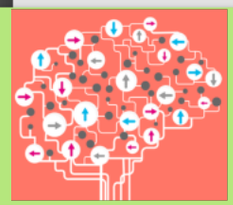


Behavioral
Data Science

UTS
UNIVERSITY OF TECHNOLOGY SYDNEY

Evolution of Privacy Loss in Wikipedia

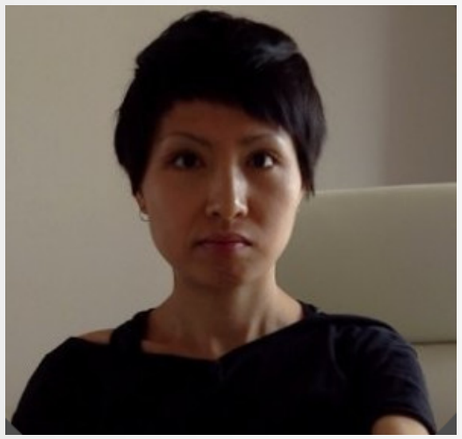
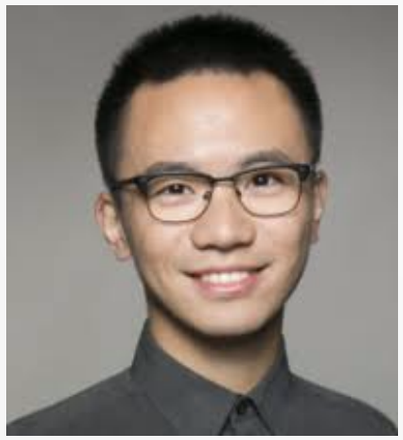
Marian-Andrei Rizoiu

WIKIMEDIA
FOUNDATION

The research group

Behavioral Data Science

1 research associate, 3 PhD students, 2 Honors students, 1 lecturer

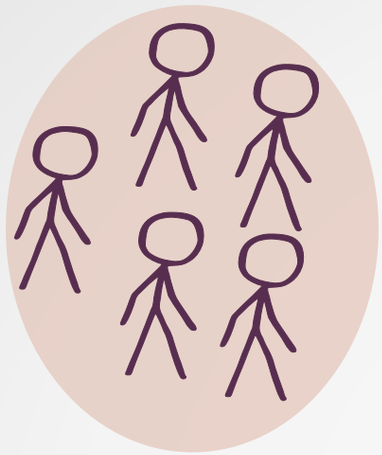





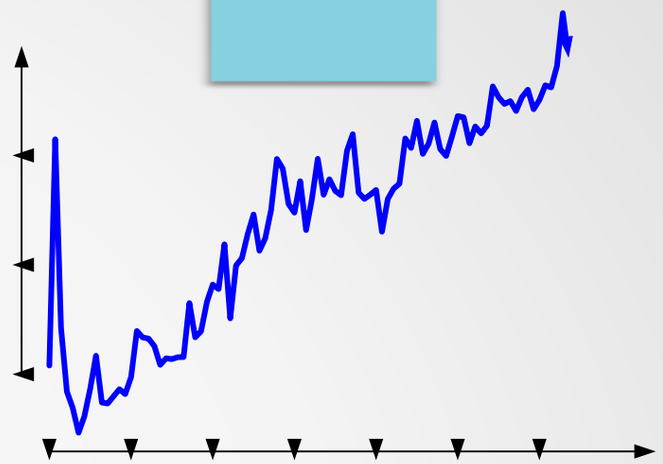

Research objectives

Behavioral Data Science

1.



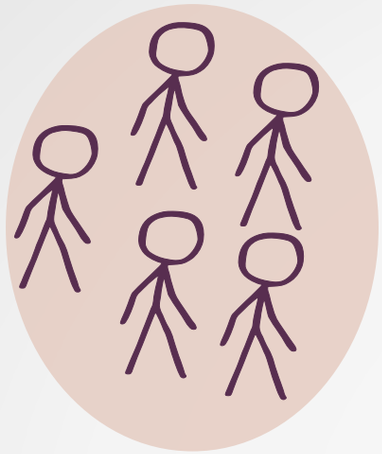
information diffusion
epidemics spreading
behavioral modeling

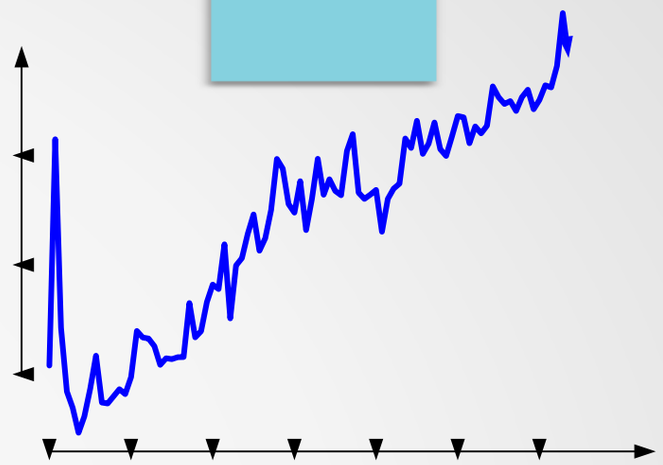


Research objectives

Behavioral Data Science

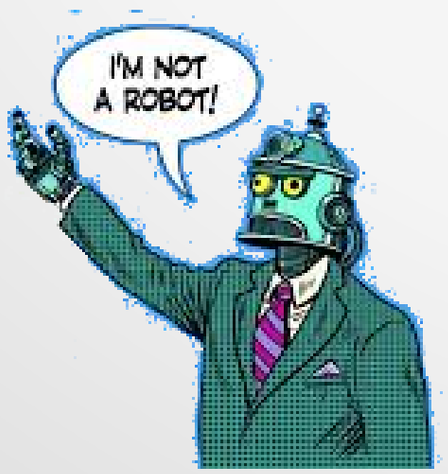
1.





information diffusion
epidemics spreading
behavioral modeling

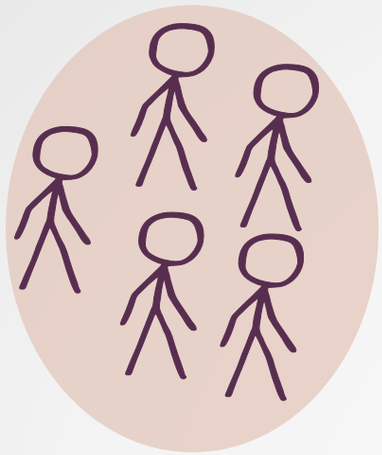
2.



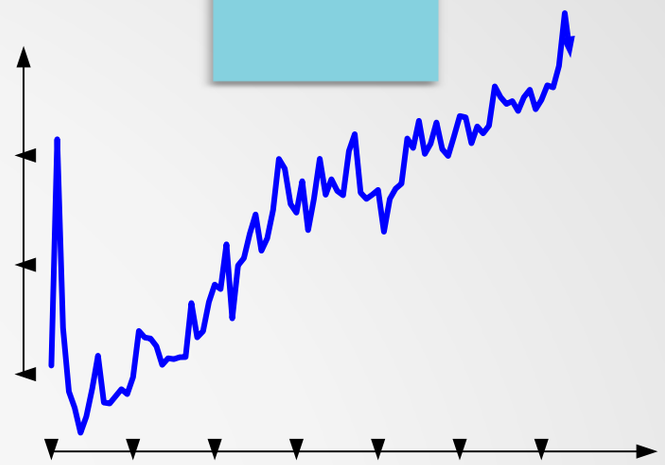
Research objectives

Behavioral Data Science

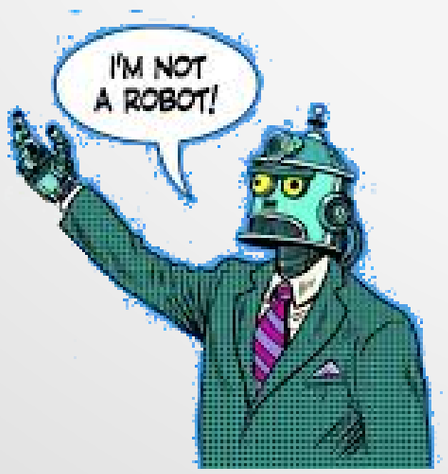
1.



information diffusion
epidemics spreading
behavioral modeling



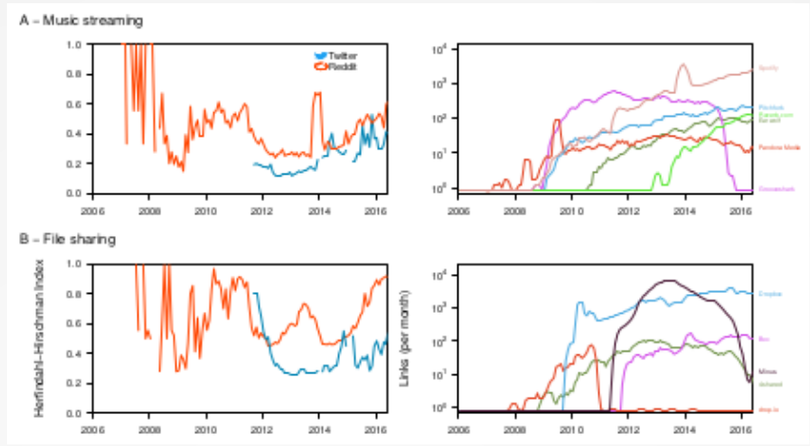
2.

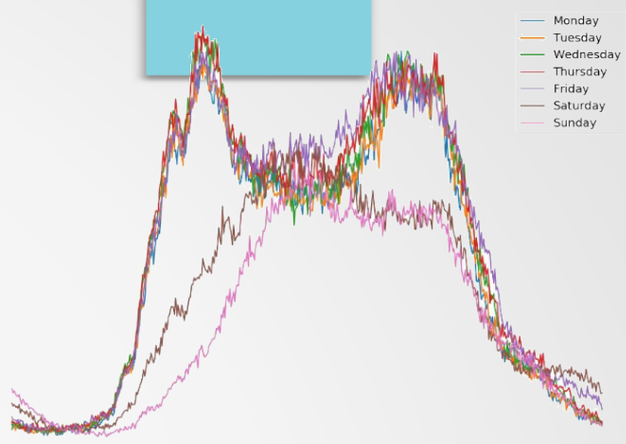


3.

Other projects

Behavioral Data Science

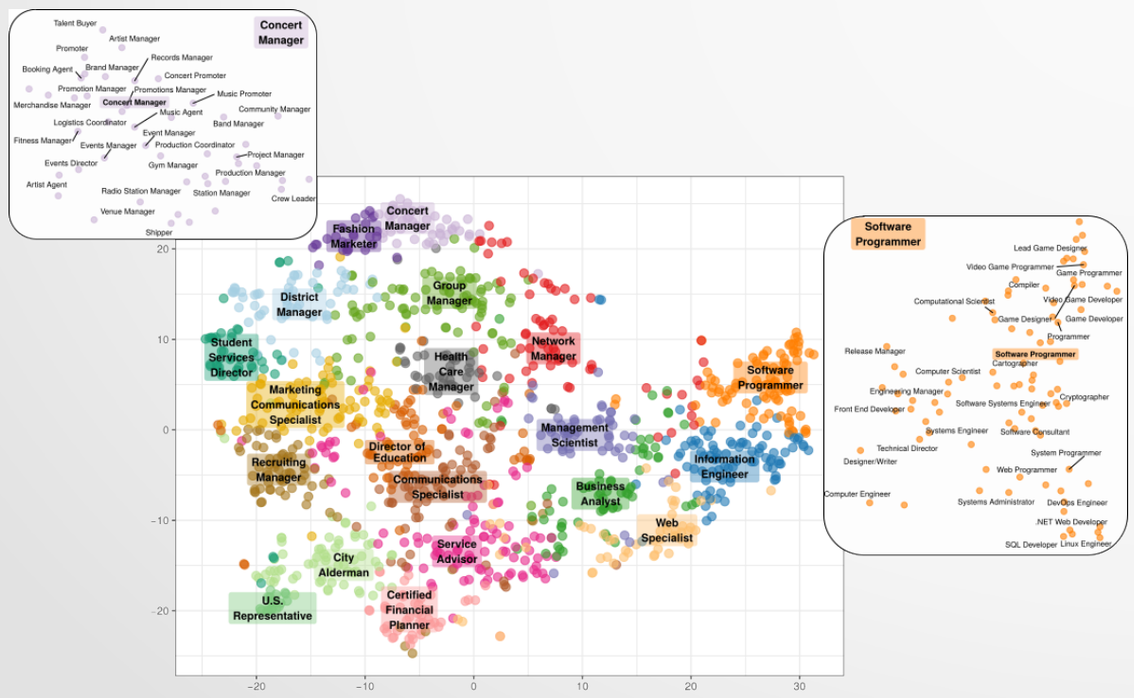




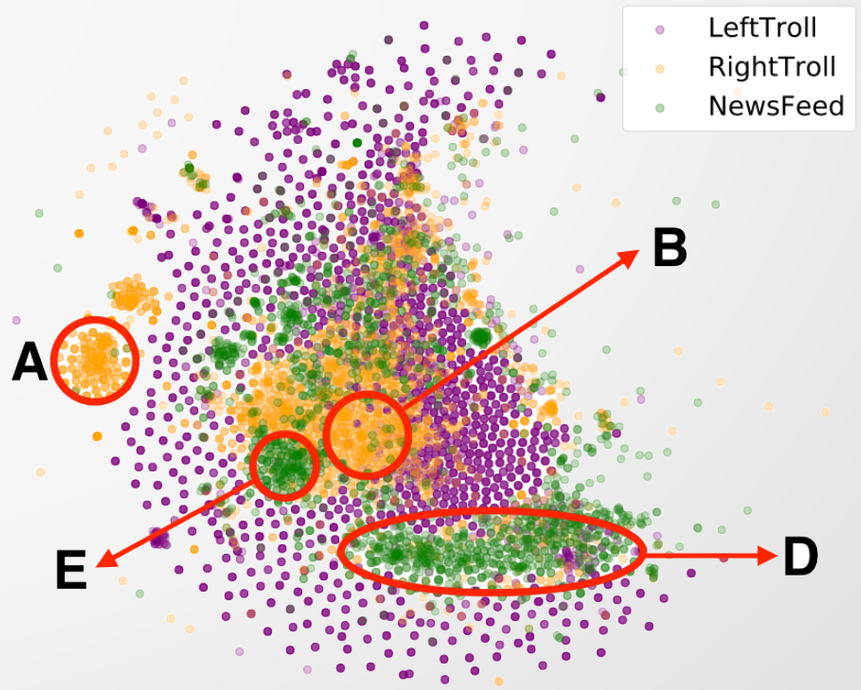
Wikipedia privacy

Online Diversity

Smart traffic

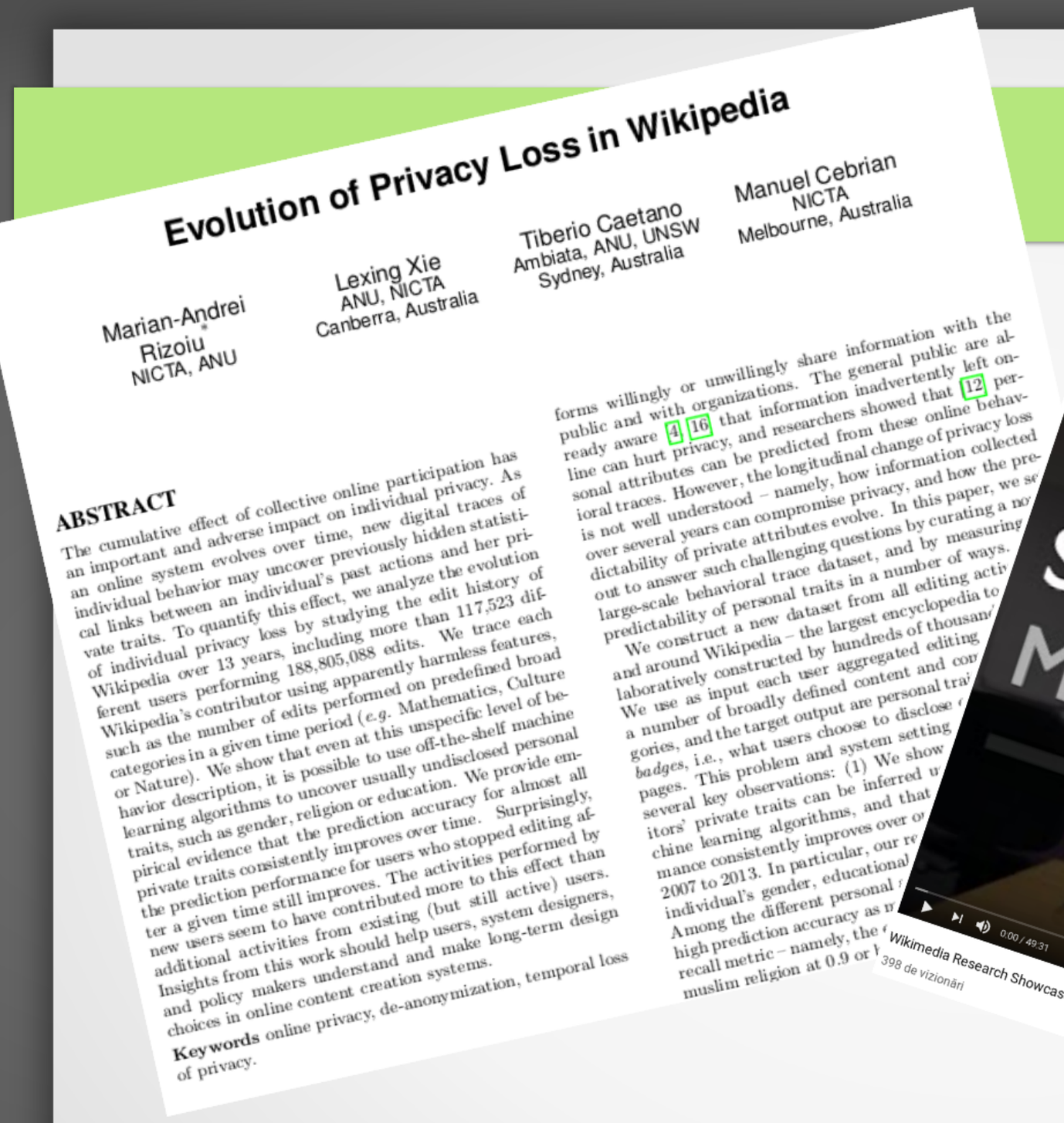


Vocation compass



Busting Russian Trolls

[Rizoiu, WSDM'16]



Evolution of Privacy Loss in Wikipedia

Marian-Andrei Rizoiu
NICTA, ANU

Lexing Xie
ANU, NICTA
Canberra, Australia

Tiberio Caetano
Ambiata, ANU, UNSW
Sydney, Australia

Manuel Cebrian
NICTA
Melbourne, Australia

ABSTRACT

The cumulative effect of collective online participation has an important and adverse impact on individual privacy. As an online system evolves over time, new digital traces of individual behavior may uncover previously hidden statistical links between an individual's past actions and her private traits. To quantify this effect, we analyze the evolution of individual privacy loss by studying the edit history of Wikipedia over 13 years, including more than 117,523 different users performing 188,805,088 edits. We trace each Wikipedia's contributor using apparently harmless features, such as the number of edits performed on predefined broad categories in a given time period (*e.g.* Mathematics, Culture or Nature). We show that even at this unspecific level of behavior description, it is possible to use off-the-shelf machine learning algorithms to uncover usually undisclosed personal traits, such as gender, religion or education. We provide empirical evidence that the prediction accuracy for almost all private traits consistently improves over time. Surprisingly, the prediction performance for users who stopped editing after a given time still improves. The activities performed by new users seem to have contributed more to this effect than additional activities from existing (but still active) users. Insights from this work should help users, system designers, and policy makers understand and make long-term design choices in online content creation systems.

Keywords online privacy, de-anonymization, temporal loss of privacy.

forms willingly or unwillingly share information with the public and with organizations. The general public are already aware [4] [16] that information inadvertently left online can hurt privacy, and researchers showed that [12] personal attributes can be predicted from these online behavioral traces. However, the longitudinal change of privacy loss is not well understood – namely, how information collected over several years can compromise privacy, and how the predictability of private attributes evolve. In this paper, we set out to answer such challenging questions by curating a no large-scale behavioral trace dataset, and by measuring predictability of personal traits in a number of ways.

We construct a new dataset from all editing activ laboratively constructed by hundreds of thousan and around Wikipedia – the largest encyclopedia to We use as input each user aggregated editing a number of broadly defined content and con gories, and the target output are personal trai *badges*, i.e., what users choose to disclose pages. This problem and system setting several key observations: (1) We show itors' private traits can be inferred u chine learning algorithms, and that mance consistently improves over o 2007 to 2013. In particular, our re individual's gender, educational Among the different personal high prediction accuracy as n recall metric – namely, the muslim religion at 0.9 or

Evolution of privacy loss in Wikipedia

Motivation

- Social media and online privacy are two of today's hot topics.
- Given that we know that digital traces reveal more than users might think **[Kosinski, PNAS 2013]**, we ask the next questions.

- Goals of this work:
  - Does online user's privacy degrade over time?
  - What factors contribute most to revealing private traits?
  - Can I stop leaking personal information if I stop posting online?

Content of this presentation

Presentation outline

- Case study: Wikipedia dataset
- Profiling of editing behavior
- Measuring predictability of personal traits
- Marginal utility of features over time
- Conclusion and the way ahead

Case study: Wikipedia

Why Wikipedia?

- 13 years long, public: ideal for longitudinal study;
- tens of thousands of editors, of different geographic locations, religious, educational and political backgrounds;
- *apparently harmless dataset*: a reservoir of knowledge, no focus on personal information.

Dataset dimensionality:

- 188,805,088 revisions
- 117,523 editors
- 8,679 editor badges
- 22,172,813 edited pages
- 430,410 page categories
- Time extent: January 2001 - July 2013.

Encoding editing behavior (1)

Editor activity profiles:

- *basic set*: #revisions over 6 predefined categories (Wiki namespaces);
- *extended set*: adding Wikipedia's 23 high level thematic categories (Math, Geography, History etc.)

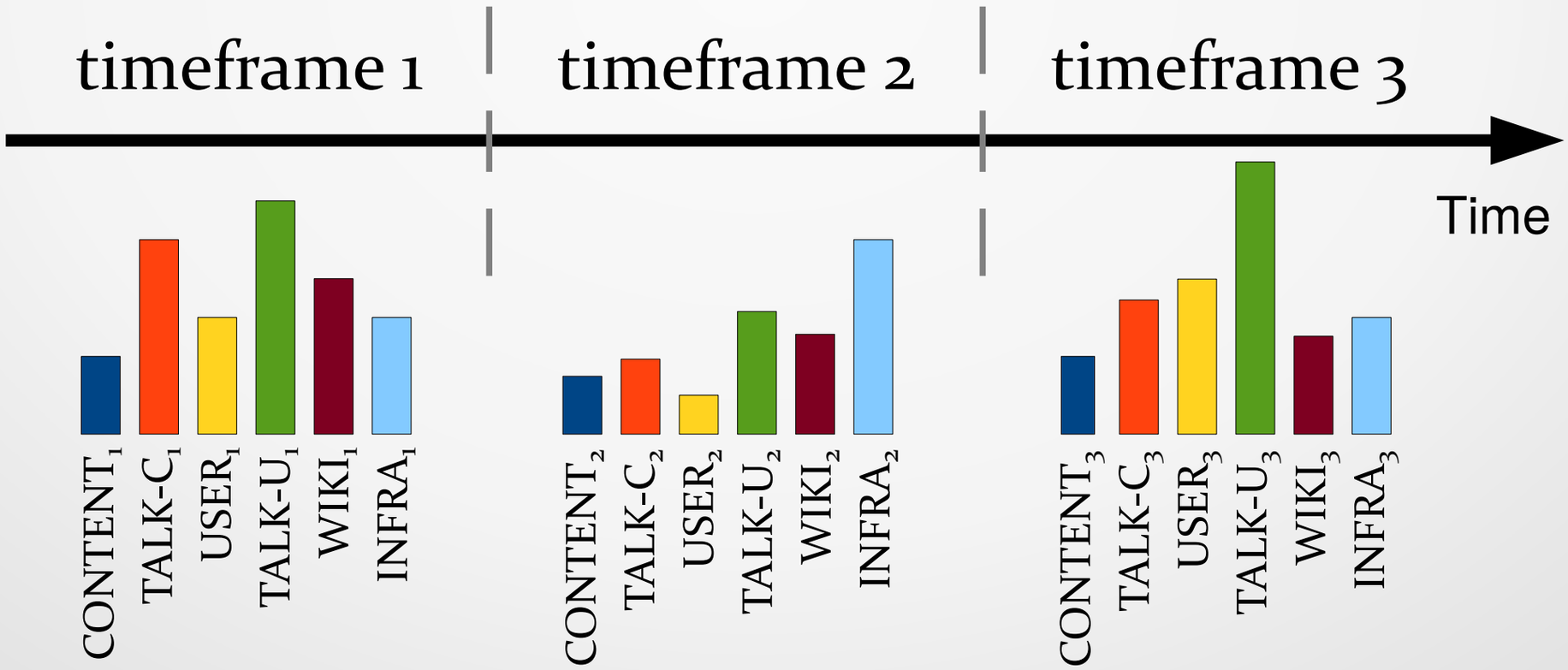
	Feature name	Namespaces
Basic feature set	CONTENT	0, 6
	TALK-C	1, 7
	USER	2
	TALK-U	3
	WIKI	4, 5
	INFRA	8, 9, 10, 11, 12, 13, 14, 15, 100, 101

Editor personal information:

- Extracted from the badges editors put on their editor pages;
- **Gender (6936 out of more than 117k)**, ethnic origin, **religious views (7685)**, **education (9224)**, sexual orientation *etc.*

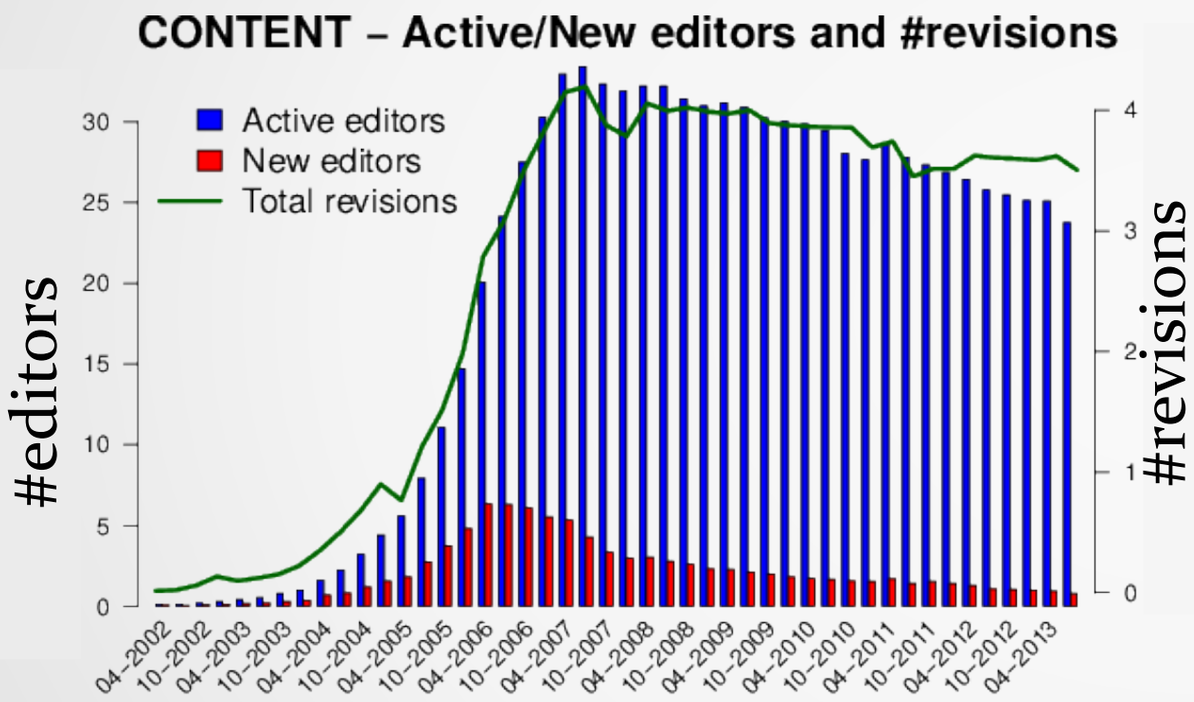
Encoding editing behavior (2)

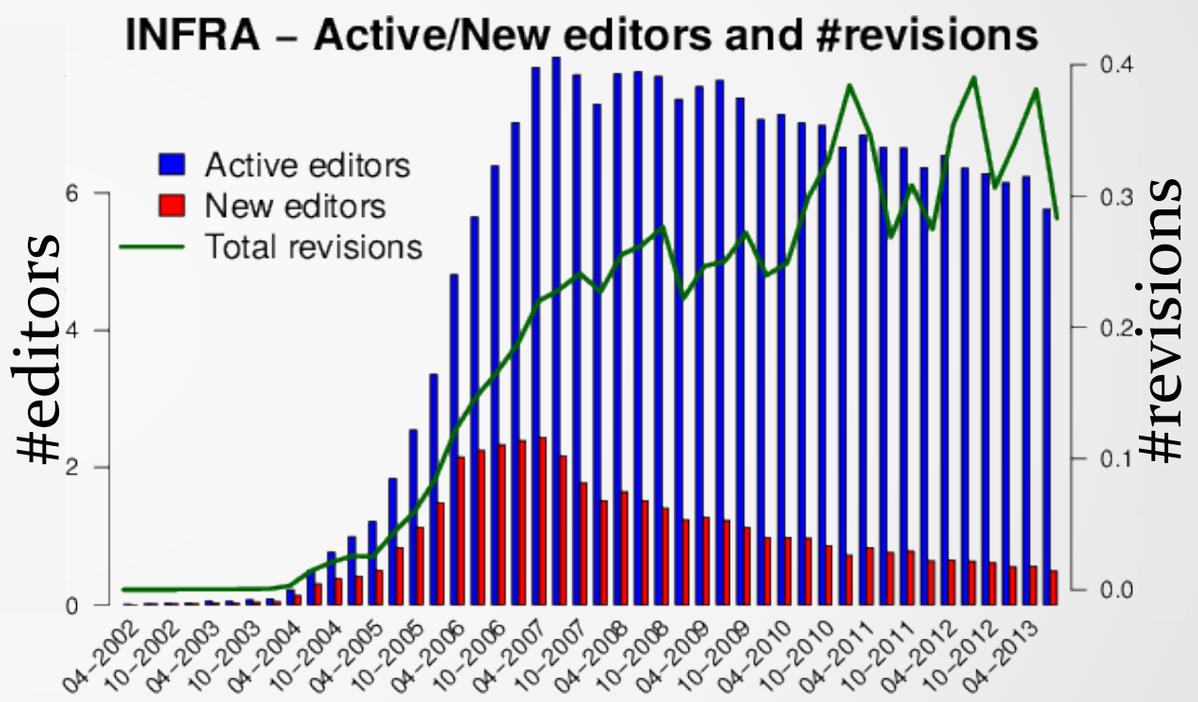
- 3-month timeframes
- description for each editor per timeframe, each feature counts revision over categories
- feature set temporally embeds increasing amounts of information



Profiling of editing behavior (1)

Decline of editorship and rise of maintenance

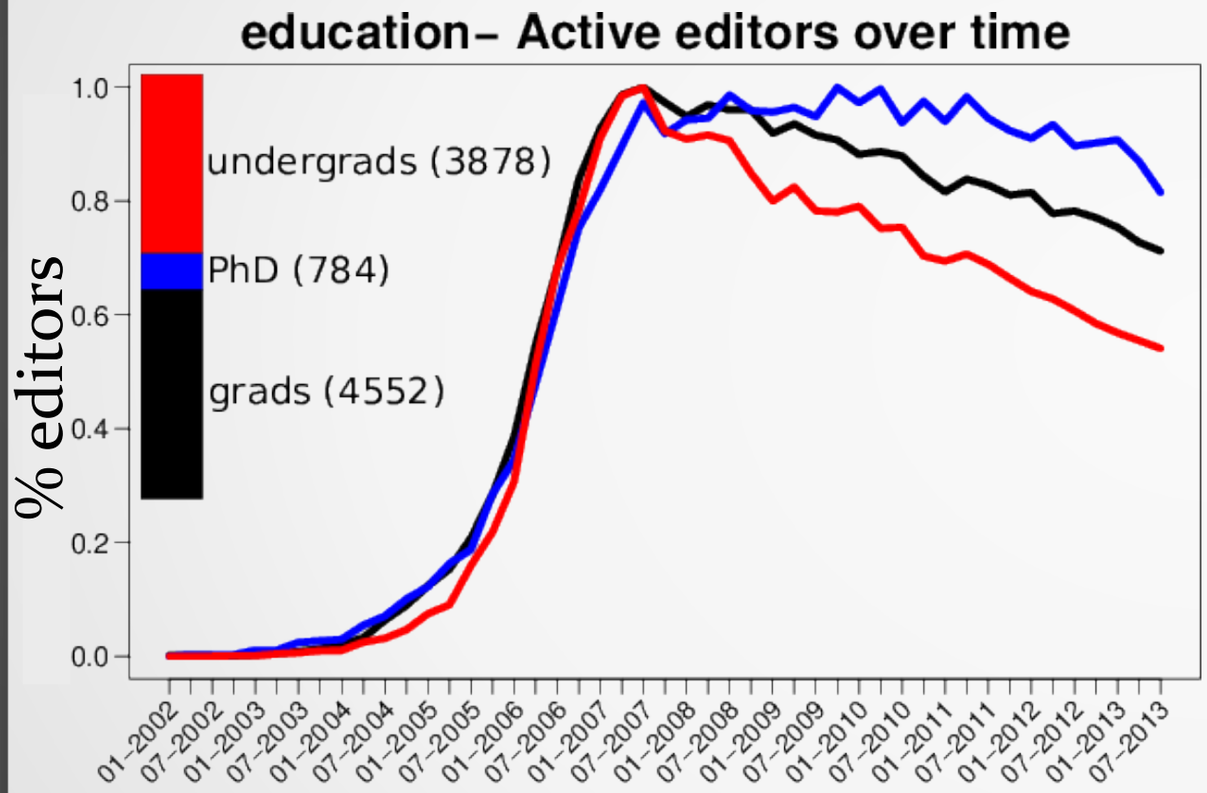


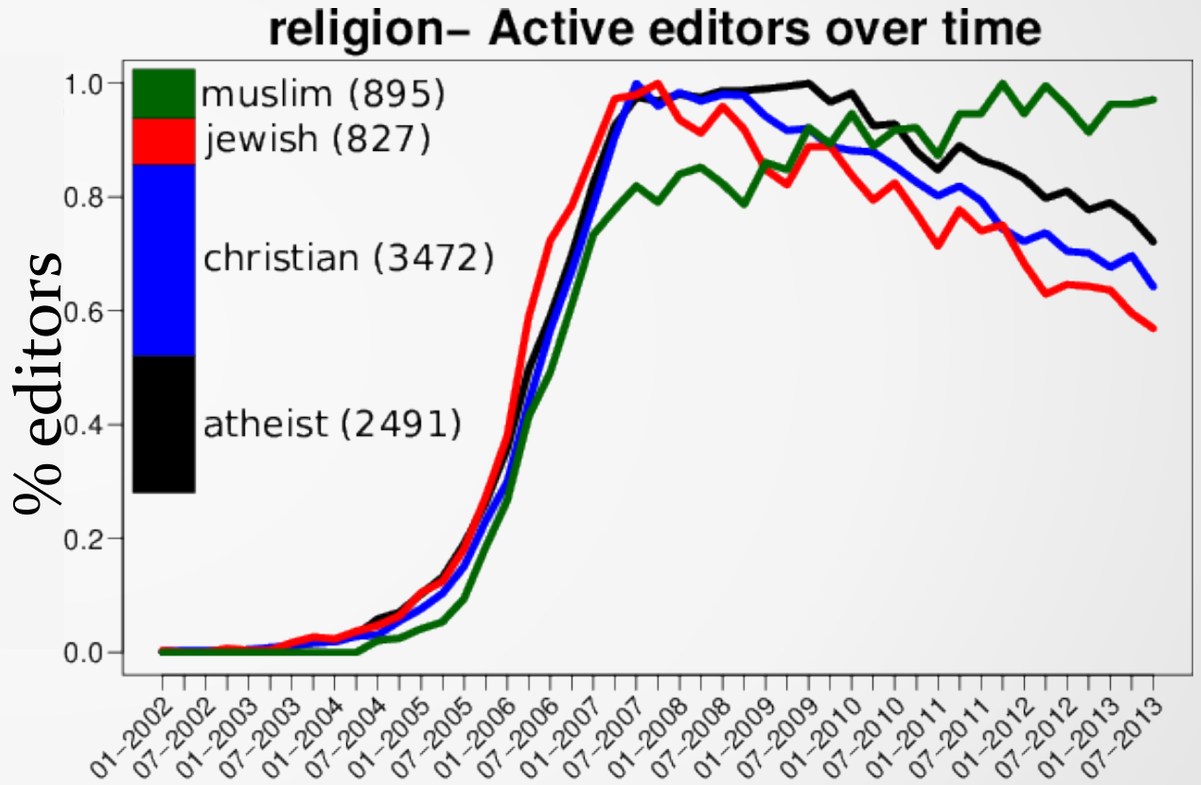


While the Wikipedia "slowdown" has been previously reported **[Suh '09, Halfaker '12]**, we break down this evolution per category and detect a *rise of maintenance effort*

Profiling of editing behavior (2)

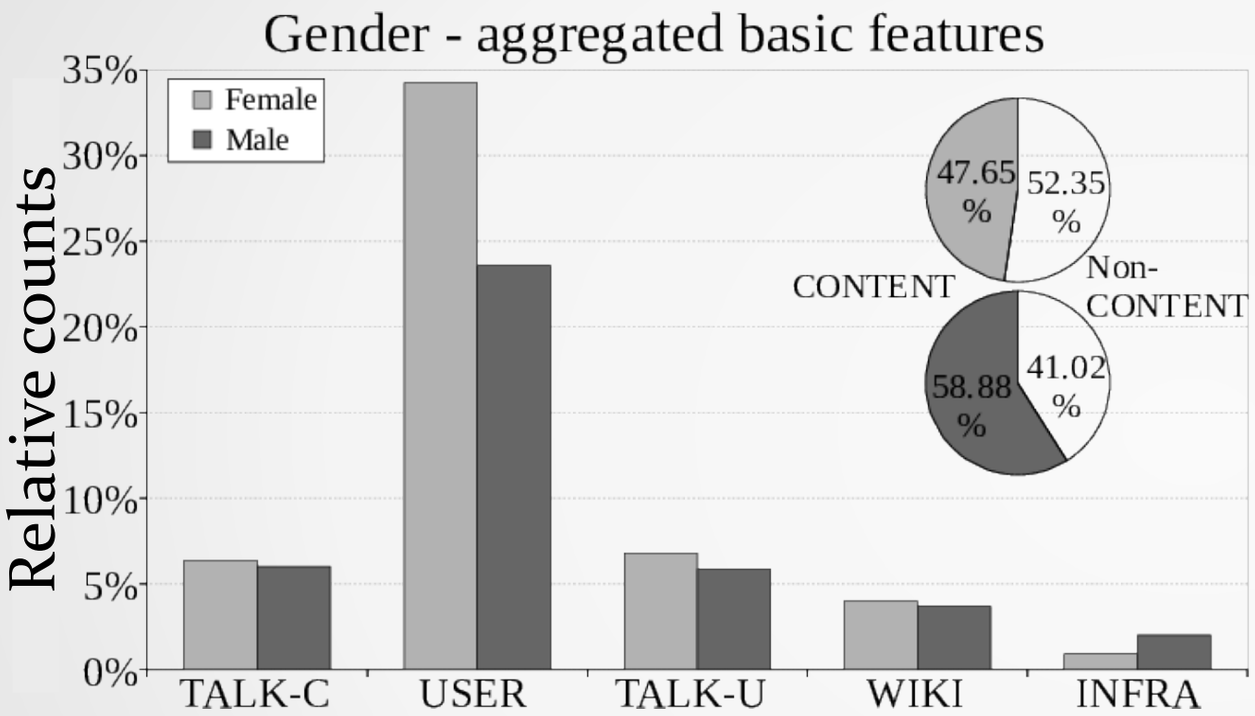
Different growth trends across editor demographics

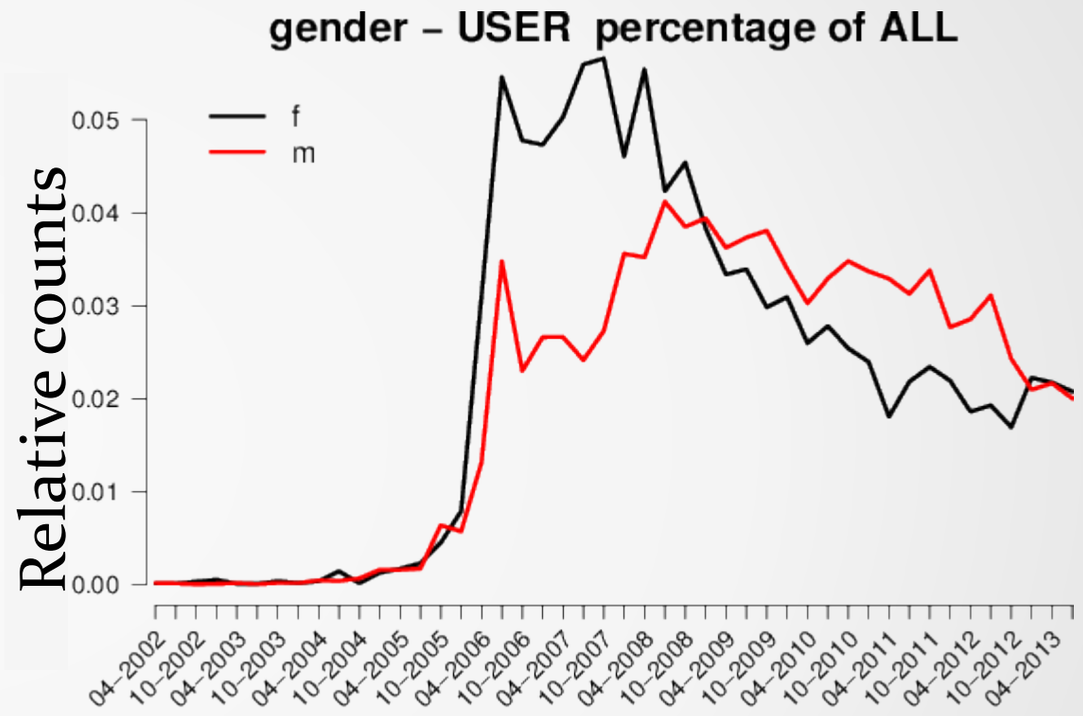




Evolution trends across editor categories are unequal, providing plausible explanations for the slowdown **[Gibbons '12]**, as well as *personal identification clues*.

Edit behavior correlates with private traits

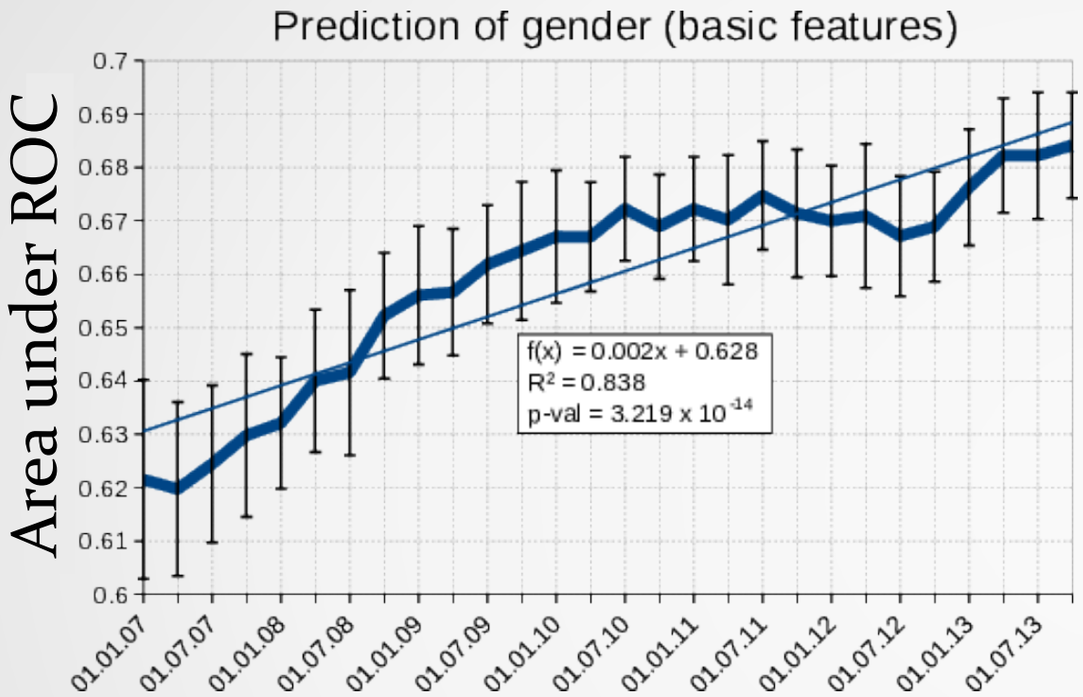


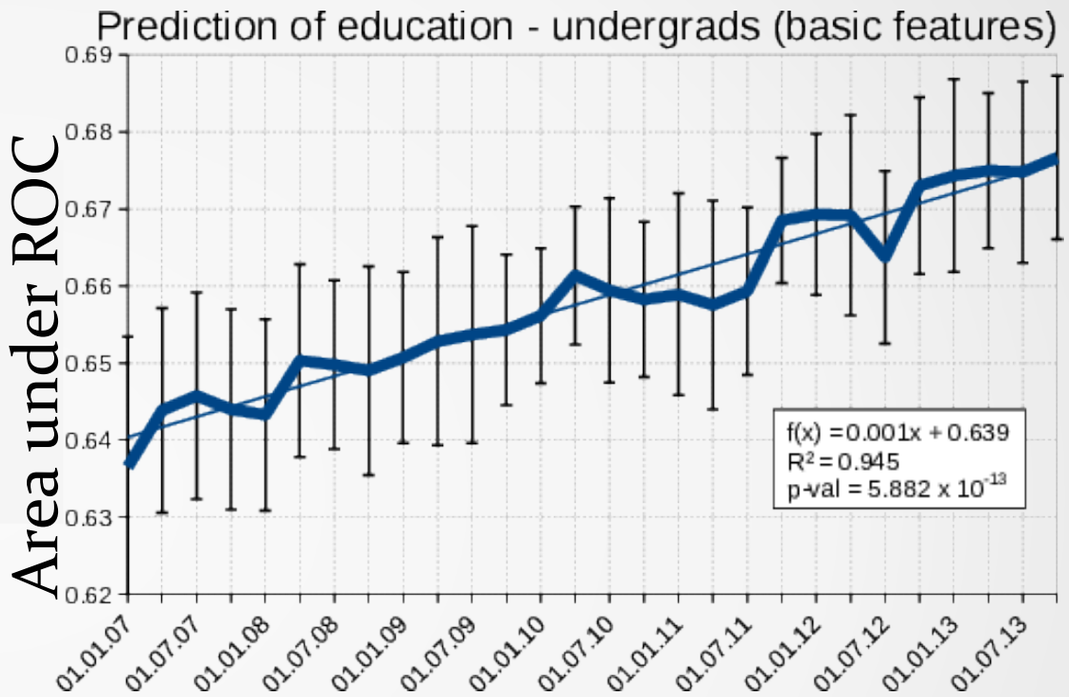


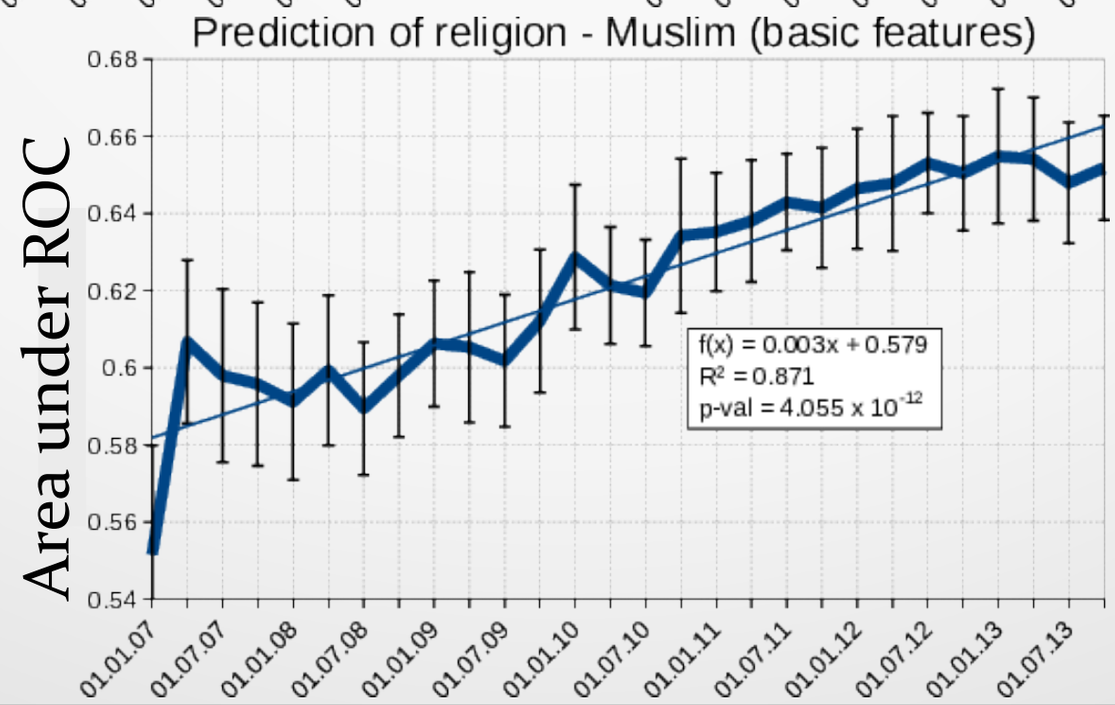
Mean editing behavior analysis shows regularities in the editing patterns for each sub-population.

Predictability improves over time

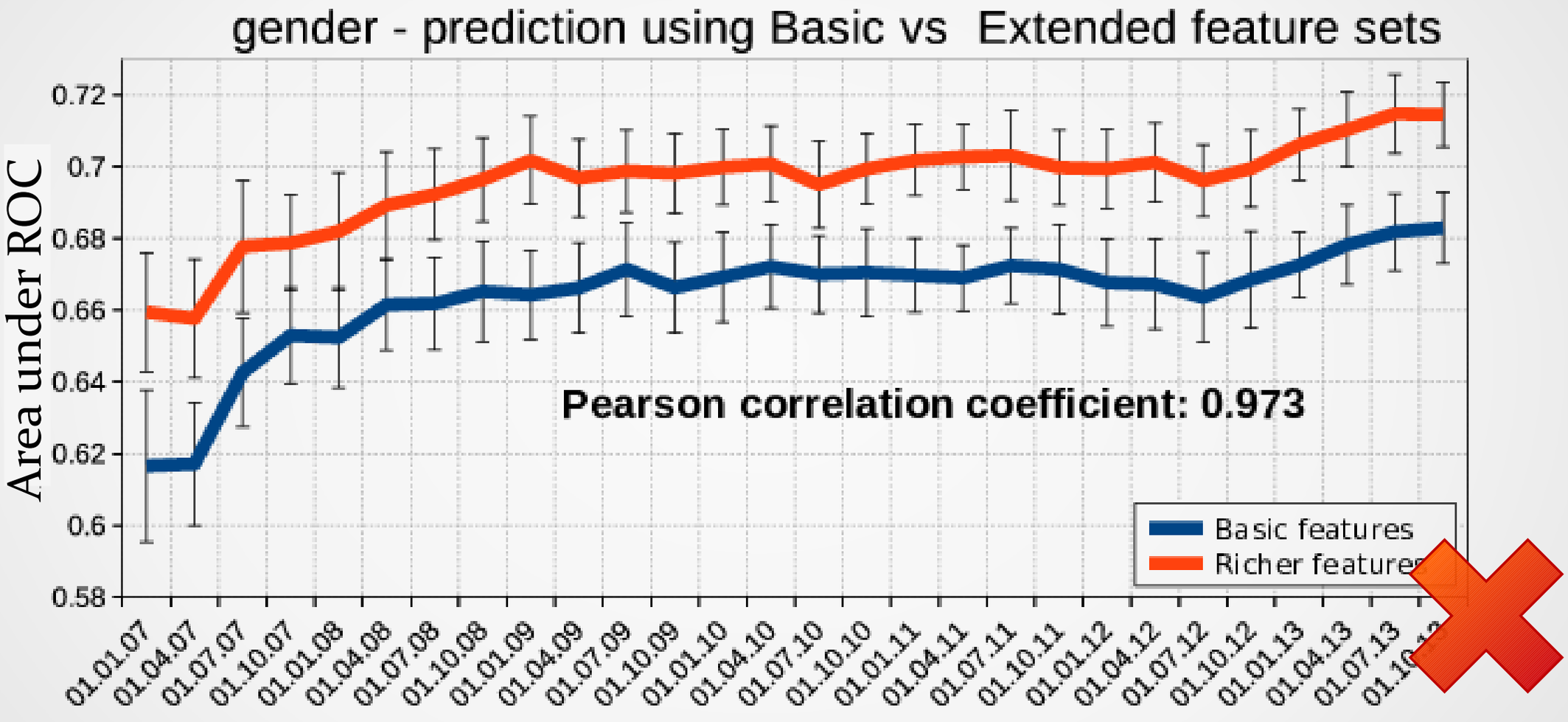
Privacy Loss as a *prediction problem*





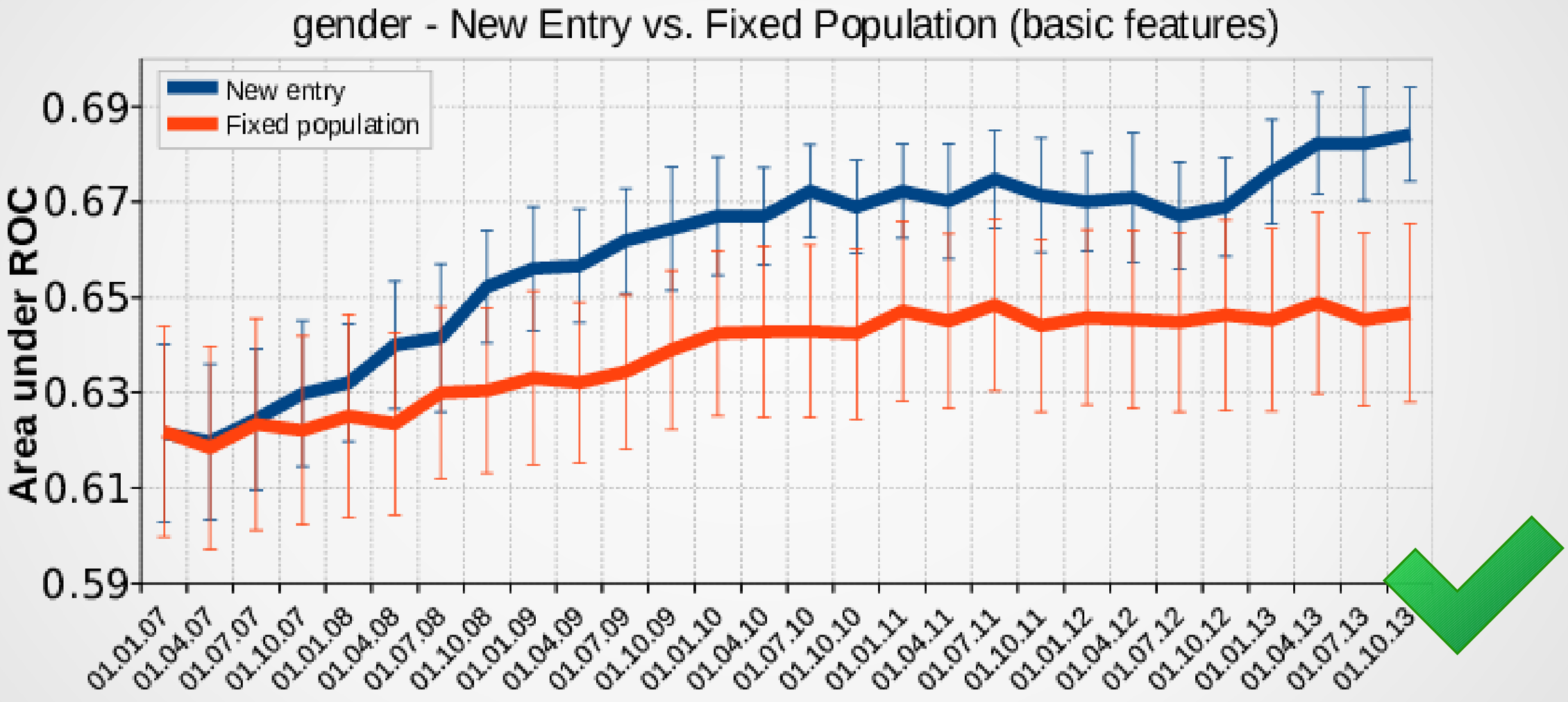


Sources of Privacy Loss (1)



Richer features: improve prediction,
but not *Privacy Loss*

Sources of Privacy Loss (2)



Newcomers: information from *newcomers* hurts privacy

Marginal utility of features over time

Information theory measures:

- Uncertainty about private information → **entropy of target variable Y**

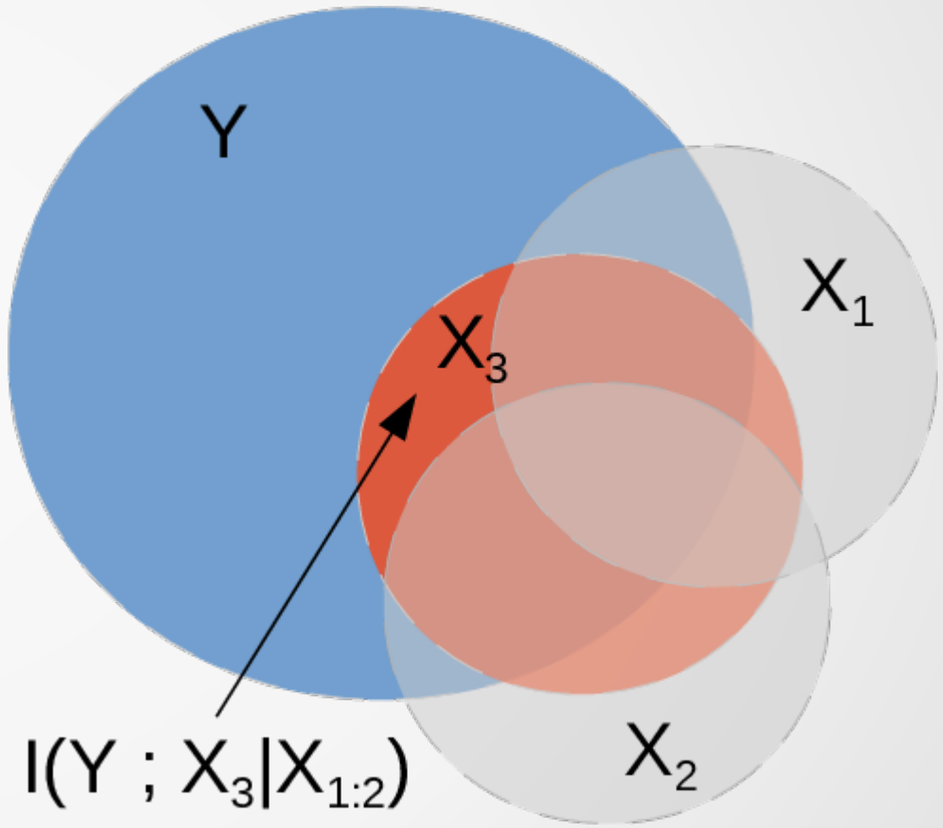
  H(Y)

- Amount of information disclosed by a feature X about Y → **mutual information of X and Y**

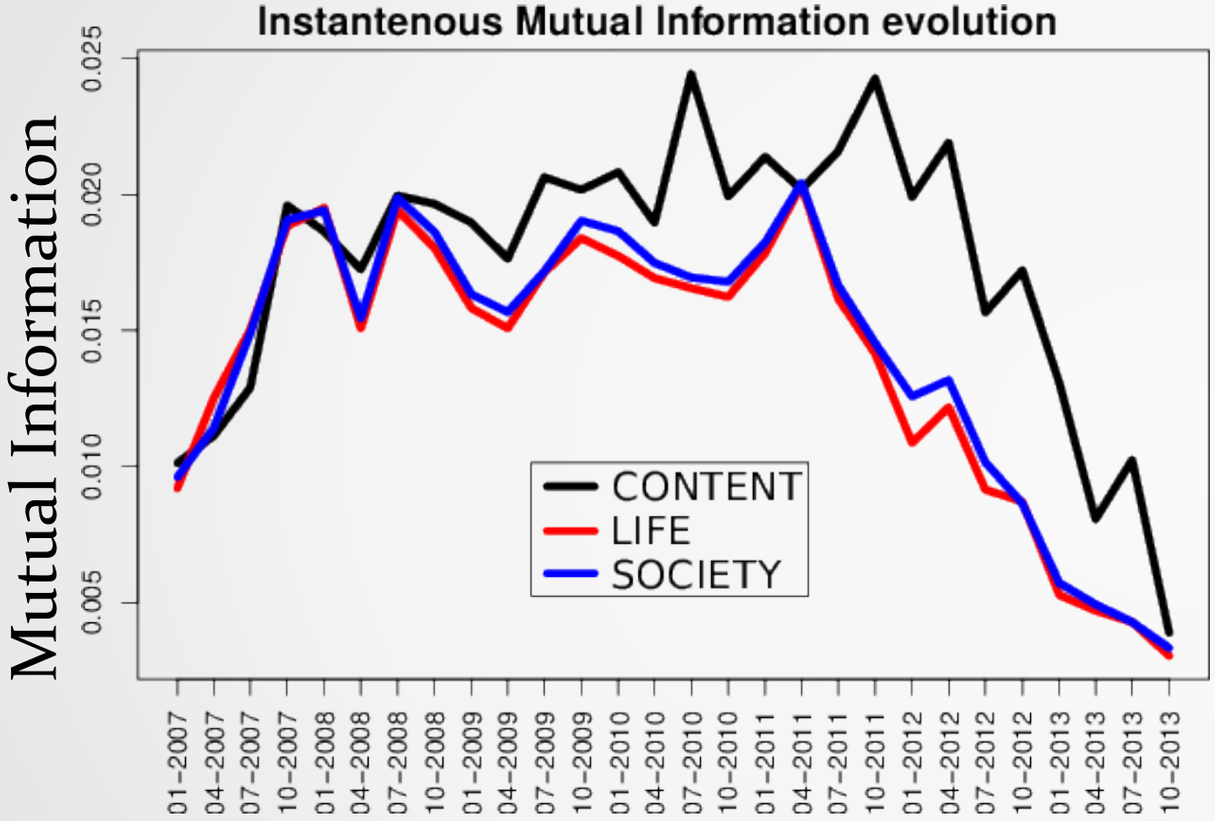
  I(Y; X)

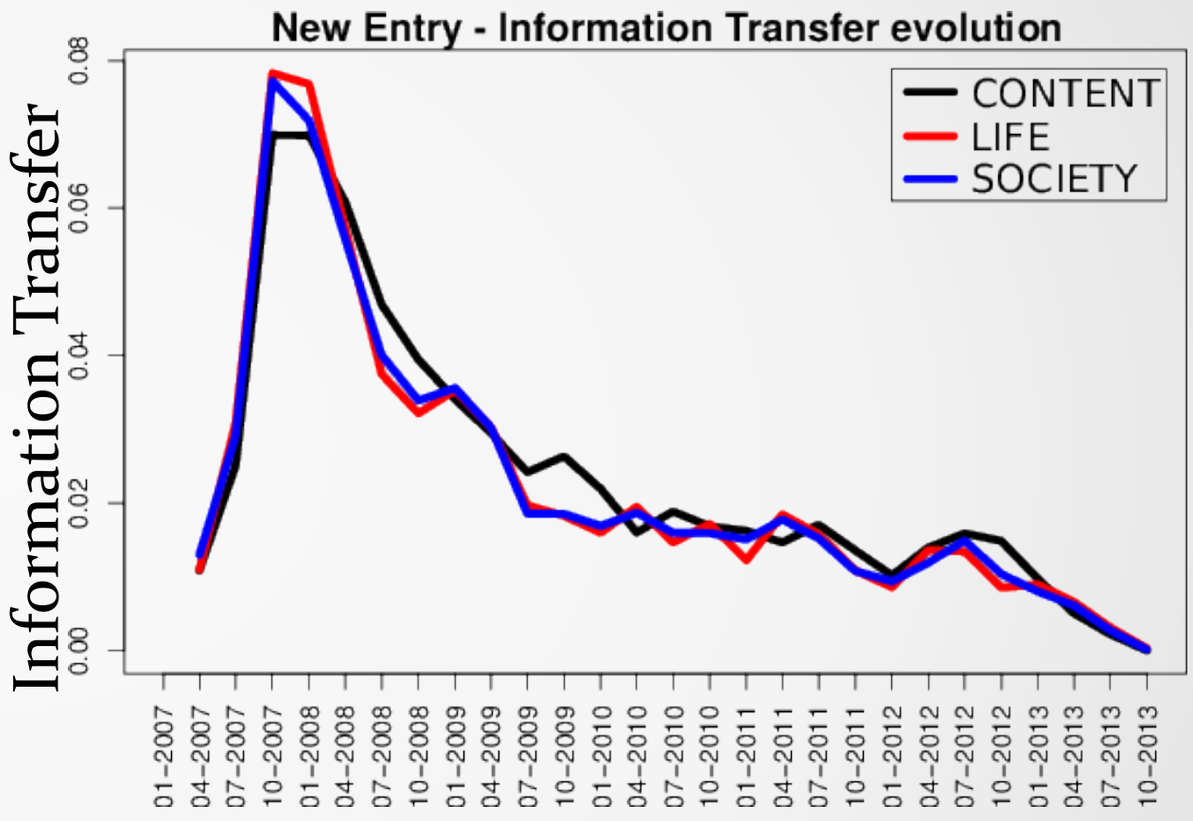
- Amount of *new information* disclosed by a feature at time t X_t → **Information Transfer**

  $I(Y; X_t \mid X_{1:t-1})$



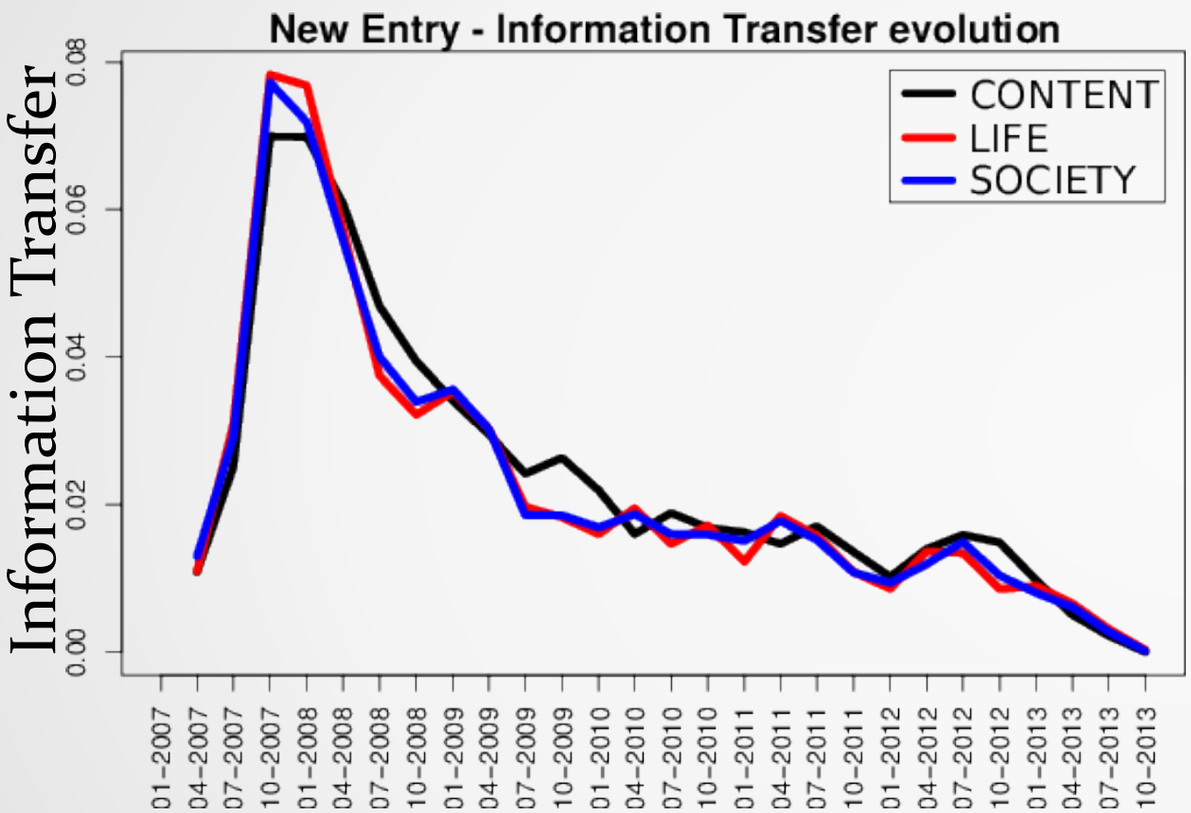
Effect of *online breadcrumbs*

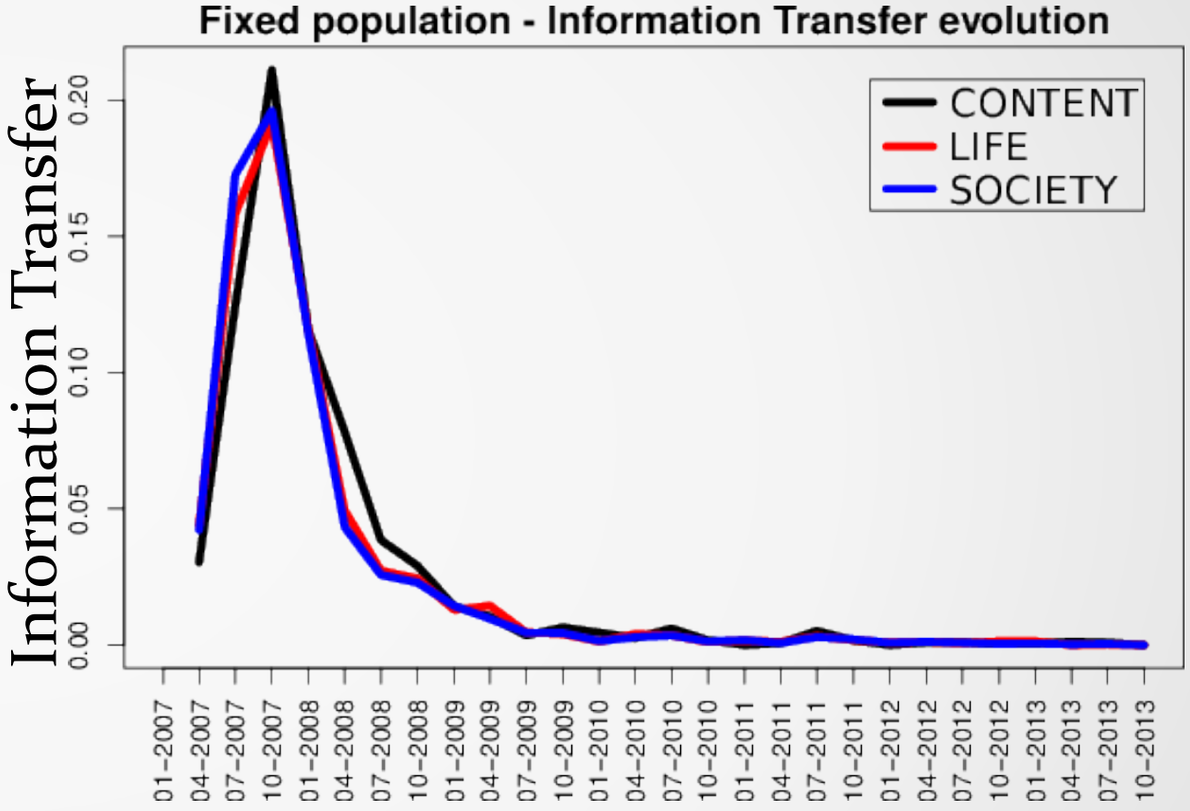




While later edits contain just as much information about a user's privacy as the earlier edits, they tend to be less harmful since most of the information they bring has already been learned.

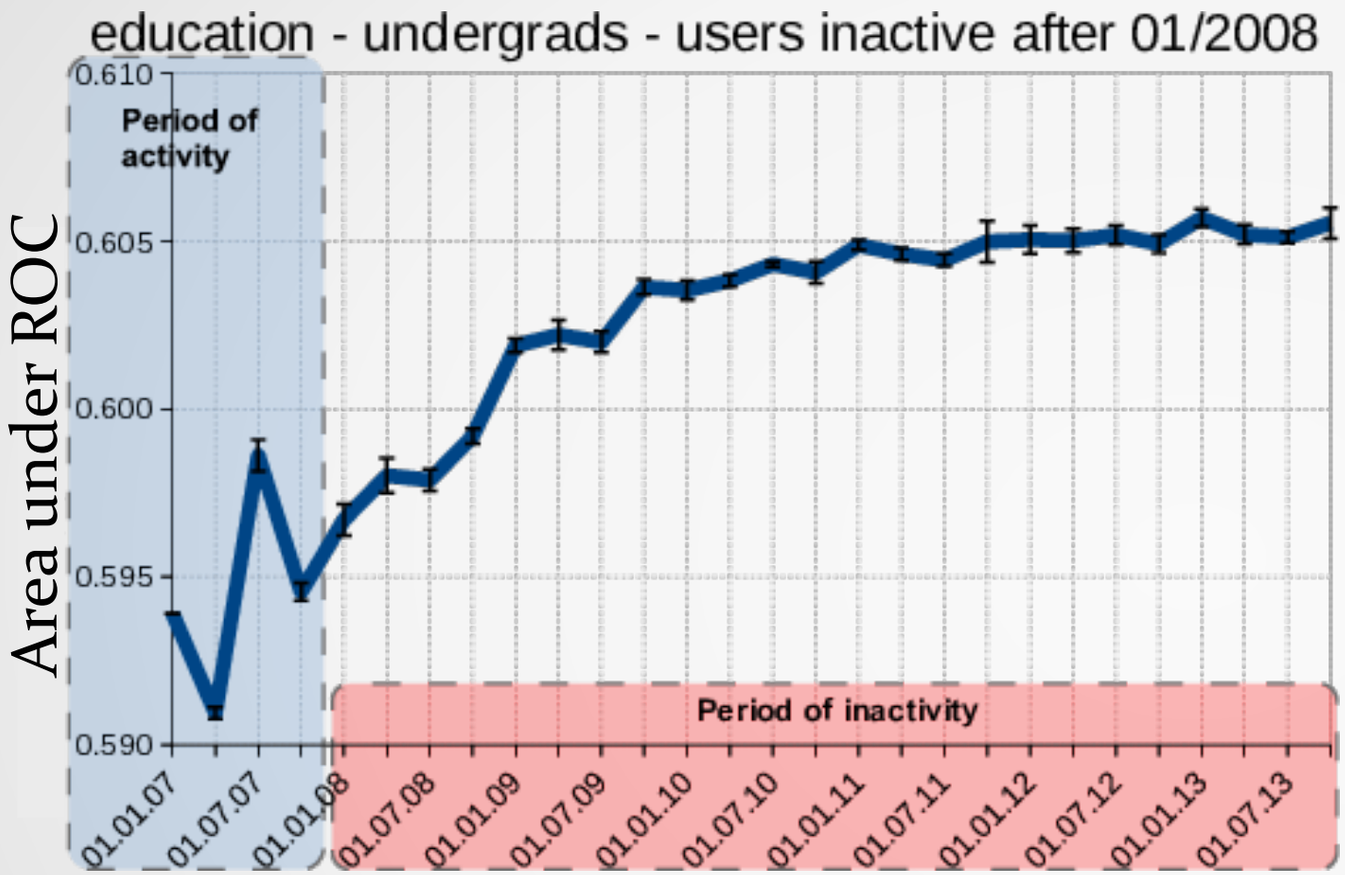
Effect of *newcomers*

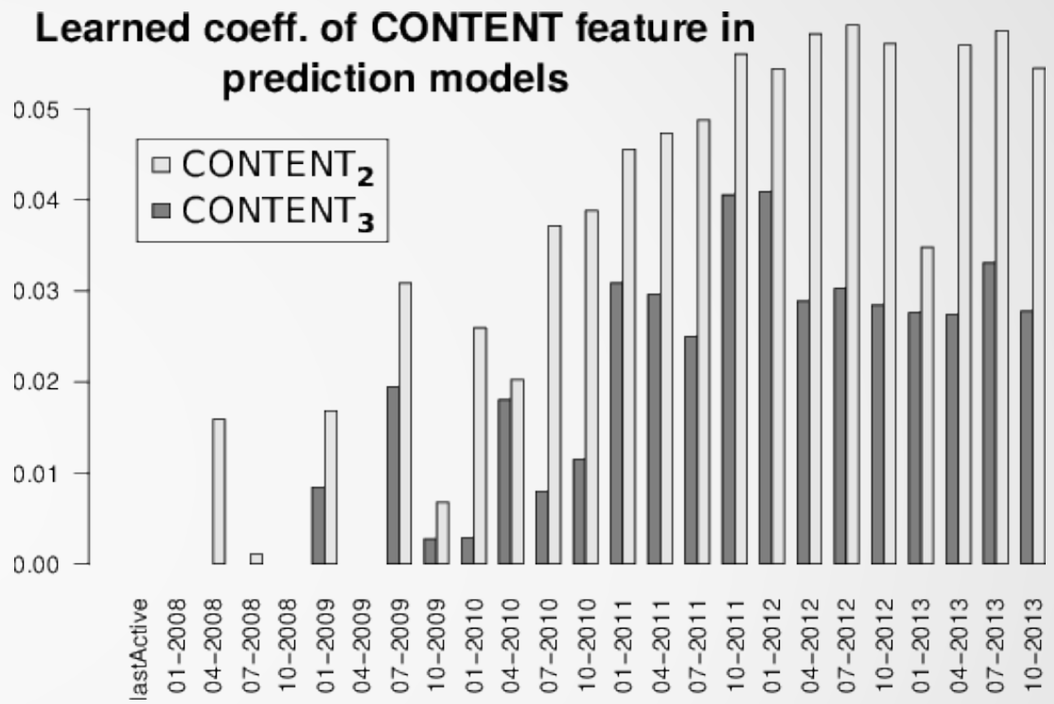




The information inferred from newcomers seems to be moderate, but consistent over time.

Privacy erodes even for *retired editors*





Plausible explanation: observed prediction improvement originates with currently active editors, whose activity overlaps with exited editors

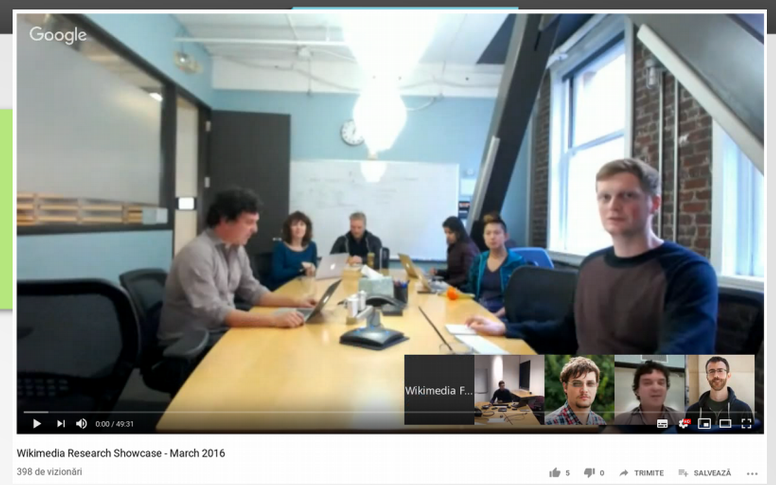
Conclusion

3 main conclusions:

- Time has an adverse effect on privacy
- Factors influencing Privacy Loss:
  - *online breadcrumbs* (i.e. editor's own activity)
  - activity of other editors and newcomers
- Privacy erodes even for *retired* editors

Users don't have complete control over the consequences of the information they release

Thank you!



https://www.youtube.com/watch?v=Xle0oOFCNnk

3 main conclusions:

- Time has an adverse effect on privacy
- Factors influencing Privacy Loss:
  - *online breadcrumbs* (i.e. editor's own activity)
  - activity of other editors and newcomers
- Privacy erodes even for *retired* editors

Users don't have complete control over the consequences of the information they release