** *to* a stimulus.
3. ***Complexity*** which refers to the ***diversity*** in a stimulus pattern.
4. ***Conflict*** occurs when the ***same stimulus triggers multiple*** *incompatible responses*.

# Information Theory

**Berlyne** propose a *methodology to quantify* **collative variables** based on information theoretical metrics [Berlyne, 1960].

**Silvia** states that *with some algebraic manipulation* all **collative variables** can be expressed by metrics from information theory [Silvia, 2006].

These arguments inspired some *recent studies* in Computer Science to propose *metrics capturing* different **collative variables**.

A collative variable can be captured by different metrics (different aspects).

# Social Curiosity

***Social curiosity*** denotes individual skills to tackle the interpersonal world.

- *It is the interesting in obtaining new information about how others think, behave or act.*

Recently introduced as a *key component* of 5DCR [Kashdan et al., 2018].

We argue that ***in addition*** to the traditional collative variables, ***social influence*** *should be considered as part of stimulus to curiosity*.

# Roadmap

Contextualization and Motivation

**Goals, Hypothesis and Research Questions**

Related Work

Problem Statement: Elements and Assumptions

Case Studies and Contributions

Conclusions and Next Steps

# 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***.

# 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?

# Roadmap

Contextualization and Motivation

Goals, Hypothesis and Research Questions

**Related Work**

Problem Statement: Elements and Assumptions

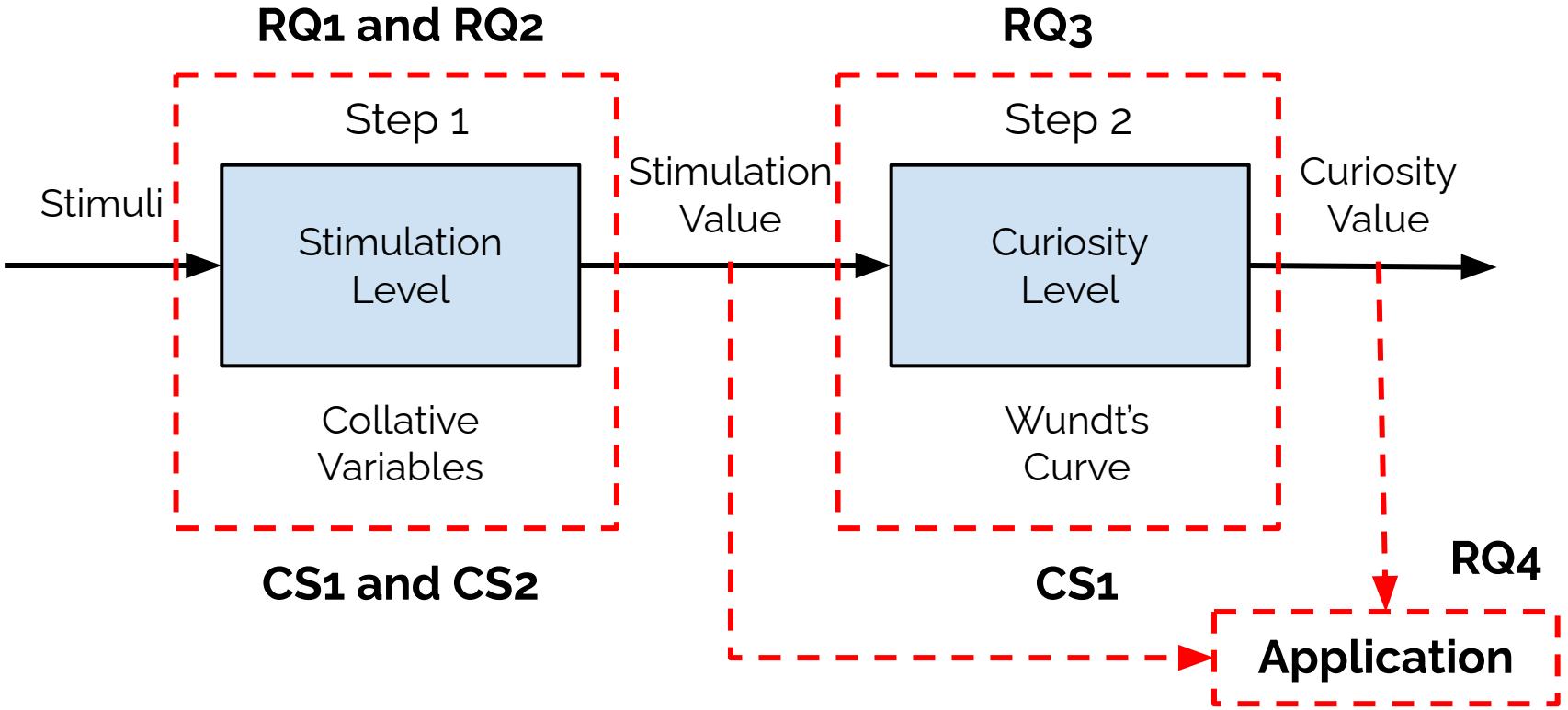
Case Studies and Contributions

Conclusions and Next Steps

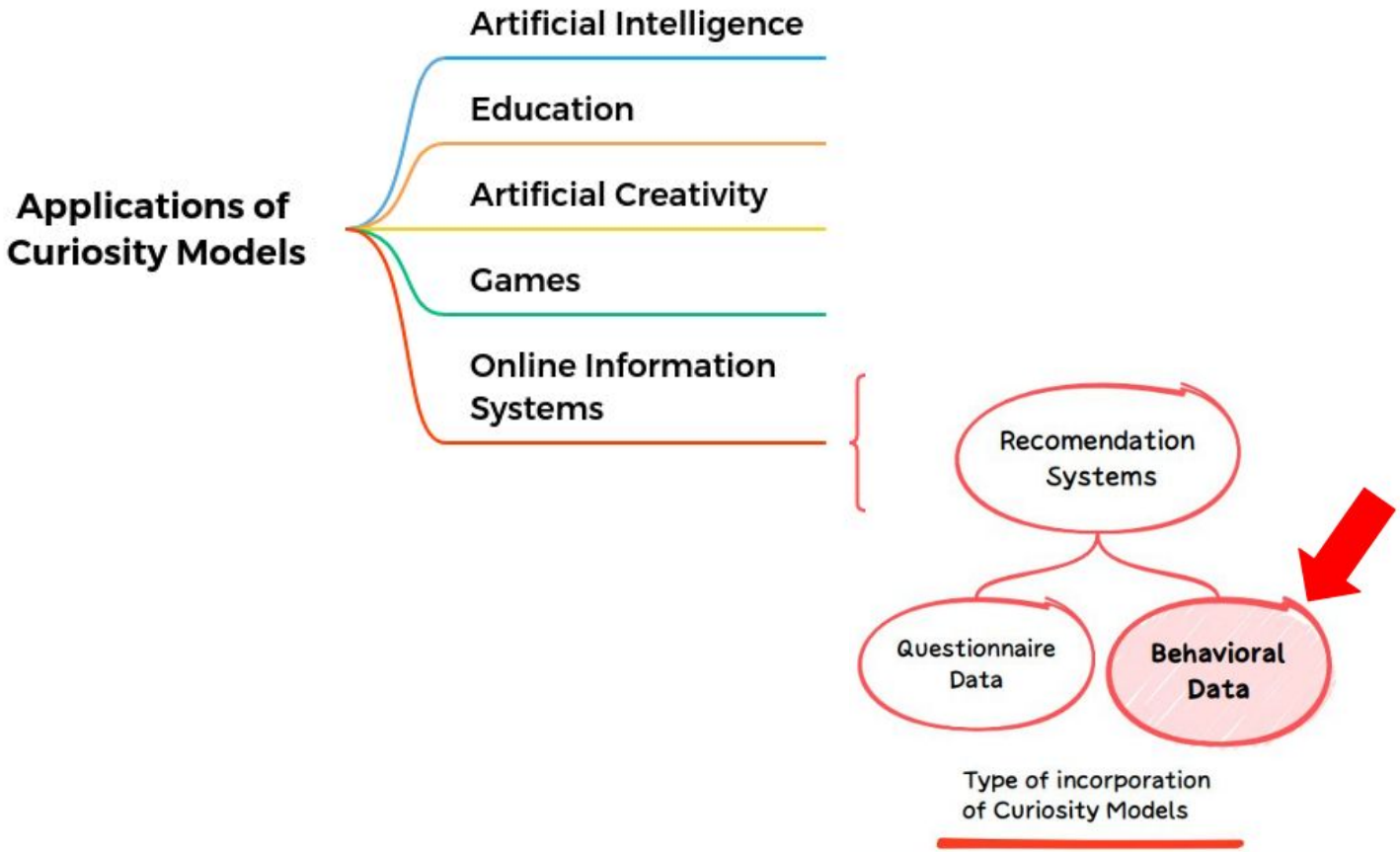
# Computational Models of Curiosity

# General Appraisal Process (GAP)

[Wu and Miao, 2013]

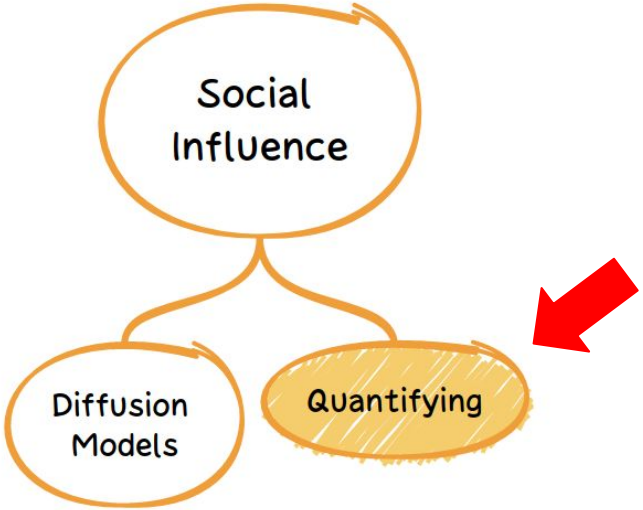


# Computational Models of Curiosity



# Information Dissemination

- *Social influence* may also be a **component** of curiosity stimulation.




- **Prior efforts** to model social influence as a stimulus to user curiosity *are quite rare* in the literature.

# Roadmap

Contextualization and Motivation

Goals, Hypothesis and Research Questions

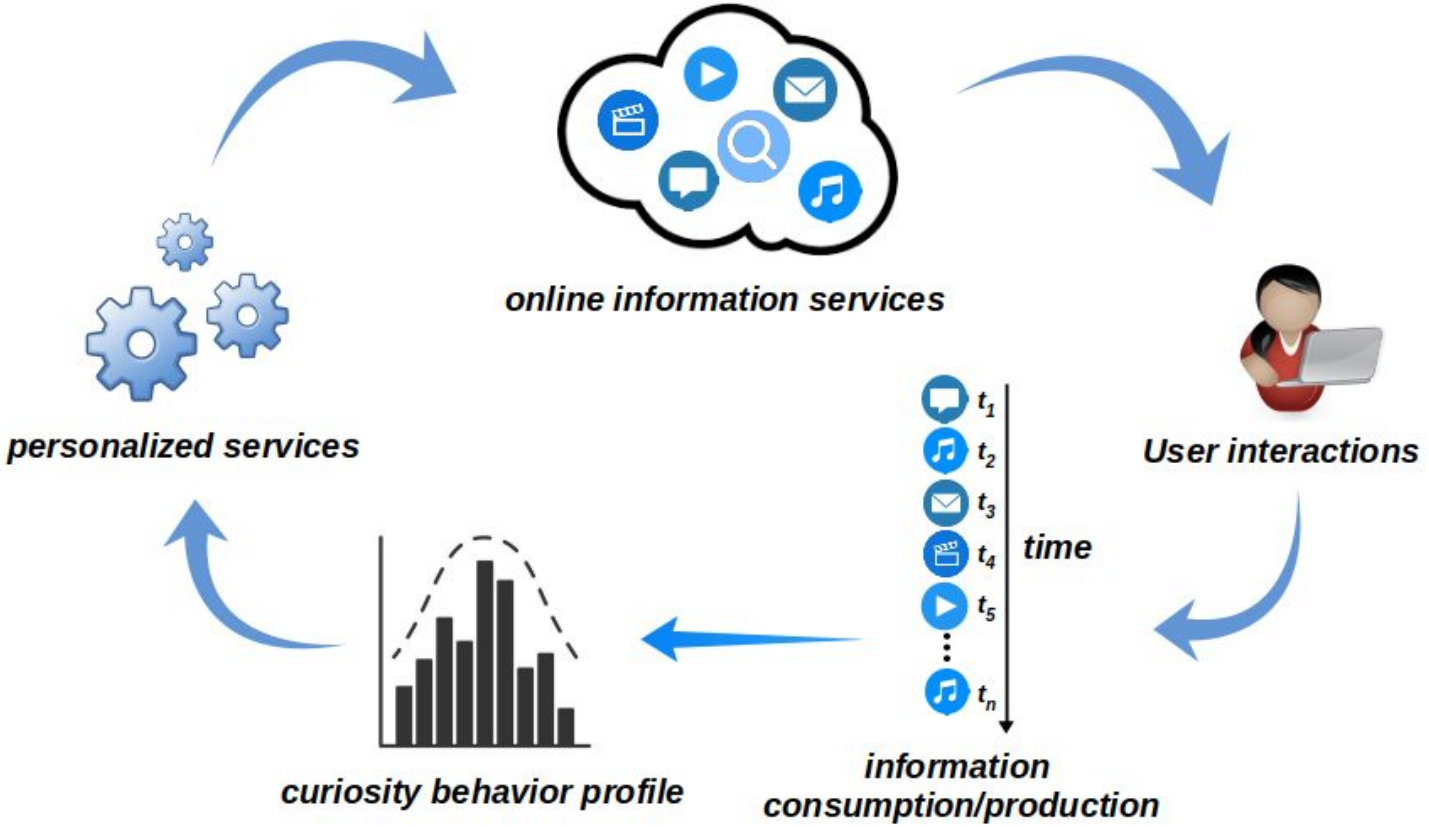
Related Work

**Problem Statement: Elements and Assumptions**

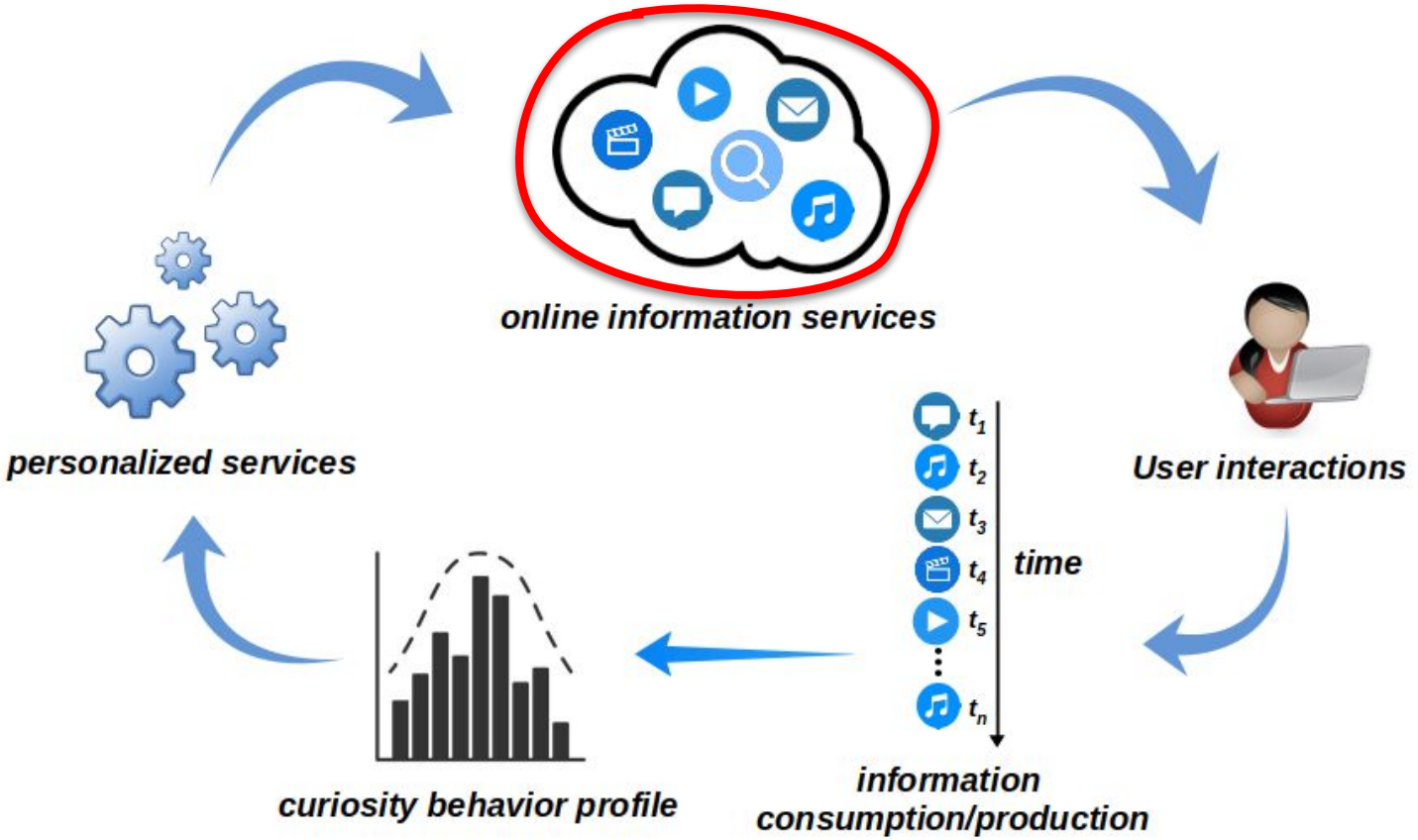
Case Studies and Contributions

Conclusions and Next Steps

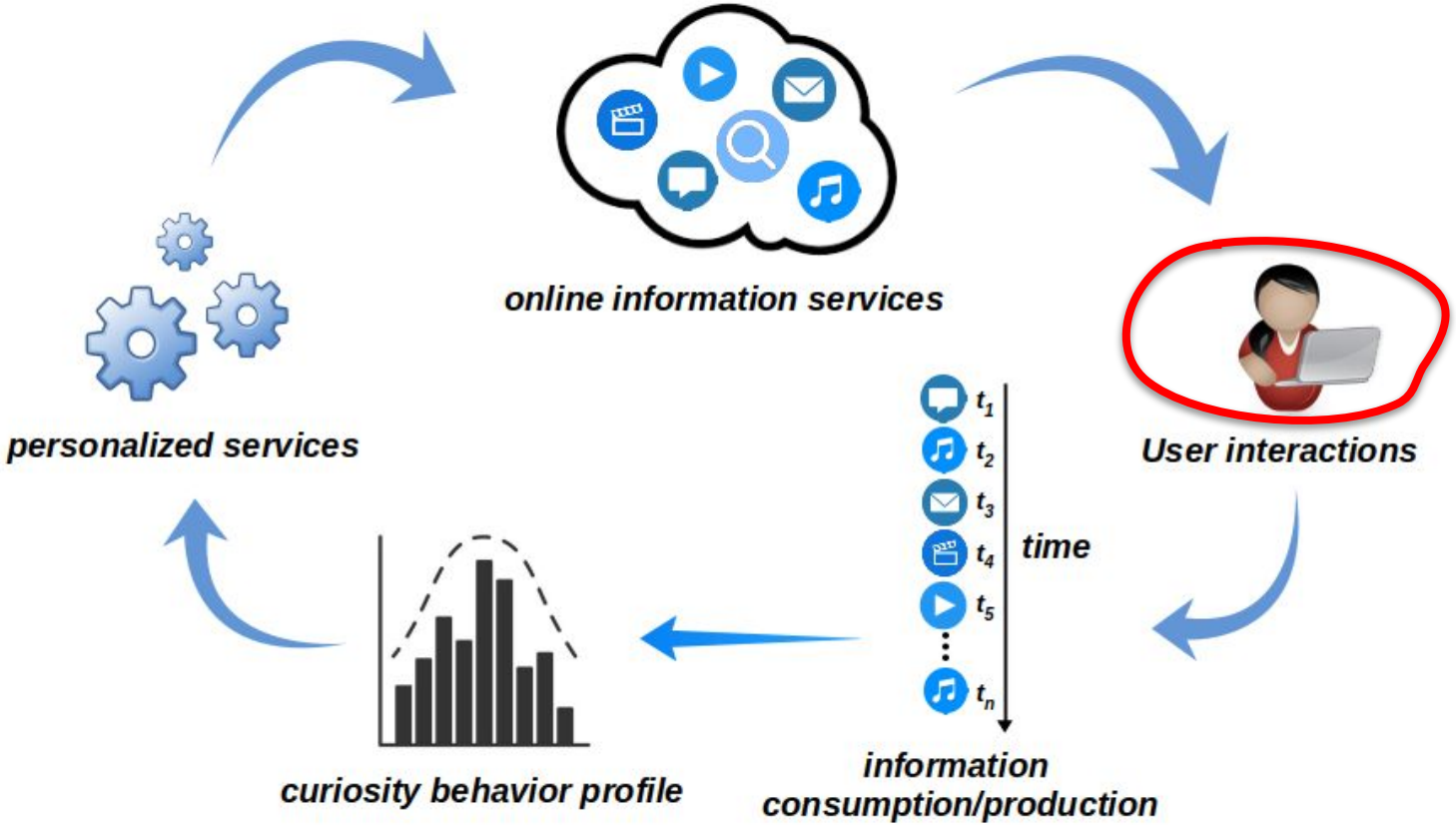
# Overview


online information services
User interactions
time
$t_1$
$t_2$
$t_3$
$t_4$
$t_5$
$t_n$
information consumption/production
curiosity behavior profile
personalized services

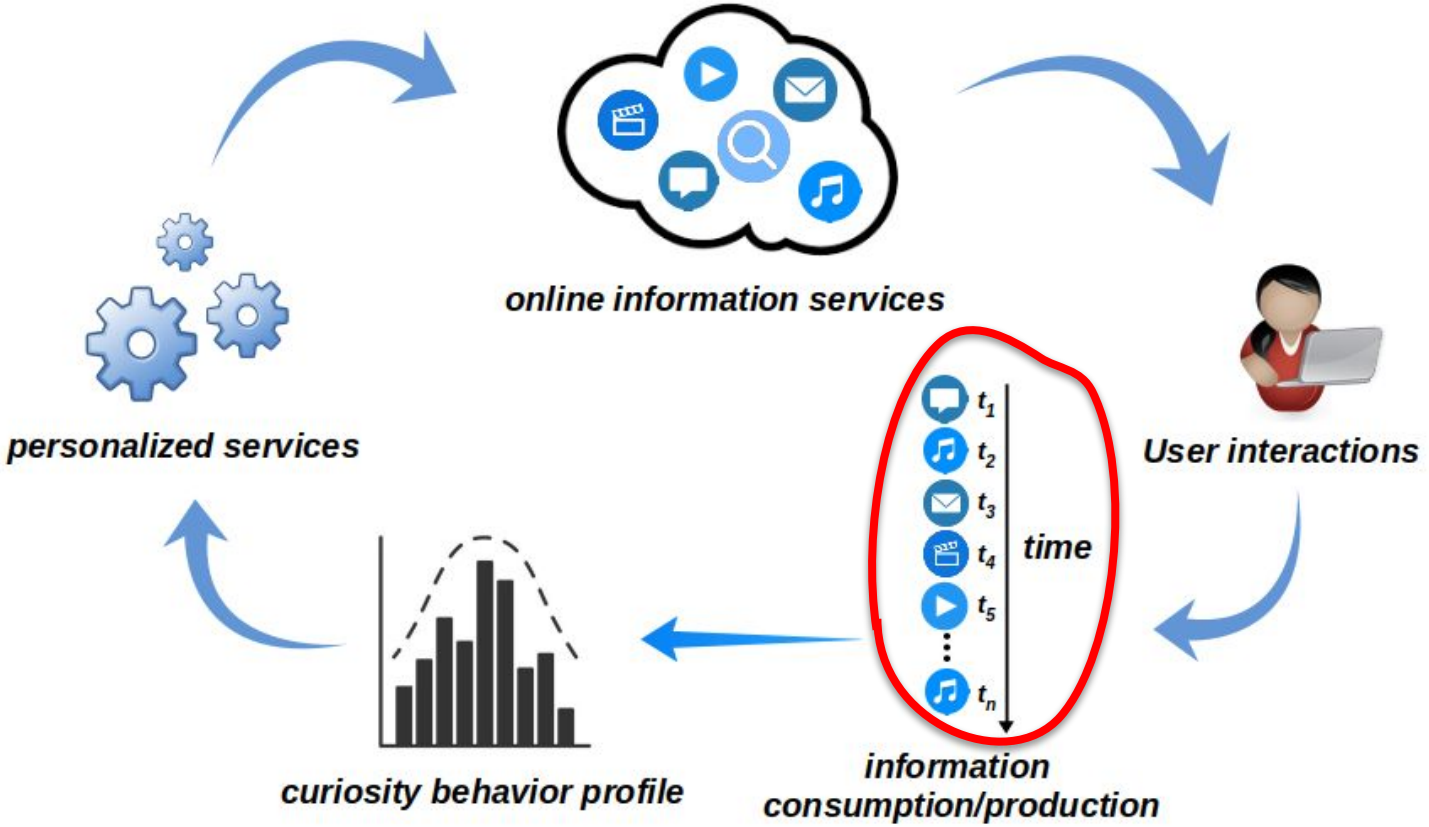
# Overview


online information services
User interactions
time
$t_1$
$t_2$
$t_3$
$t_4$
$t_5$
$t_n$
information consumption/production
curiosity behavior profile
personalized services

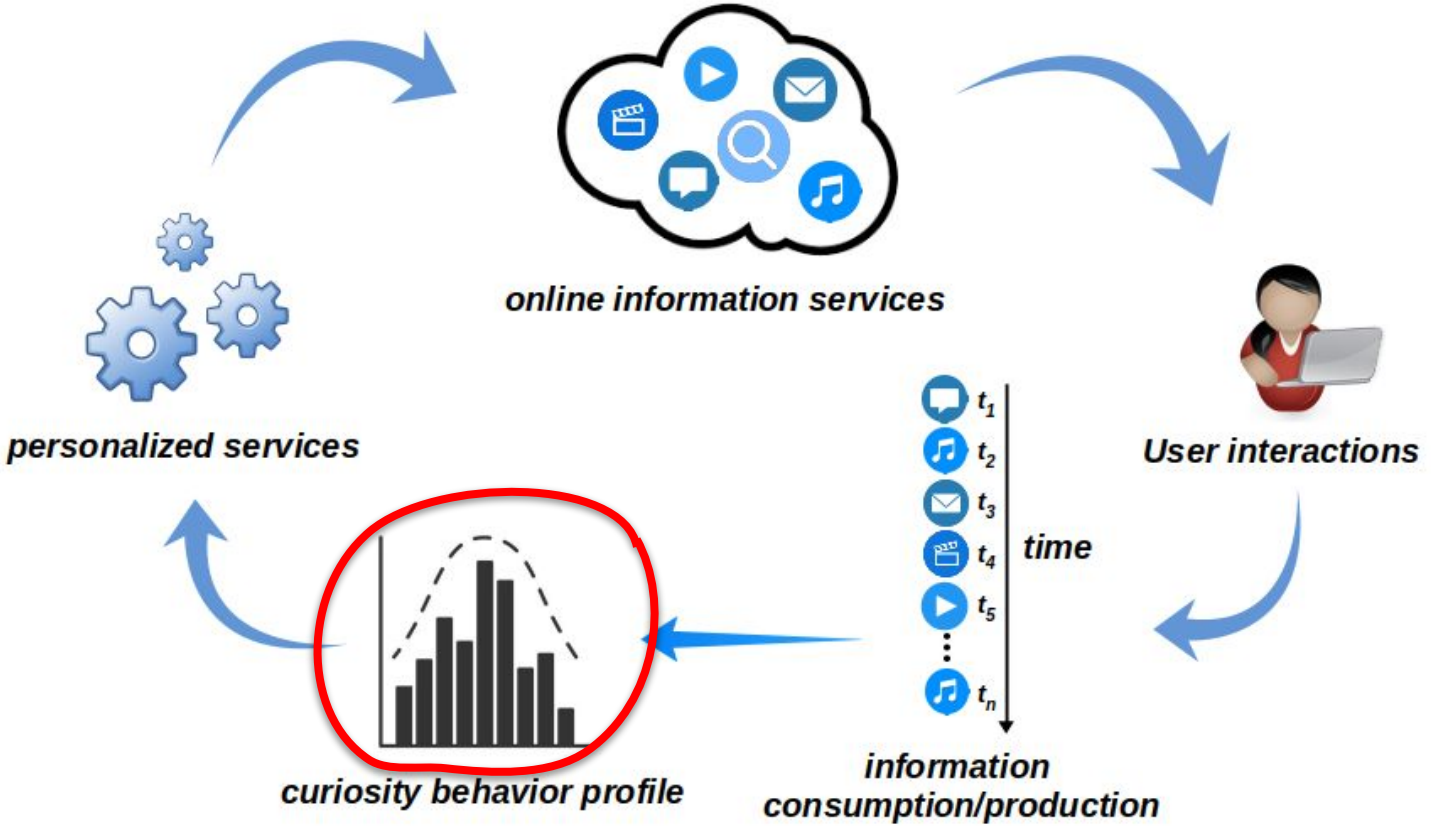
# Overview


online information services
User interactions
time
$t_1$
$t_2$
$t_3$
$t_4$
$t_5$
$t_n$
information consumption/production
curiosity behavior profile
personalized services

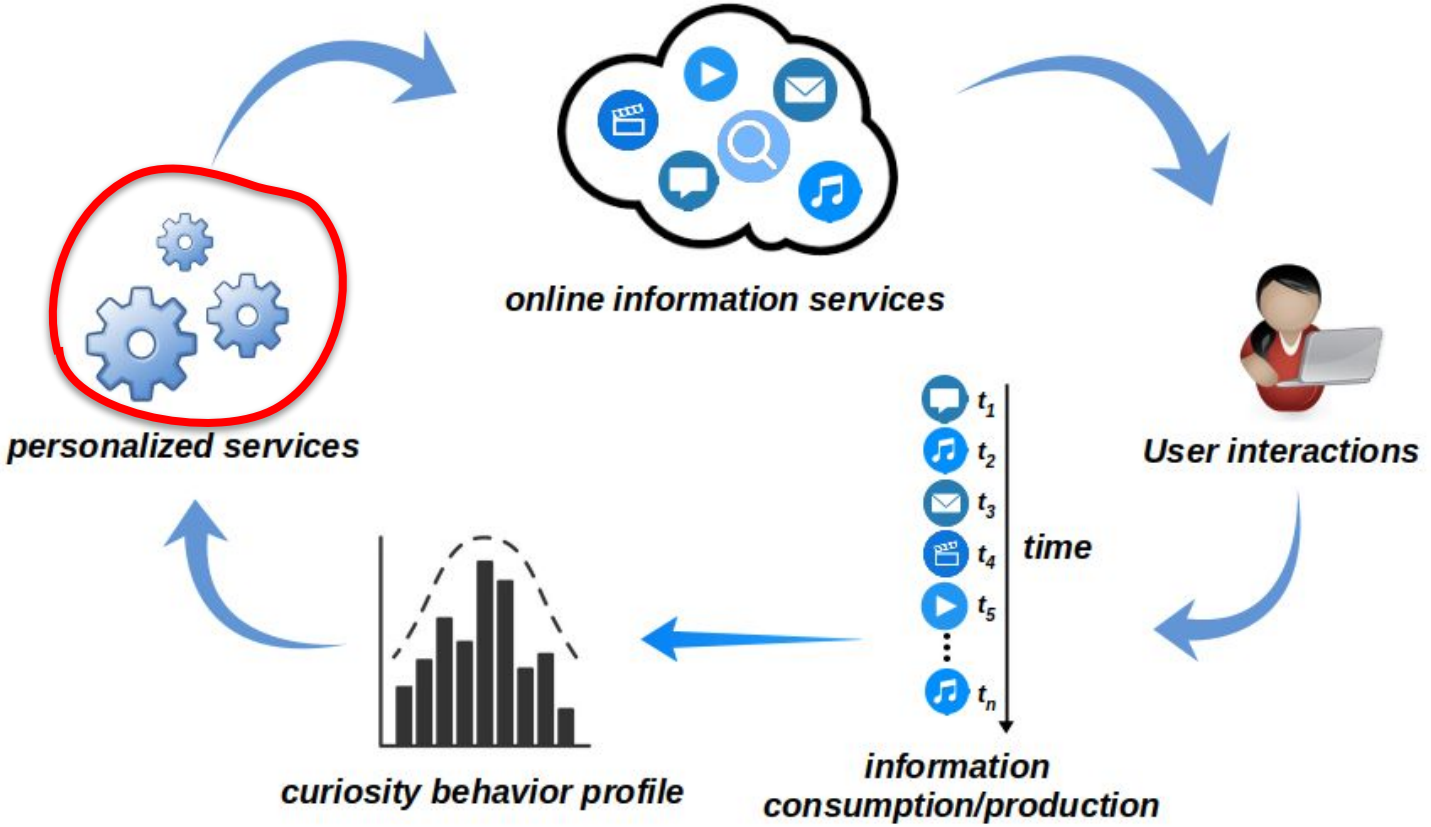
# Overview


online information services
User interactions
time
$t_1$
$t_2$
$t_3$
$t_4$
$t_5$
$t_n$
information consumption/production
curiosity behavior profile
personalized services

# Overview


online information services
User interactions
time
$t_1$
$t_2$
$t_3$
$t_4$
$t_5$
$t_n$
information consumption/production
curiosity behavior profile
personalized services

# Overview


online information services
User interactions
time
$t_1$
$t_2$
$t_3$
$t_4$
$t_5$
$t_n$
information consumption/production
curiosity behavior profile
personalized services

# Overview

Online Information Services

*Information Consumption*

*User interactions*

*Information Production*

# Fundamental Elements

***Content items*** of information that are *produced*, *consumed or shared*.

***Users*** who interact with the information system by consuming or producing (sharing) content items.

***User actions*** are the ***different types of actions events*** which a user may have while contributing to information spread.

***History of user action events*** ***recorded by system*** as logs of user interactions over time.

# Case Studies

We focus on **two case studies** covering platforms with *different features* which *foster different types* of user interactions.

- The two case studies cover **different and complementary perspectives** of information dissemination process.

***Case Study 1***: we investigate the role of user curiosity as a driving force behind users listening online music in ***LastFM*** (*information consumption*).

***Case Study 2***: we investigate how one's curiosity is stimulate by actions of other users in absence of explicit links in ***Whatsapp*** (*information production*).

# Case Study 1: LastFM

(1) content item

| | Song | Artist | Timestamp |
|---|---|---|---|
| ▶ | **9**. Have a Drink on me | AC/DC | 12:00 |
| ▶ | **8**. Live and Let Die | Guns N'Roses | 11:30 |
| ▶ | **7**. Patience | Guns N'Roses | 11:20 |
| ▶ | **6**. Live and Let Die | Guns N'Roses | 11:05 |
| ▶ | **5**. Don't cry (original) | Guns N'Roses | 10:55 |
| ▶ | **4**. November Rain | Guns N'Roses | 10:45 |
| ▶ | **3**. Black in Black | AC/DC | 10:30 |
| ▶ | **2**. Welcome to The Jungle | Guns N'Roses | 10:25 |
| ▶ | **1**. T.N.T. | AC/DC | 10:15 |

(a) Songs listened by User 1 in online music consumption.

# Case Study 1: LastFM

(2) categorization

**Musical Genres of current song**

Rock Hard rock Classic rock 90s

| | Song | Artist | Timestamp |
|---|---|---|---|
| ▶ | **9**. Have a Drink on me | AC/DC | 12:00 |
| ▶ | **8**. Live and Let Die | Guns N'Roses | 11:30 |
| ▶ | **7**. Patience | Guns N'Roses | 11:20 |
| ▶ | **6**. Live and Let Die | Guns N'Roses | 11:05 |
| ▶ | **5**. Don't cry (original) | Guns N'Roses | 10:55 |
| ▶ | **4**. November Rain | Guns N'Roses | 10:45 |
| ▶ | **3**. Black in Black | AC/DC | 10:30 |
| ▶ | **2**. Welcome to The Jungle | Guns N'Roses | 10:25 |
| ▶ | **1**. T.N.T. | AC/DC | 10:15 |

(a) Songs listened by User 1 in online music consumption.

# Case Study 1: LastFM


(a) Songs listened by User 1 in online music consumption.

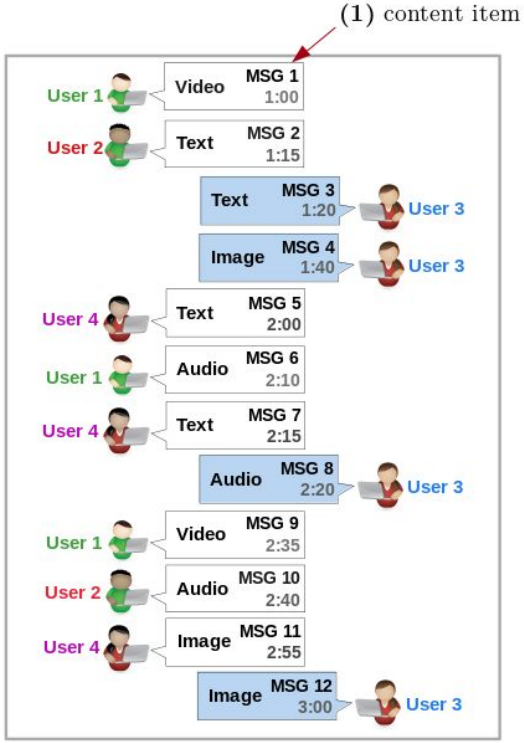
# Case Study 1: LastFM

| | Song | Artist | Timestamp |
|---|---|---|---|
| ▶ | **9**. Have a Drink on me | AC/DC | 12:00 |
| ▶ | **8**. Live and Let Die | Guns N'Roses | 11:30 |
| ▶ | **7**. Patience | Guns N'Roses | 11:20 |
| ▶ | **6**. Live and Let Die | Guns N'Roses | 11:05 |
| ▶ | **5.** Don't cry (original) | Guns N'Roses | 10:55 |
| ▶ | **4.** November Rain | Guns N'Roses | 10:45 |
| ▶ | **3**. Black in Black | AC/DC | 10:30 |
| ▶ | **2**. Welcome to The Jungle | Guns N'Roses | 10:25 |
| ▶ | **1**. T.N.T. | AC/DC | 10:15 |

**(4)** history of user action events.

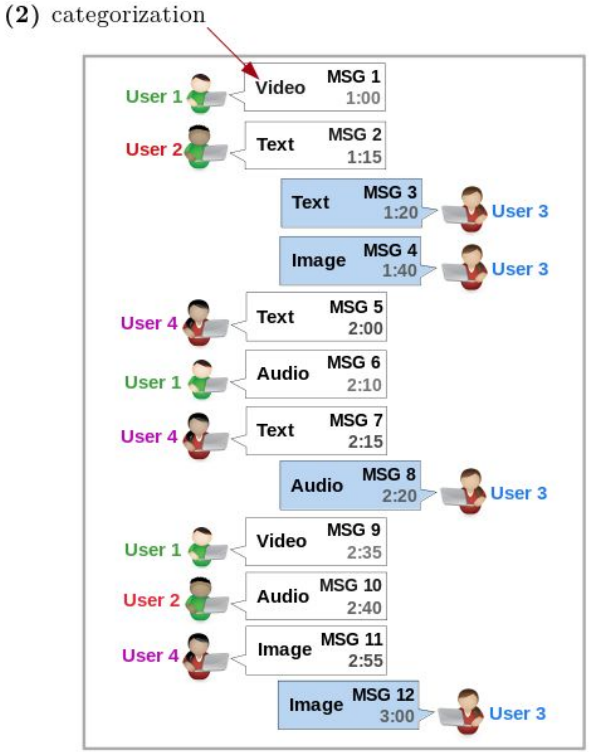
(a) Songs listened by User 1 in online music consumption.

# Case Study 2: WhatsApp



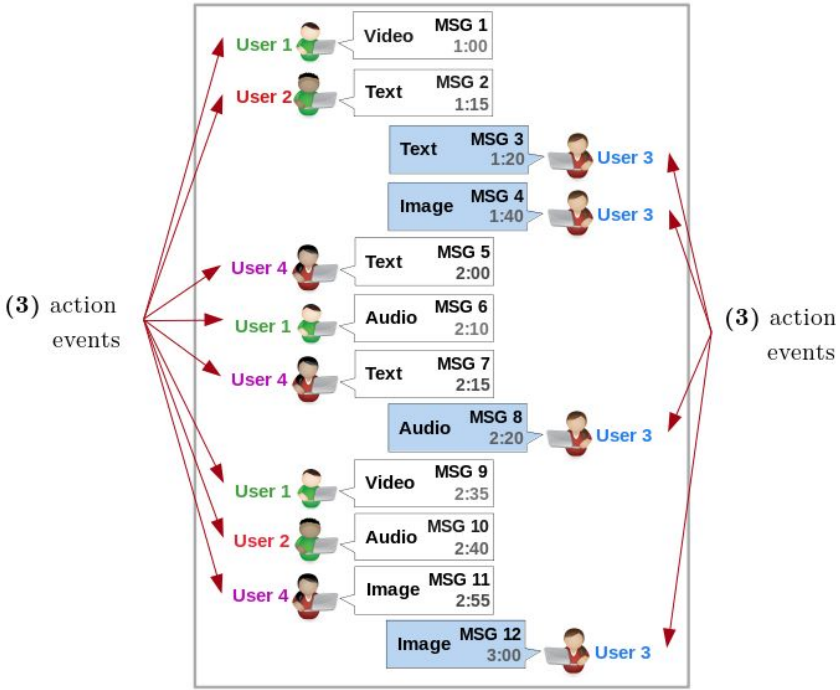

**(a)** Messages Shared by 4 members of a group.

# Case Study 2: WhatsApp



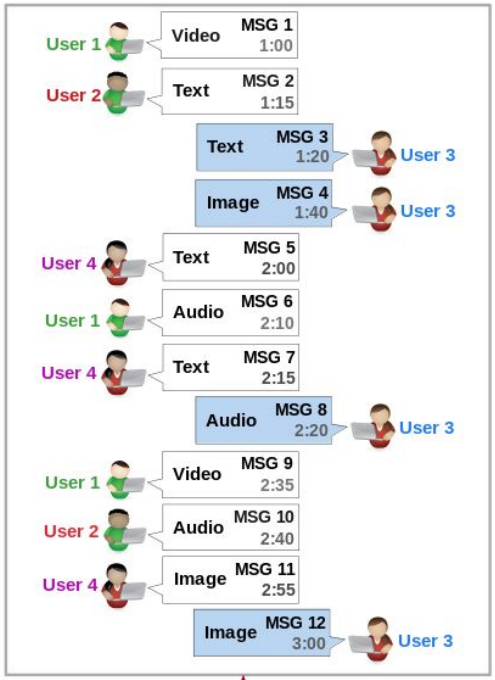

**(a)** Messages Shared by 4 members of a group.

# Case Study 2: WhatsApp




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

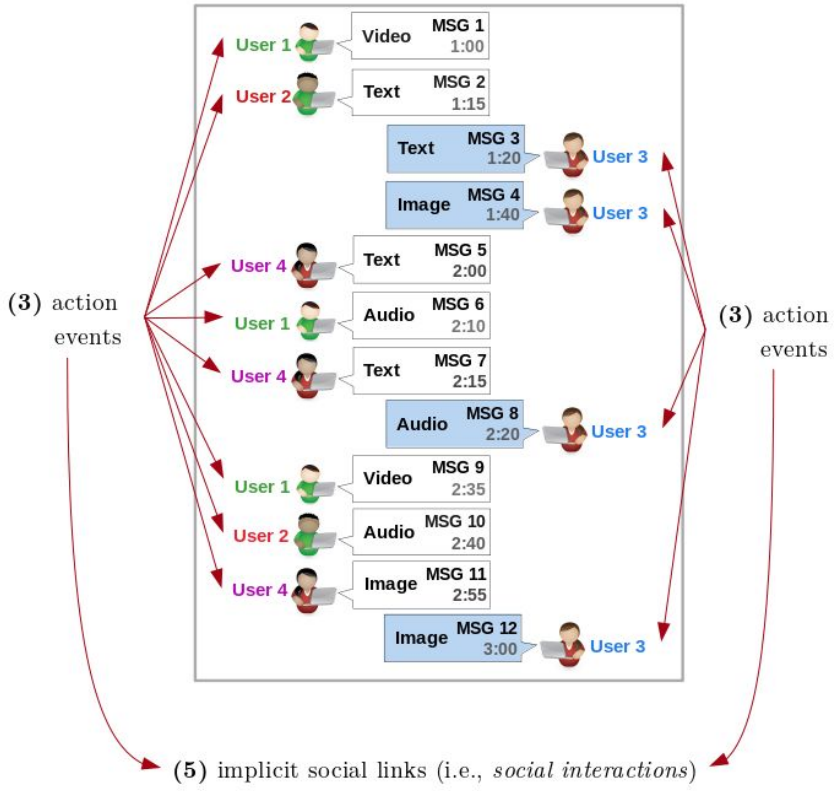
# Case Study 2: WhatsApp




**(4)** history of user action events.

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

# Case Study 2: WhatsApp




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

# Assumptions

1. The curiosity of a user may be *triggered differently* depending on various features of the ***target platform*** and the ***types*** of user actions.

2. The ***way*** that the curiosity of a user ***reacts*** to a given ***stimulus*** may *change over time*.

3. The curiosity driving an action by a given user ***u*** at a given time ***t*** has a ***period of activation*** $\delta_T$ (*window of interaction*).

# Additional Assumptions for Social Curiosity

4. The curiosity of a user may be **triggered differently** depending on the *people with whom is interacting* and the *ongoing discussions* among them.

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*.

6. The extent to which user's curiosity is *stimulated by other users* can be estimated by *historical patterns*.

# Problem Definition

*Given a set of users* $\mathcal{U}$ *of an information 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 on a content item* $i \in \mathcal{I}$*, characterized by (one or more) categories* $c \in \mathcal{C}$*, which serves as proxy for representing content properties.*

***chronologically ordered tuple!***

$(u, i, \mathcal{C}_s, t)$, where $\mathcal{C}_s \subseteq \mathcal{C}$.

# Roadmap

Contextualization and Motivation

Goals, Hypothesis and Research Questions

Related Work

Problem Statement: Elements and Assumptions

**Case Studies and Contributions**

Conclusions and Next Steps

# Case Study 1

## Analyzing and Modeling User Curiosity in **Online Content Consumption**

# Contextualization

1. We investigate the role of user ***curiosity*** as a *driving force* behind users listening online music in ***LastFM*** (i.e., *information consumption*).
   - ***musical tastes*** *are related to personality characteristics* [Greasley et al., 2013].
2. All four basic *collative variables* are taken as **sources of stimulus** to curiosity.
3. Investigation whether a *single Wundt's curve* is a **reasonably model** of user curiosity.

Focus only on **RQ1** and **RQ3**.

# Metrics for Collative Variables

- We **propose metrics** that *capture different aspects* of a stimuli which ***govern*** how the curiosity in ***information consumption*** is stimulated.

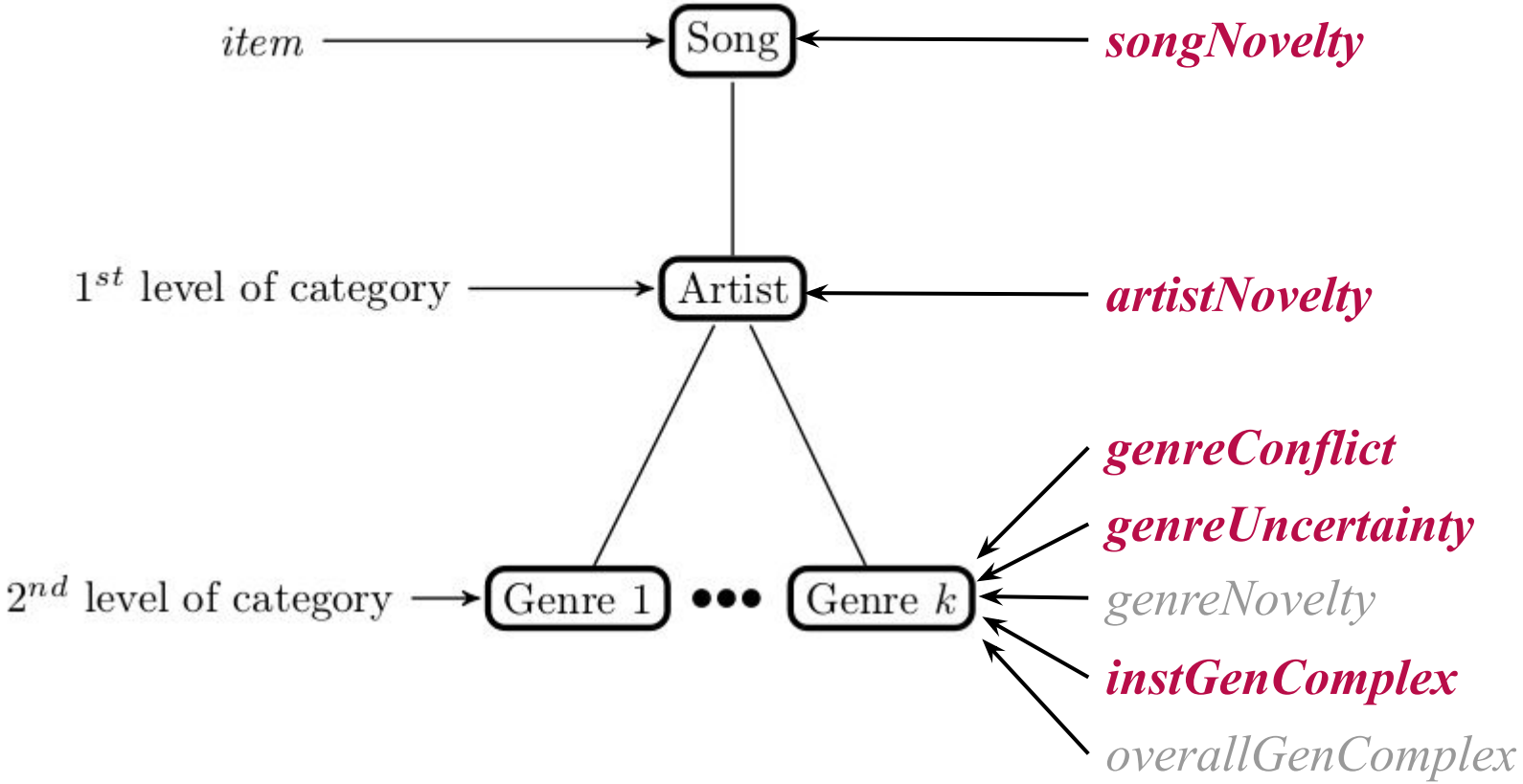
- We make use of **measures** from ***information theory*** to *capture different collative variables* associated with **curiosity stimulation**:

  - *Proposal of 7 metrics for all components of collative variables (novelty, uncertainty, conflict and complexity).*

  - *They are based on concept of **Entropy** and **Surprisal**.*

[Cover and Thomas, 2006; MacKay, 2005]

# Content Categorization & Proposed Metrics



# Proposed Metrics

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

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

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

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

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

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

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

# Dataset

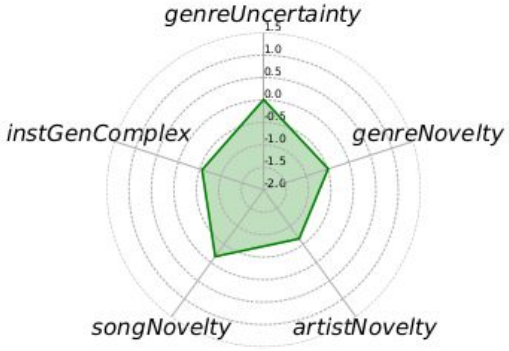
**LFM-1B** Dataset [Schedl, 2016]:

- Capturing the ***listening behavior*** of users.
  - *84K users, 219K artists, 243M listening events.*
- **Period** from *January 2013 to August 2014*.
- ***Action event*** recorded: *user*, *artist*, *album*, *track* and *timestamp*.
- Incorporation of ***artist's genre*** [Schedl and Ferwerda, 2017].

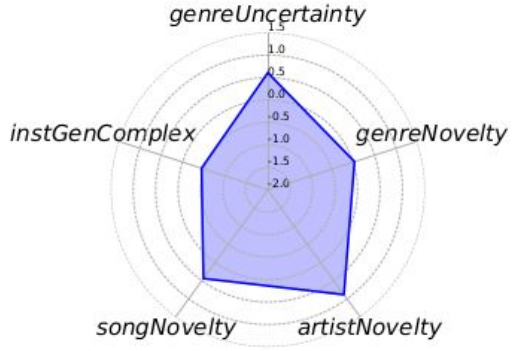
***RQ1**: Are there distinct user behavior profiles in terms of curiosity stimuli, as captured by **multiple collative variables**?*

# Clustering by Access Profile

32%



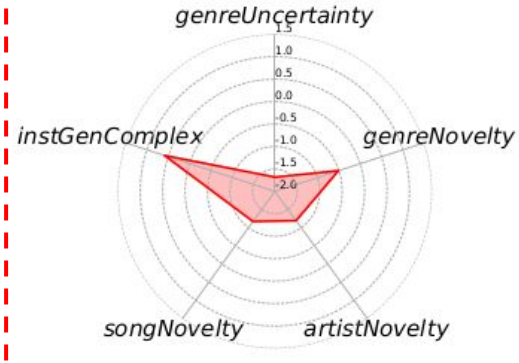

(a) $C_0$: Underground.



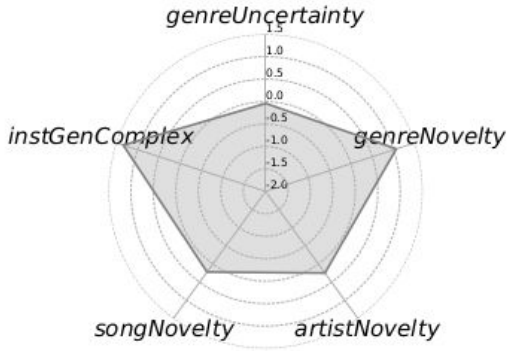

(b) $C_1$: Eclectic.

35%

12%



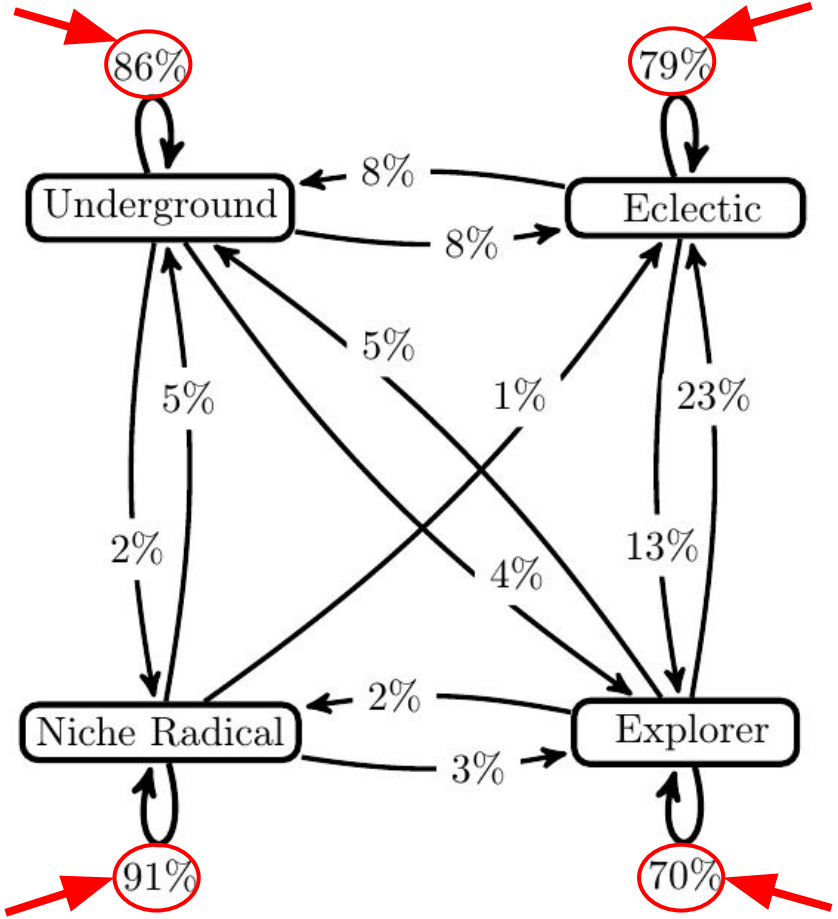

(c) $C_2$: Niche Radical.




(d) $C_3$: Explorer.

20%

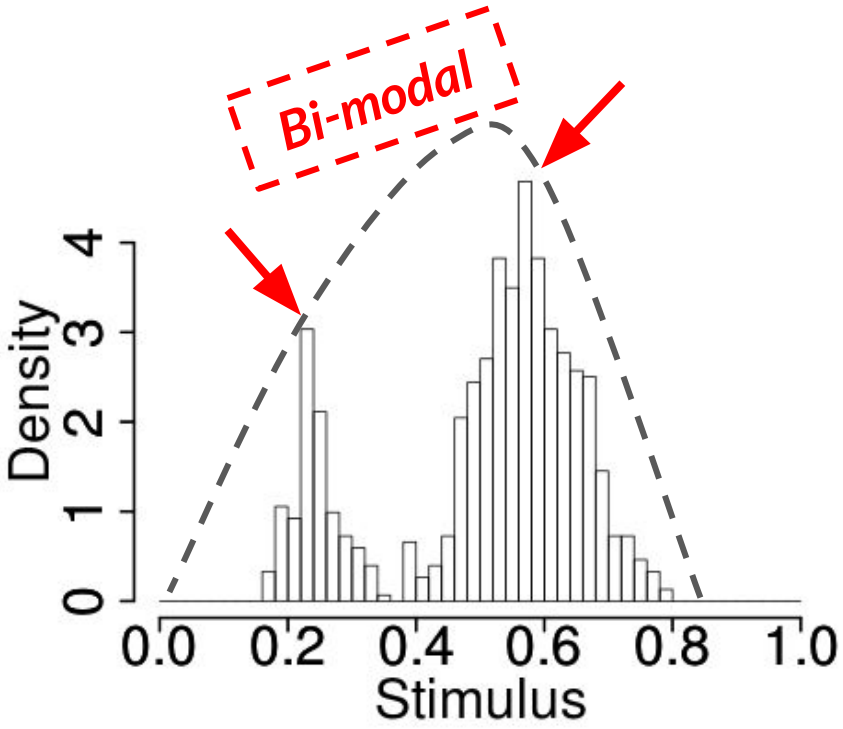
# Dynamic of User Behavior

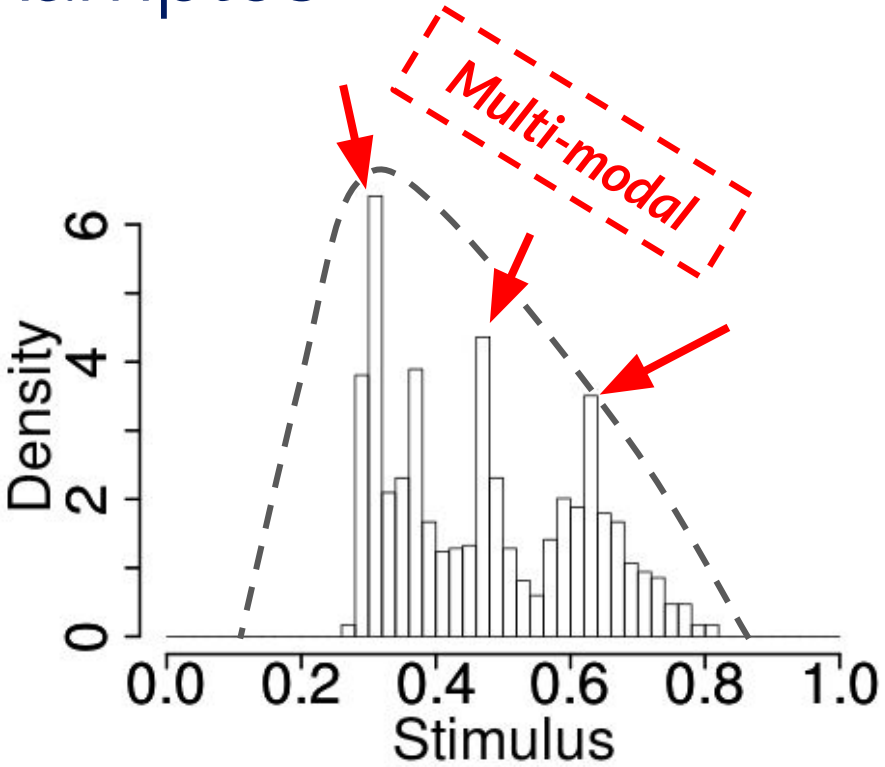


***RQ3**: To which extent, user curiosity driving online information dissemination can be accurately modeled by a **Wundt's curve**?*

**Prior work**: assumption of ***single peak*** of curiosity and suggested use of a ***single beta distribution***.

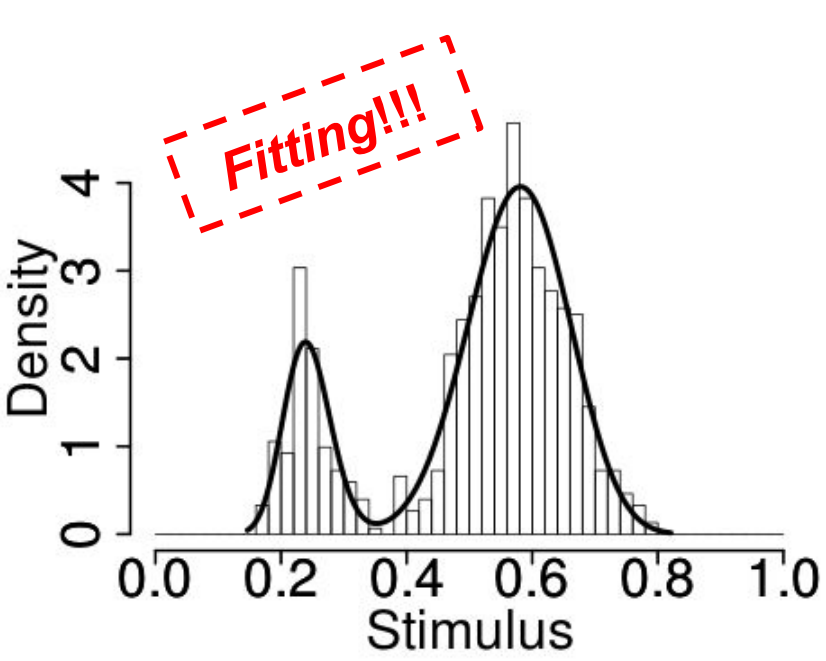
# Real Curiosity Profile: Examples

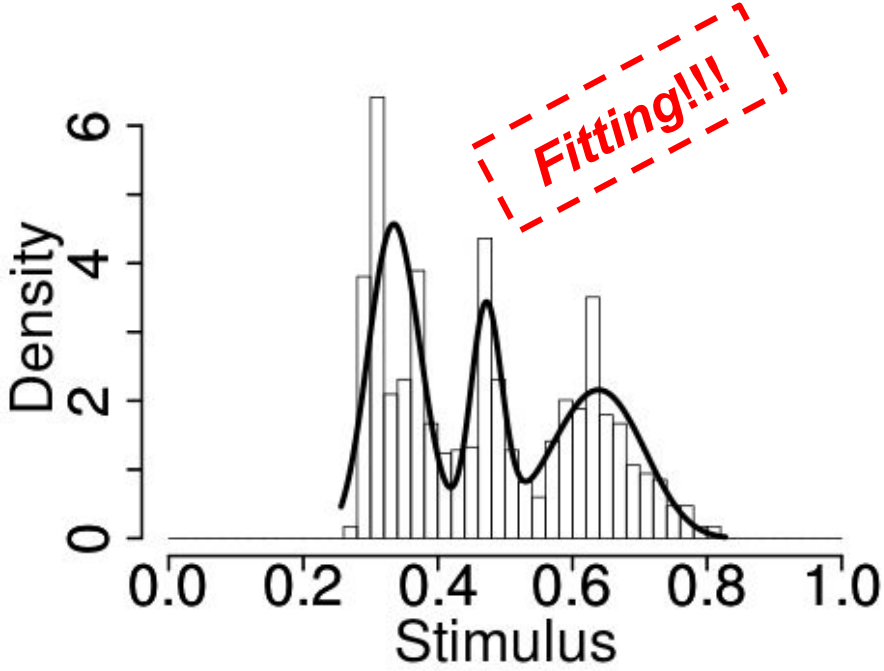



*Stimulus = **average** of all five metrics*

# Real Curiosity Profile: Examples




*Stimulus = **average** of all five metrics*

# Modeling User Curiosity

| Description | Users (Percent.) |
|---|---|
| 1 Beta | 24,578 (57.84%) |
| 2 Beta mixture | 15,498 (36.46%) |
| 3 Beta mixture | 1,684 ( 3.96%) |
| No valid users | 751 ( 1.76%) |
| **Total of users** | 42,511 (100.00%) |

# Modeling User Curiosity

| Description | Users (Percent.) |
|---|---|
| 1 Beta | 24,578 (57.84%) |
| 2 Beta mixture | 15,498 (36.46%) |
| 3 Beta mixture | 1,684 ( 3.96%) |
| No valid users | 751 ( 1.76%) |
| **Total of users** | 42,511 (100.00%) |

**58%** users is well modeled by **single** beta distribution
(*as assumed by prior work*)

# Modeling User Curiosity

| Description | Users (Percent.) |
|---|---|
| 1 Beta | 24,578 (57.84%) |
| 2 Beta mixture | 15,498 (36.46%) |
| 3 Beta mixture | 1,684 ( 3.96%) |
| No valid users | 751 ( 1.76%) |
| **Total of users** | 42,511 (100.00%) |

But a *large percentage* of users (**40%**) is *better modeled* by mixture of ***2 or 3 betas***.

- ***More betas*** *needed to capture* ***more complex behavior****: transitions across profiles occur more often (particularly niche radical).*

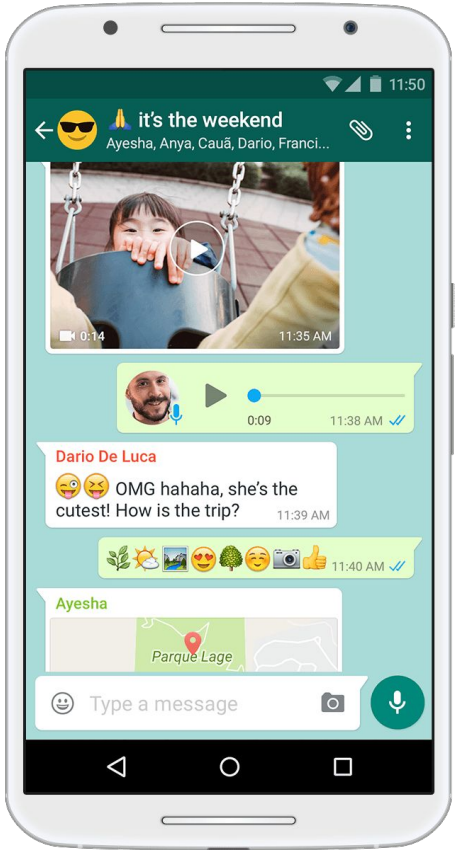
# Case Study 1: Summary of Results

1. Our findings are important as it **unveils** some of the *flaws of prior endeavors*.
2. Proposal of **new metrics** for all 4 collative variables:
   - *We inspect which collative variables **should be used** to model curiosity stimuli.*
3. *Modeling of dynamics of curiosity stimulus as function of metrics.*
4. *For **large fraction** of users we require more than one Wundt's curve to model curiosity:*
   - *use of **mixture of beta distributions (2 or 3 betas) for 40% of users**.*

# Case Study 2

## Metrics of **Social Curiosity**

# Contextualization: WhatsApp




- We consider a **universe** consisting of *sequences of messages shared* by **users** in various **groups**.
- *A **message** is composed of pieces of content of different **media types** (text, URL, image, audio or video).*
- We are not aware of any **prior attempt** to investigate *how to model social aspects of curiosity.*
- **Our focus:** *model the **social stimulus** to the **curiosity** that drive users to **share content** in WhatsApp groups.*

Focus only on **RQ2**.

# Supplementary Research Questions

**RQ2.1**: How to *quantify social influence* as a stimulus 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*?

# Novel Metrics of User Curiosity

- We **propose metrics** that *capture different aspects of the stimuli* one which is ***driven*** by when choosing to **share a piece of content** in group.

- We make use of **measures** from ***information theory*** to *capture different collative variables* associated with **social curiosity stimulation**:
  - *Proposal of 4 individual metrics of social curiosity (direct and indirect).*
  - *And 1 group metric of social curiosity.*
  - *They are based on concept of **Mutual Information** (MI).*

[Cover and Thomas, 2006; MacKay, 2005]

# Direct Influence

- *Pointwise Mutual Information (PMI):*

$$PMI^{\leftarrow}_{t,g}(D{=}d, O{=}o) = \begin{cases} \log_2\left(\dfrac{P^{\leftarrow}_{t,g}(D{=}d \,|\, O{=}o)}{P^{\leftarrow}_{t,g}(D{=}d)}\right), & \text{if } P^{\leftarrow}_{t,g}(D{=}d) > 0. \\ 0, & \text{otherwise.} \end{cases}$$

We derive from PMI: *average* and *maximum* ***direct*** *influence.*

# Indirect Influence

- ***Mutual Information*** *of the* ***destinations*** *conditioned on a* ***particular origin****:*

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

We derive from MI: *average and maximum* ***indirect*** *influence.*

# Metric of Social Curiosity for Group

- We note that **Mutual Information** can also be used to quantify *how social influence drives curiosity of a group*.
- It may *offer an aggregate view* of social **curiosity stimulation** in the **ecosystem** of a particular group.
- *Group Mutual Information*:

$$H(D) - H(D|O)$$

*entropy of all **destinations** regardless of social influence*

*entropy of destinations **conditioned** on origins*

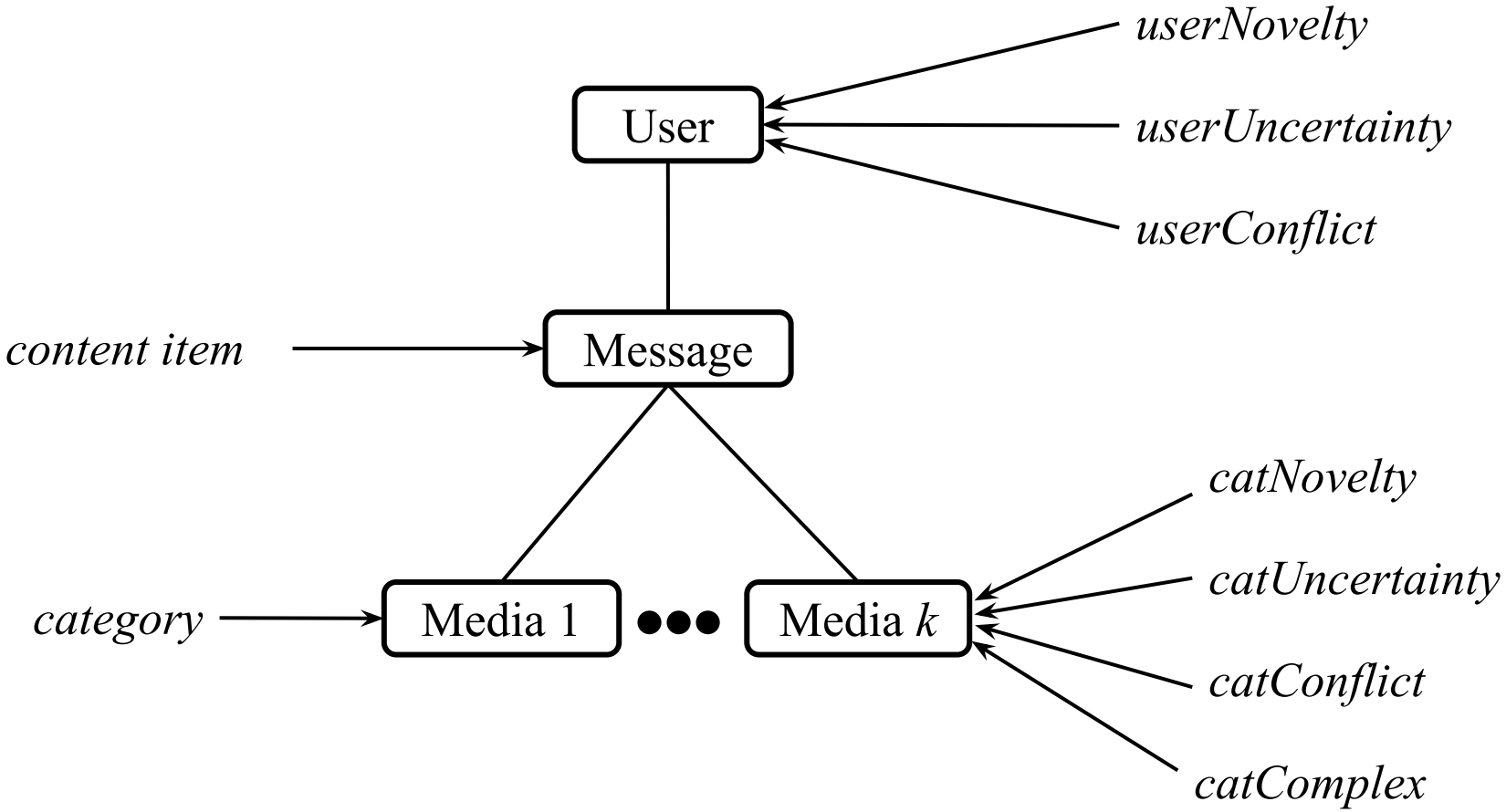
# Metric of Social Curiosity for Group

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

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

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

# Traditional Collative Variables



# Traditional Collative Variables

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

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

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

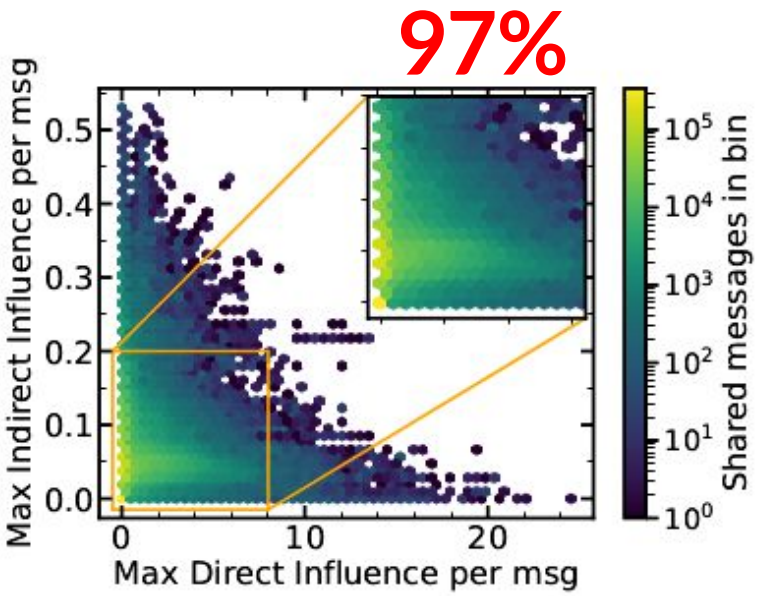
$$catUncertainty(t{=}t_{i|d,g}, g){=}-\sum_{c\in\mathcal{C}^{\rightarrow}_{t,g}} P^{\rightarrow}_{t,g}(C{=}c)\log_2\left(P^{\rightarrow}_{t,g}(C{=}c)\right)$$

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

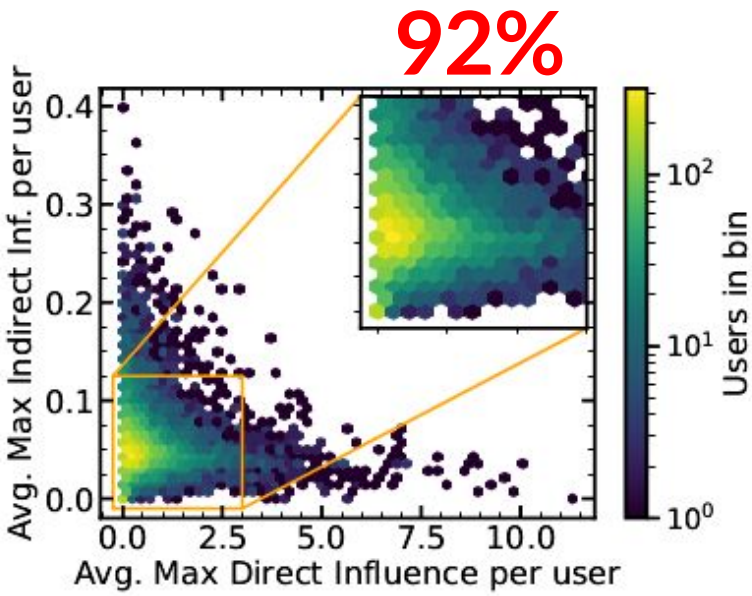
$$catConflict(t{=}t_{i|d,g}, g){=}-\log_2\left(\frac{1}{|\mathcal{C}^{\rightarrow}_{t,g}|}\sum_{c\in\mathcal{C}^{\rightarrow}_{t,g}} P^{\rightarrow}_{t,g}(C{=}c)\right)$$

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

# Social Curiosity: Diversity and Dynamics




(a) Stimuli associated with messages.




(b) Average stimuli for users.

The **social curiosity** is stimulated quite differently for different users and even for the same user.

# Social Curiosity at the Message Level

- ***Social stimulation profiles:***

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

# Social Curiosity at the Message Level

- ***Social 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 |

# Social Curiosity at the Message Level

- ***Social 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 |

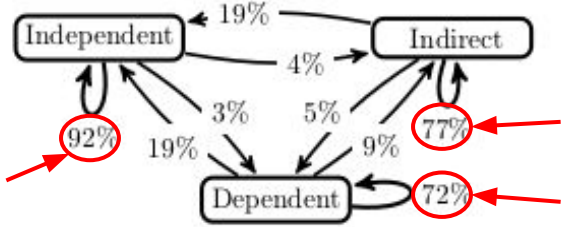
# Social Curiosity at the Message Level

- ***Social stimulation profiles:***

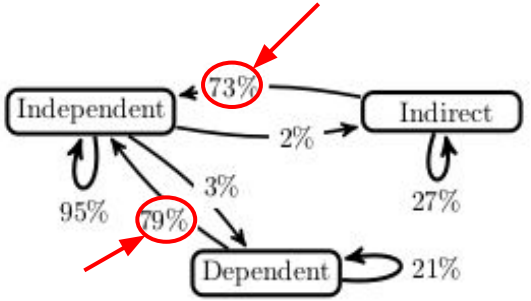
| Cluster | $\%msgs$ | Max. Dir. Influence (average ±95% C.I.) | Max. Ind. Influence (average ±95% C.I.) |
|---|---|---|---|
| Independent | 72.6% | 0.26 ± 0.000750 | 0.04 ± 0.000044 |
| Indirect | 14.4% | 0.43 ± 0.002320 | 0.15 ± 0.000210 |
| Dependent | 13.0% | 3.52 ± 0.006623 | 0.05 ± 0.000099 |

# Social Curiosity at the User Level

30%

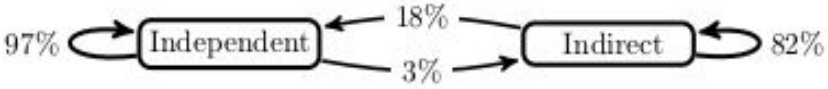



(a) $U_0$: 2,476 elements (383 msgs/user).




11%

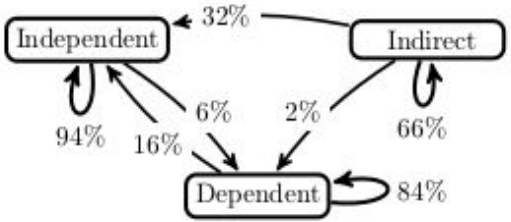
(b) $U_1$: 889 elements (90 msgs/user).



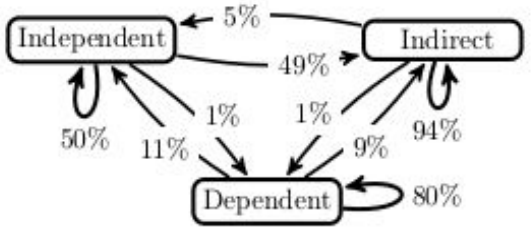

31%

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

26%




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




3%

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

# Social Curiosity at Group Level

- We focus on two aspects:
  - The ***role of the group*** on the curiosity stimulation of its members.
  - The ***overall social curiosity*** of each group.

# Social Curiosity at Group Level

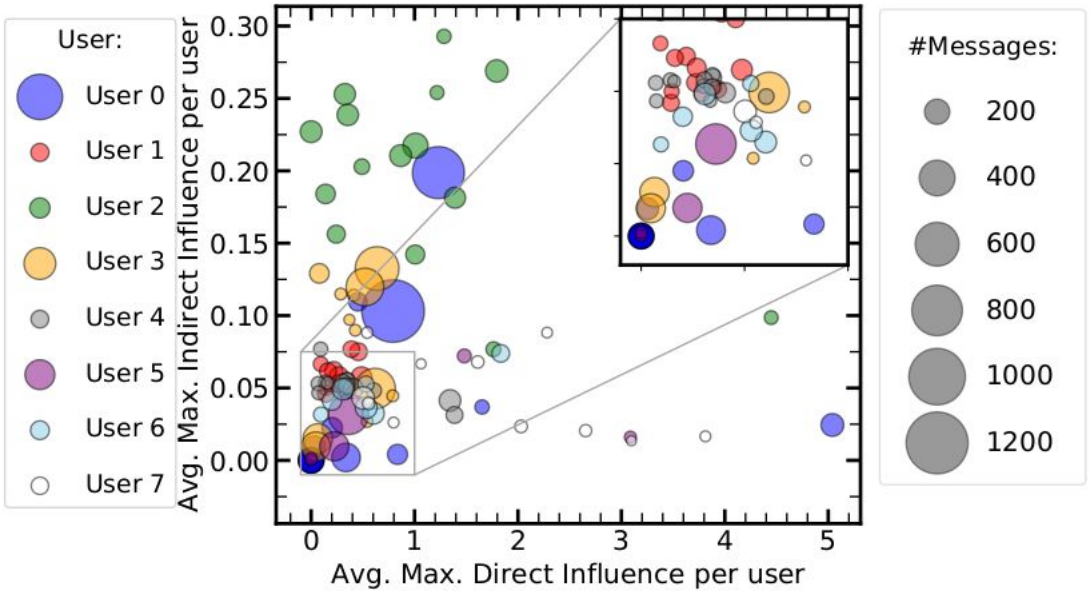
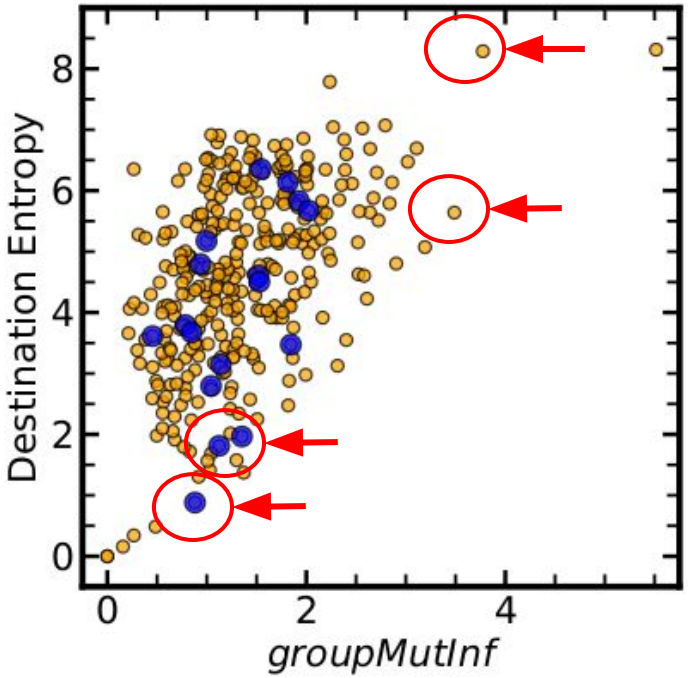


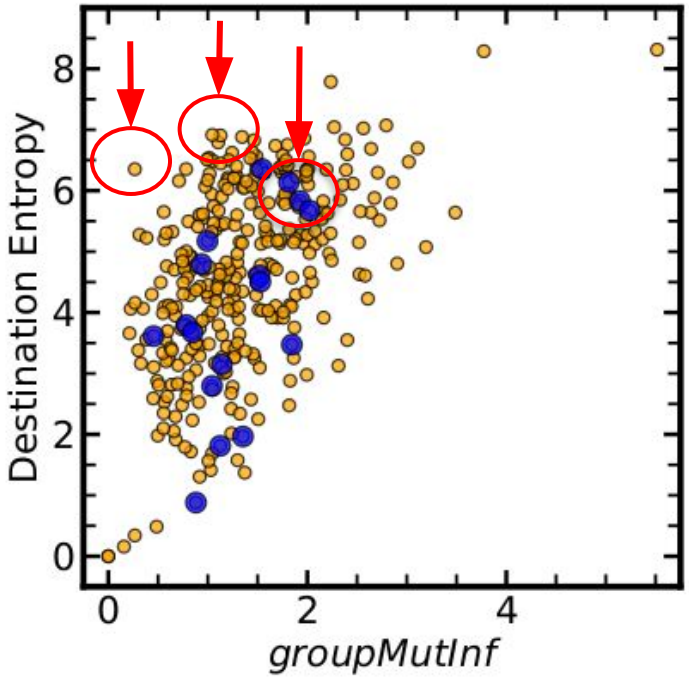

Figure 5.9: Diversity of social stimulation of selected users in different groups: the same user in different groups is represented by the same color.

# Social Curiosity at Group Level



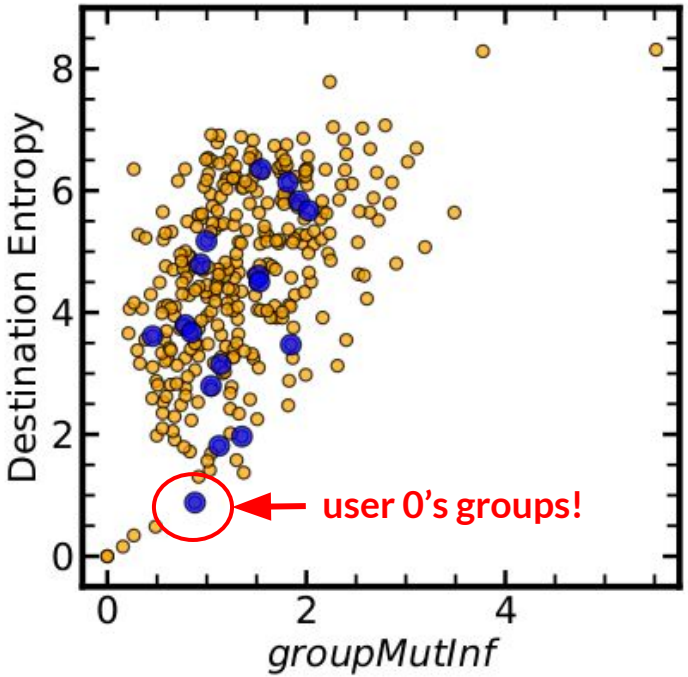

*We recall that* ***groupMutInf*** is *reduction in uncertainty* associated with **destinations** *due to the knowledge* of social influence from **origin** users.

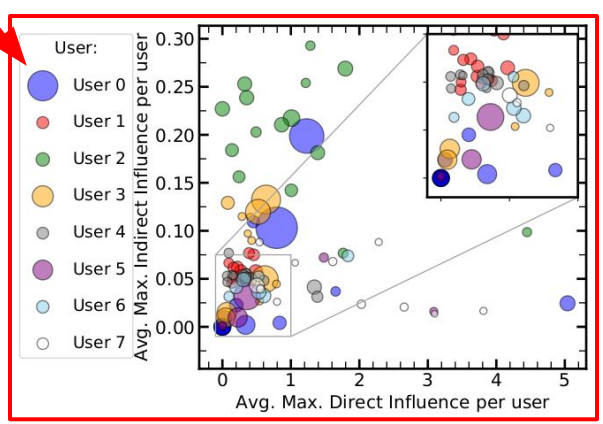
# Social Curiosity at Group Level




*We recall that* ***groupMutInf*** is *reduction in uncertainty* associated with **destinations** *due to the knowledge* of social influence from **origin** users.

# Social Curiosity at Group Level




*We recall that* ***groupMutInf*** is *reduction in uncertainty* associated with **destinations** *due to the knowledge* of social influence from **origin** users.

# Case Study 2: Summary of Results

*We proposed **4 new metrics** to capture social influence in publicly political-oriented **WhatsApp** groups.*

- *they are founded on* [Berlyne, 1960] ⟶ *challenges!!!*
- *relevance for individual users and groups!*

*The **proposed metrics** captures aspects of curiosity stimulation **not covered** by traditional collative variables.*

*We performed a **broad characterization** of social curiosity at 3 levels of aggregation (messages, users and groups).*

- ***our findings are novel** with respect to prior efforts!*
- *our metrics can be **applied** to **other environments.***

# Roadmap

Contextualization and Motivation

Goals, Hypothesis and Research Questions

Related Work

Problem Statement: Elements and Assumptions

Case Studies and Contributions

**Conclusions and Next Steps**

# Conclusions

- Computational models of curiosity have ***neglected*** the use of ***multiple collative variables***.
- Although modern psychological studies argue for the existence of ***social curiosity***, that aspect of curiosity stimulation ***has not been considered***.
- ***Curiosity models*** of Wundt's curve ***neglect*** the ***heterogeneity and dynamics*** of online user's behavior stimulation.
- Curiosity models used in ***recommendation system*** still ***need to be improved*** to provide *better personalized* and *diverse* services.

# Conclusions

**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?

# Conclusions on **RQ1**

1. We ***proposed metrics*** to capture ***different aspects*** related to stimulating user curiosity.

2. At the **access level**, we uncovered four ***different profiles*** of curiosity stimulation: *underground*, *radical niche*, *eclectic and explorer*.

3. At the **user level**, we uncovered ***dynamic patterns*** that show how stimulation process of user's curiosity changes across different profiles.

4. *Most curiosity stimulation profiles prefer **new songs** from **different artists** with **different music genres** related to those they have recently listened to.*

# Conclusions on **RQ2**

1. The stimulation of social curiosity ***varies*** not only between individual *messages*, but also between *different users* and *different groups*.
2. At the **message level**, we uncovered three profiles and in almost 30% of the cases, social influence is indeed the reason for user to share content.
3. At the **user level**, we have shown that the evolution of user curiosity is very heterogeneous over time with five different user profiles.
4. We also demonstrated that the stimulation of curiosity can **change significantly** ***depending on the group*** to which the user belongs.
5. At the **group level**, the role of social curiosity is **more evident** when group members are more ***engaged*** in ongoing discussions (sharing content).

# Conclusions on **RQ3**

1. The conventional ***single*** Wundt's curve ***does not*** *adequately model the process of curiosity elicitation for a large fraction of users* (**40%**).
2. For such users, **we propose** a ***combination*** of 2 or 3 Beta distributions.
3. These observations can be applied to ***other domains*** besides music.
   - *e.g., communication platforms and content production and sharing services.*

# Next Steps

1. We focus on ***applying*** our *model of human curiosity* in recommendation systems to answer **RQ4**.
2. We intend to begin developing a hybrid ***Recommendation System (RS) framework*** that includes:
   - *General Appraisal Process.*
   - *Machine learning-based recommendation model.*
3. To **evaluate the performance** of our proposed recommendation framework **against** the main ***state–of-the-art baselines*** [***Zhao and Lee***, 2016; ***Shrestha et al.***, 2020; ***Xu et al.***, 2021].
   - *LFM-1B Dataset* [***Schedl***, 2016] *and LastFM Social Connections* [***Duricic et al.***, 2021].

# Planned Schedule

| 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 | | | | | | | | | ✓ |

# 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), Elsevier, 2022.

# Thanks!

e-mail: amagnosousa@dcc.ufmg.br

# Backup

# Traditional Collative Variables

## Online Music Consumption in **LastFM**

# Proposed Metrics

Metrics grounded in Surprisal:

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

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

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

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

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

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

Metrics grounded in Entropia:

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

| Song | Artist | Timestamp |
|---|---|---|
| **9**. Have a Drink on me | AC/DC | 12:00 |
| **8**. Live and Let Die | Guns N'Roses | 11:30 |
| **7**. Patience | Guns N'Roses | 11:20 |
| **6**. Live and Let Die | Guns N'Roses | 11:05 |
| **5**. Don't cry (original) | Guns N'Roses | 10:55 |
| **4**. November Rain | Guns N'Roses | 10:45 |
| **3**. Black in Black | AC/DC | 10:30 |
| **2**. Welcome to The Jungle | Guns N'Roses | 10:25 |
| **1**. T.N.T. | AC/DC | 10:15 |

(a) Songs listened by User 1 in online music consumption.

- (User 1, **Song 6**, 11:05am)
- (User 1, **Song 7**, 11:20am)
- (User 1, **Song 8**, 11:30am)
- (User 1, **Song 9**, 12:00pm)

window: [11:00am; 12:00pm]

- (User 1, **Song 1**, 10:15am)
- (User 1, **Song 2**, 10:25am)
- (User 1, **Song 3**, 10:30am)
- (User 1, **Song 4**, 10:45am)
- (User 1, **Song 5**, 10:55am)

window: [ 9:55am; 10:55am]

(b) Windows of interaction for songs 9 and 5 listened by User 1 ($\delta_T = 1$ hour).

| Song | Artist | Timestamp |
|---|---|---|
| **9**. Have a Drink on me | AC/DC | 12:00 |
| **8**. Live and Let Die | Guns N'Roses | 11:30 |
| **7**. Patience | Guns N'Roses | 11:20 |
| **6**. Live and Let Die | Guns N'Roses | 11:05 |
| **5**. Don't cry (original) | Guns N'Roses | 10:55 |
| **4**. November Rain | Guns N'Roses | 10:45 |
| **3**. Black in Black | AC/DC | 10:30 |
| **2**. Welcome to The Jungle | Guns N'Roses | 10:25 |
| **1**. T.N.T. | AC/DC | 10:15 |


(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).

| | Song | Artist | Timestamp |
|---|---|---|---|
| ▶ | 9. Have a Drink on me | AC/DC | 12:00 |
| ▶ | 8. Live and Let Die | Guns N'Roses | 11:30 |
| ▶ | 7. Patience | Guns N'Roses | 11:20 |
| ▶ | 6. Live and Let Die | Guns N'Roses | 11:05 |
| ▶ | **5.** Don't cry (original) | Guns N'Roses | 10:55 |
| ▶ | 4. November Rain | Guns N'Roses | 10:45 |
| ▶ | 3. Black in Black | AC/DC | 10:30 |
| ▶ | 2. Welcome to The Jungle | Guns N'Roses | 10:25 |
| ▶ | 1. T.N.T. | AC/DC | 10:15 |

(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).

# Social Curiosity Metrics

## Communication Platform: **WhatsApp**

# Direct Influence

- *Pointwise Mutual Information (PMI):*

$$PMI^{\leftarrow}_{t,g}(D{=}d, O{=}o) = \begin{cases} \log_2\left(\dfrac{P^{\leftarrow}_{t,g}(D{=}d \,|\, O{=}o)}{P^{\leftarrow}_{t,g}(D{=}d)}\right), & \text{if } P^{\leftarrow}_{t,g}(D{=}d) > 0. \\ 0, & \text{otherwise.} \end{cases}$$

We derive from PMI: *average* and *maximum* ***direct*** *influence.*

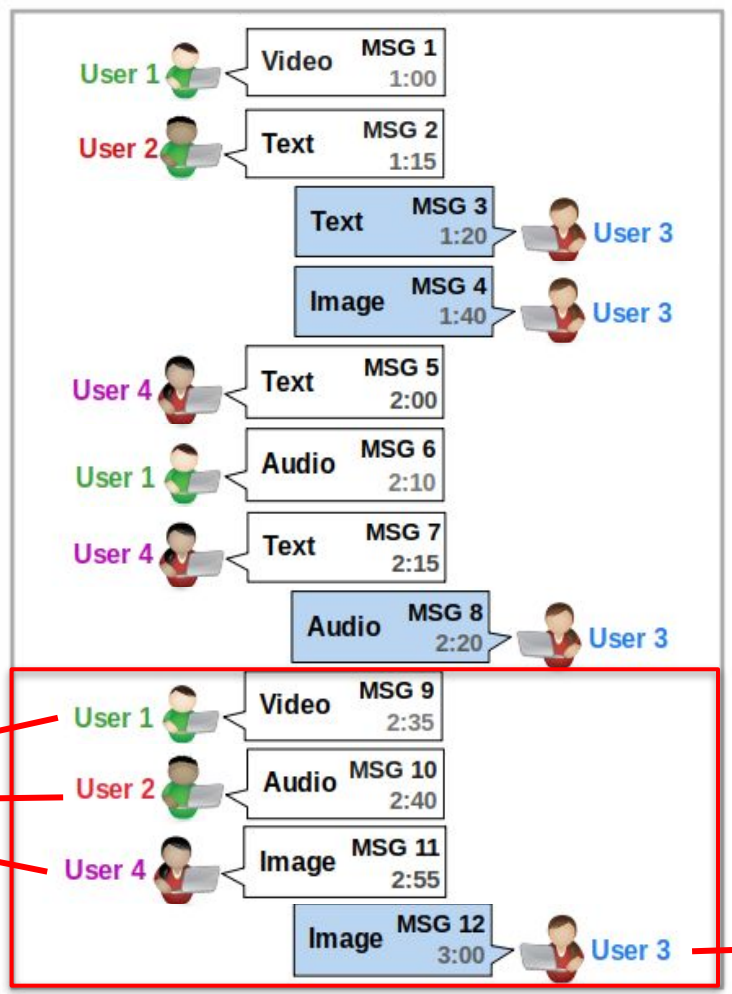
# Direct Influence

- *Influência Social Direta Máxima e Média:*

$$socInf^{\leftarrow}_{t,g}(D=d,O=o)=\max\left(PMI^{\leftarrow}_{t,g}(D=d,O=o),0\right)$$

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

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




*historical patterns*

*possible origins of direct influence*

*destination of influence*

*current window of interaction!*

# Indirect Influence

- ***Mutual Information*** *of the* ***destinations*** *conditioned on a* ***particular origin****:*

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

We derive from MI: *average and maximum* ***indirect*** *influence.*

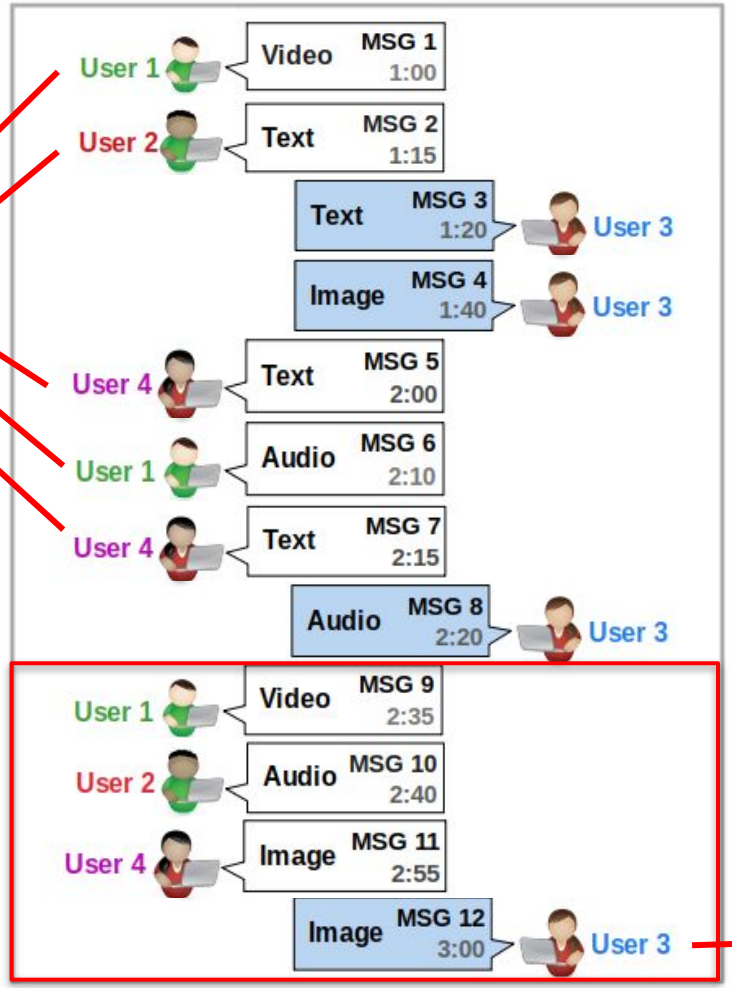
# Indirect Influence

- *Influência Social Indireta Máxima e Média:*

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

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

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




*possible origins of indirect influence*

*historical patterns*

*destination of influence*

*current window of interaction!*

# Metric of Social Curiosity for Group

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

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

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

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

# Metric of Social Curiosity for Group

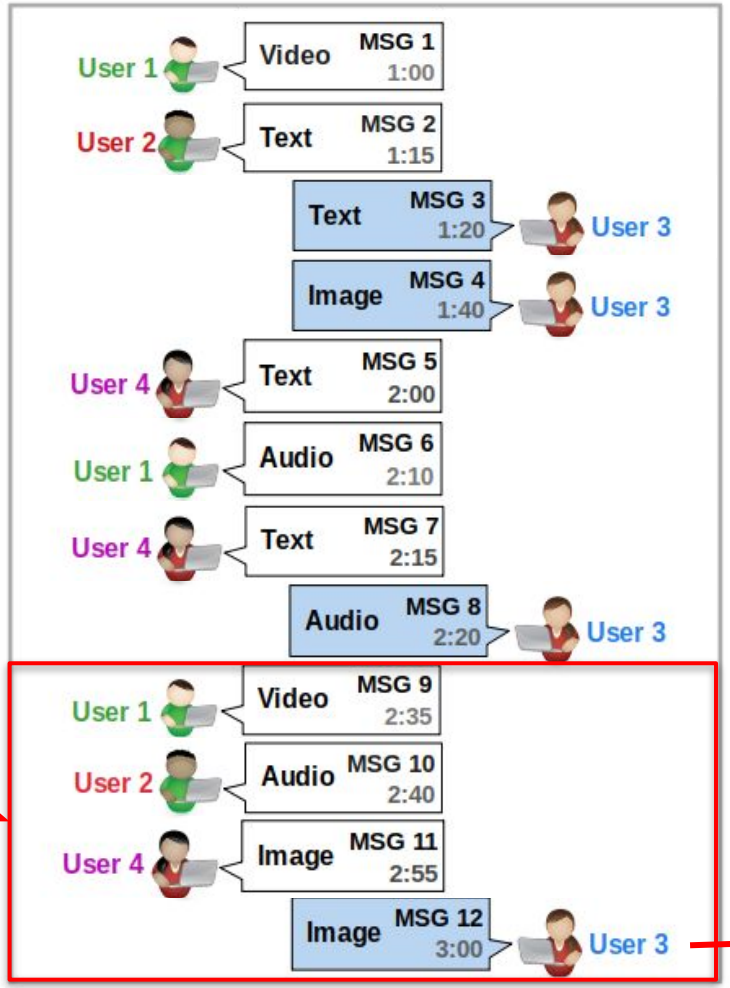
- We note that **Mutual Information** can also be used to quantify *how social influence drives curiosity of a group*.
- It may *offer an aggregate view* of social **curiosity stimulation** in the **ecosystem** of a particular group.
- *Group Mutual Information*:

$$H(D) - H(D|O)$$

*entropy of all destinations D regardless of social influence*

*entropy of destinations D conditioned on origins O*

*traditional collative variables are computed only here*




*destination of influence*

*current window of interaction!*

# Traditional Collative Variables

## Metrics grounded in Surprisal:

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

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

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

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

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

## Metrics grounded in Entropia:

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

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

# Next Steps

Preliminary Results

# Setup os Experiments

**Note:**
- Total of 10 runs
- Each run with 7 tests
- Total of 70 experiments

1st day randomly selected from dataset

1st week | 2nd week

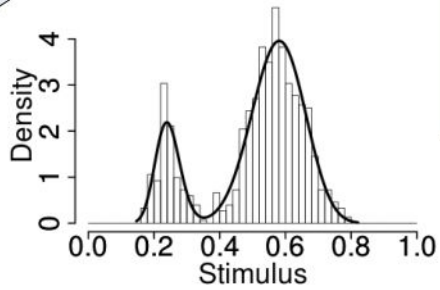
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

Training dataset | Test dataset

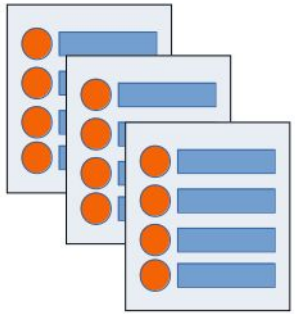
Computing stimulus

Collative Variables:
Novelty
Complexity
Uncertainty

Estimation of Curiosity Curve




Recommendation

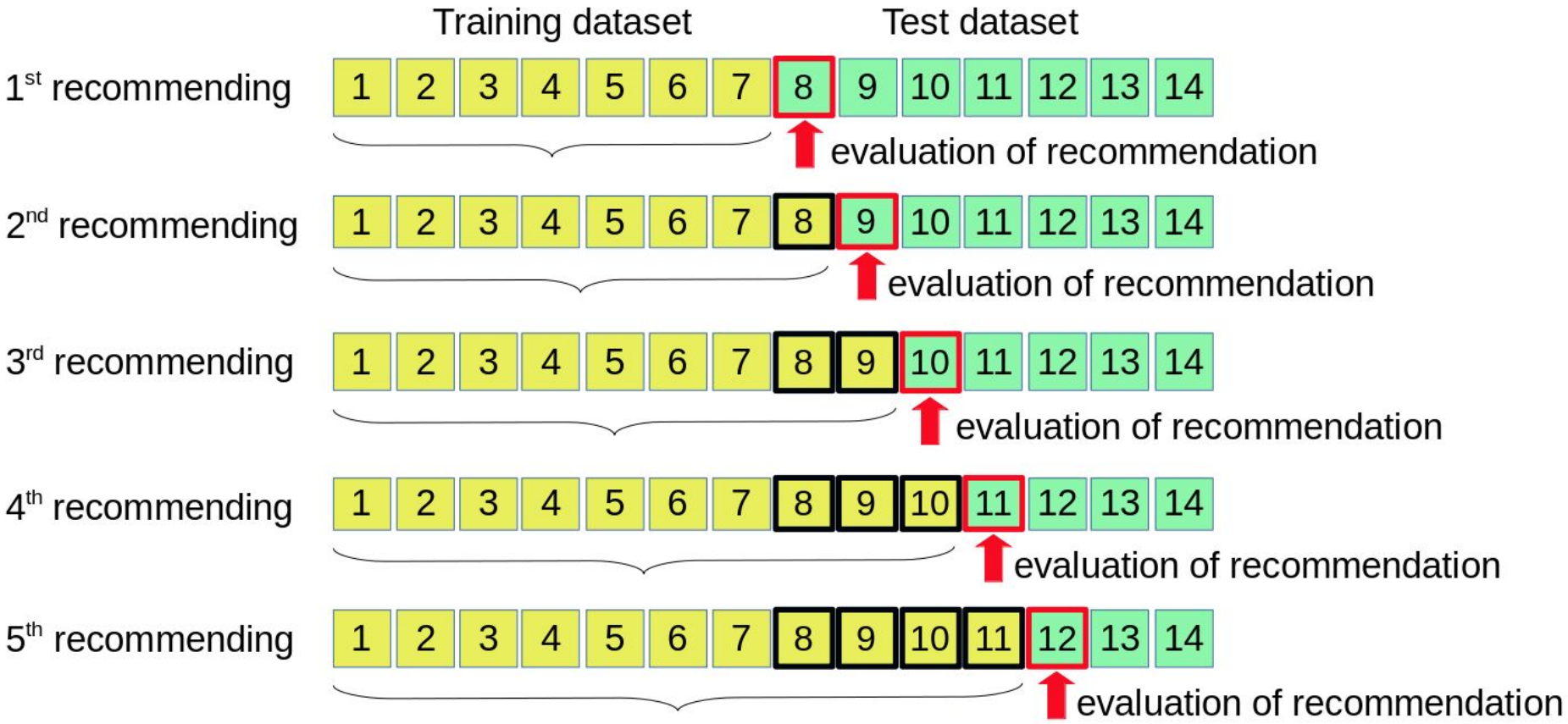


Evaluation of recommendation

Each user has at least:
- 100 events into train
- 50 events for each day of Test

**Note**:
- ≈1,450 users per run

# Setup os Experiments



# Setup of Experiments

State-of-art **Baselines**:

- *Curiosity based Recommendation System* (CBRS) [**Zhao and Lee**, 2016]
- *Curiosity-drive Recommendation Framework* (CdRF) [**Xu et al**, 2021]

Our proposal for ***Wundt's Curve Model*** estimation (probability distribution):

- Unique **Beta** distribution and Kernel Density Estimation (**KDE**)

Algorithms: ***Collaborative Filtering***, *Matrix Factorization and Popularity*

**Top K** Recommendation: K = {5, 20, 50}

***Incorporation of Curiosity***: 10%, 50% and 90%

# Setup of Experiments

Incorporation of Curiosity:

$$FC_i = (1-\theta)\, Rel_i + \theta CS_i$$

- $Rel_i$: *relevancy* of item i (song)
- $CS_i$: *curiosity score* of item *i* (song)
- $\theta$: incorporation *proportion* of curiosity
- $FC_i$: *final curiosity score*

# Setup of Experiments

Legend:

- **CBRS_Z**: CBRS of Zhao and Lee (2016) **original** implementation
- **SECM**: CdRF of Xu et al. (2021)
- **CF**: Collaborative Filtering without curiosity incorporation
- **New_UB**: ***our proposal*** with unique **Beta** distribution for curiosity curve
- **New_UK**: ***our proposal*** with unique **KDE** for curiosity curve

# Evaluation Metrics

***Precision***: measure *accuracy* of recommendations

***Intra List Dissimilarity (ILS)***: evaluate the impact of personalization in terms of *diversity*

***Aggregation Diversity (AD)***: measure the distinct number of recommended items as a measure of aggregate *diversity*.

***Curiosity Fitness***: evaluate the *curiosity fitting* of recommended item with respect the curiosity curve profile.

# Evaluation Metrics

$$Precision = \frac{1}{|U|} \sum_{u \in U} \frac{|L_u \cap T_u|}{|L_u|}$$

$$AD = \left| \cup_{u \in U} L_u \right|$$

$$CF_u = \frac{1}{K} \sum_{i \in L_u} \frac{CS_i}{CS_{max}}$$

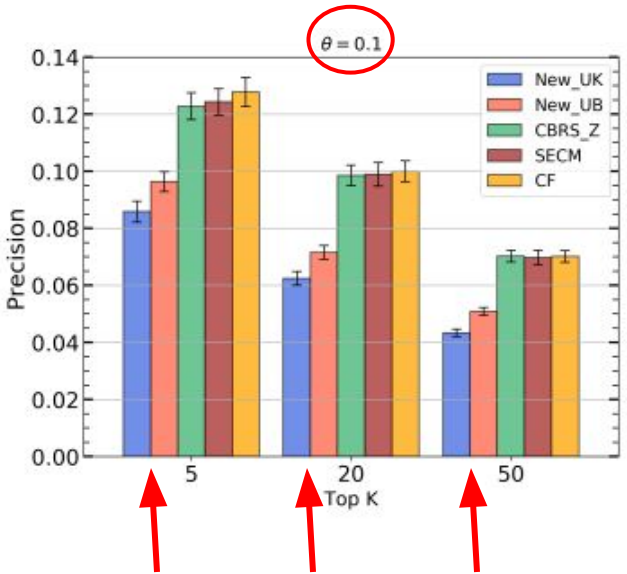
How much **LARGER**, better!

$$ILS_{uv} = \frac{L_u \cap L_v}{K}$$

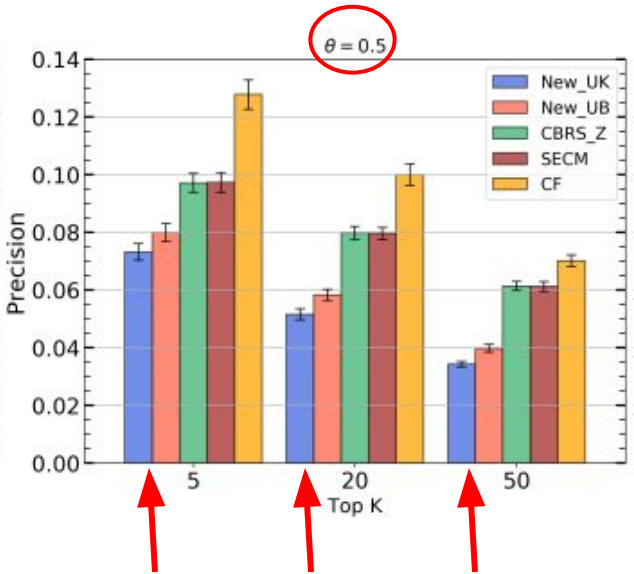
How much **SMALLER**, better!

# Results: Precision

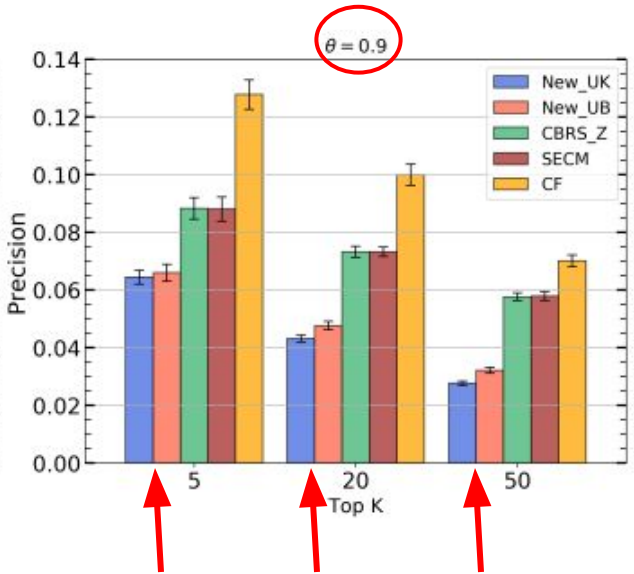
Curiosity 10%



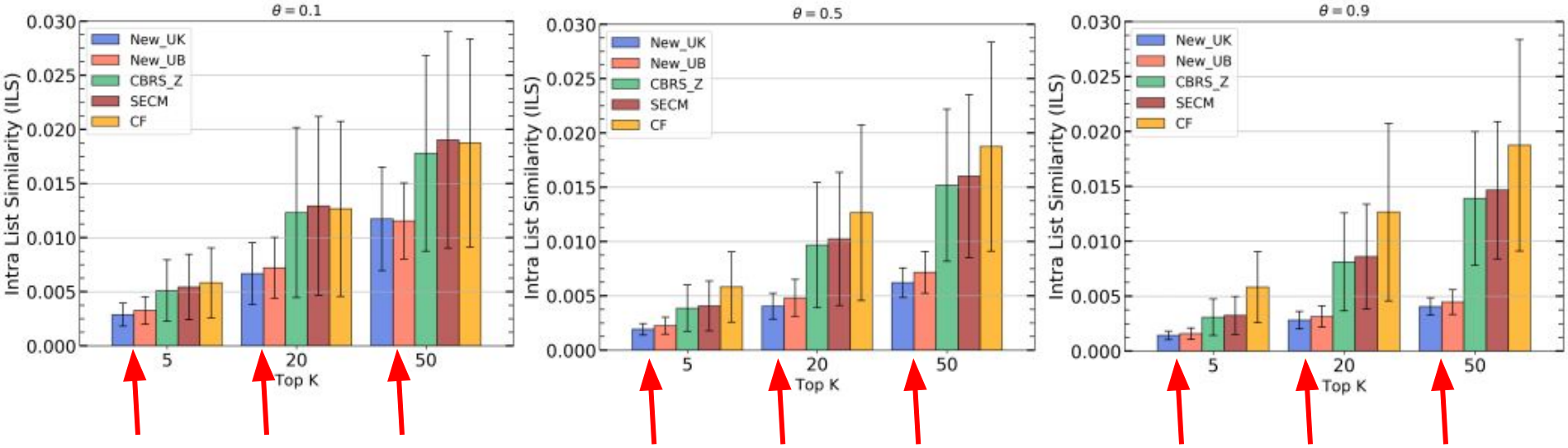

Curiosity 50%




Curiosity 90%

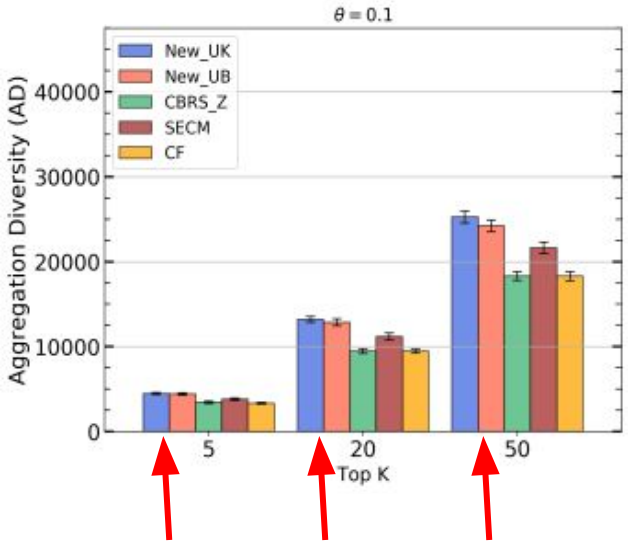


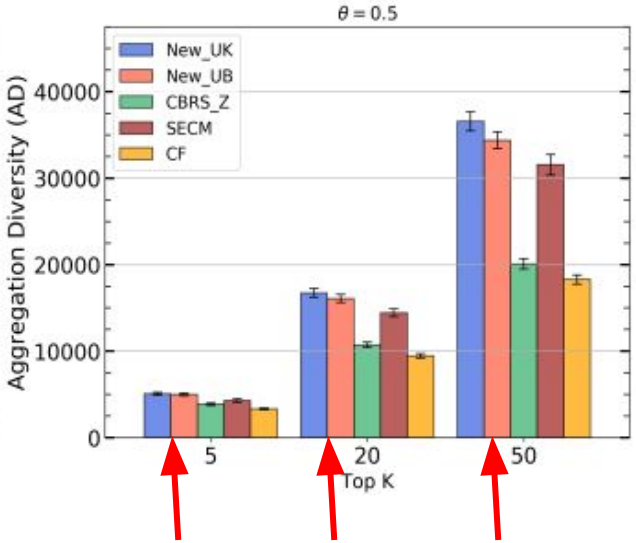
# Results: ILS (diversity)

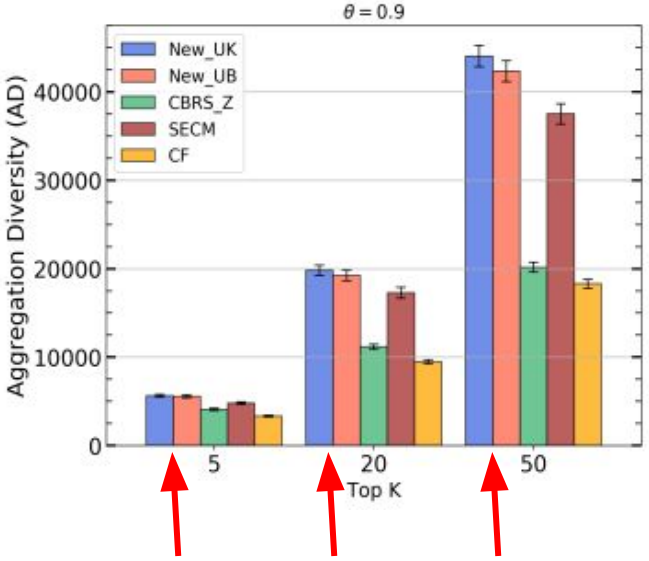



How much **SMALLER**, better!

# Results: AD (diversity)



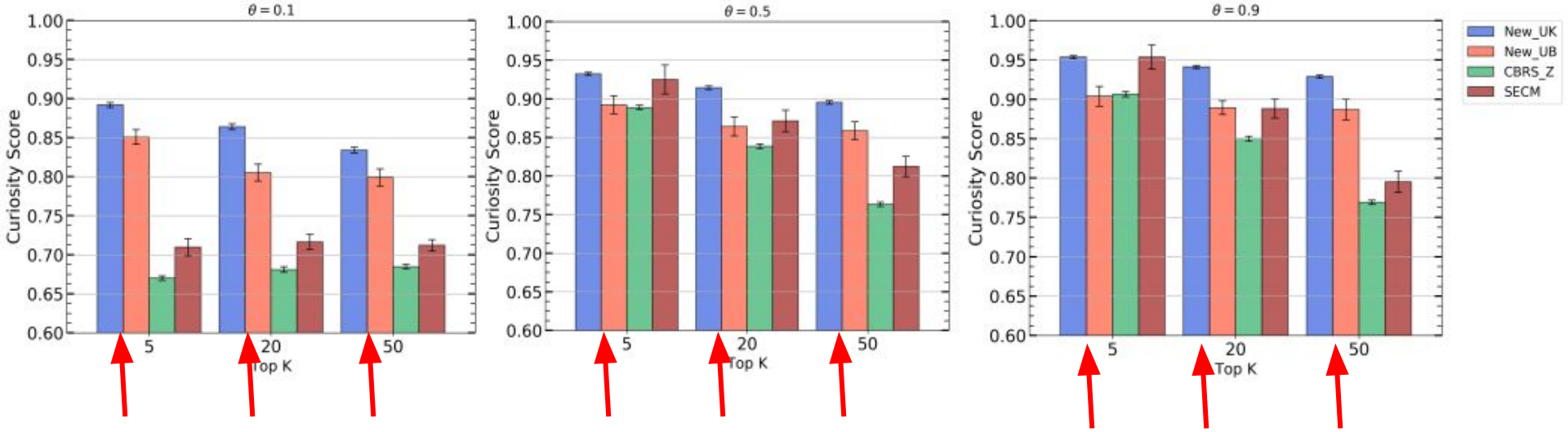






How much **LARGER**, better!

# Results: Curiosity Fitness




How much **LARGER**, better!