Evolution of Privacy Loss in Wikipedia

Marian-Andrei Rizoiu, Lexing Xie, Tiberio Caetano & Manuel Cebrian
Australian National University & Data61

1 The problem

- Does personal traits prediction improve with time? **Yes!**
- What factors contribute to inferring private information? **one's own activity (*online breadcrumbs*) and newcomers in the population**
- Can I stop leaking personal information if I stop posting online? **No! Prediction sometimes improves even after users retire.**

2 Case study

Wikipedia dataset – ideal for longitudinal study, editing focuses on content, not social information.

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

Editing behavior description

Editor badge examples

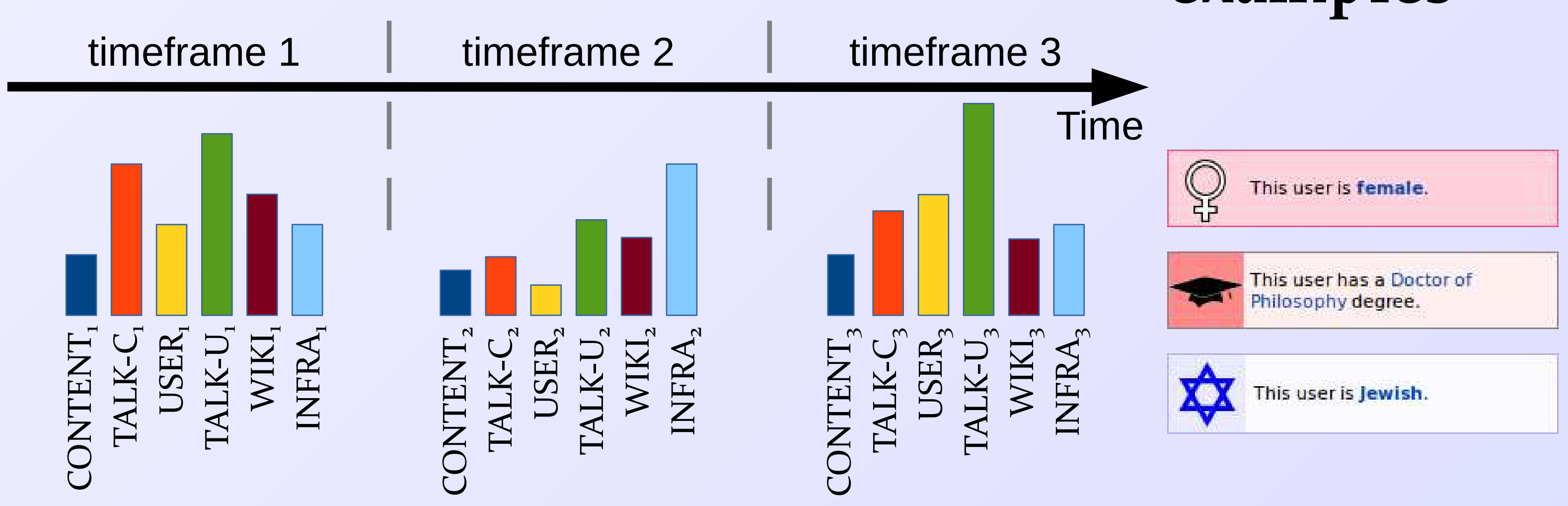


Figure 1: Left: example of constructing user descriptions which embed temporally increasing amounts of information. Features count the number of revisions made by a given editor during a given timeframe, over 6 predefined categories (basic set). **Right:** example personal information extracted from editor badges: gender (6936 ed.), religious views (7685), education (9224) and other not shown (ethnic origin, sexual orientation *etc.*)

Profiling editing behavior

The slowdown of Wikipedia (*#editors*, *#revisions*) is known, we profile editing effort per category and we detect the rise of maintenance effort.

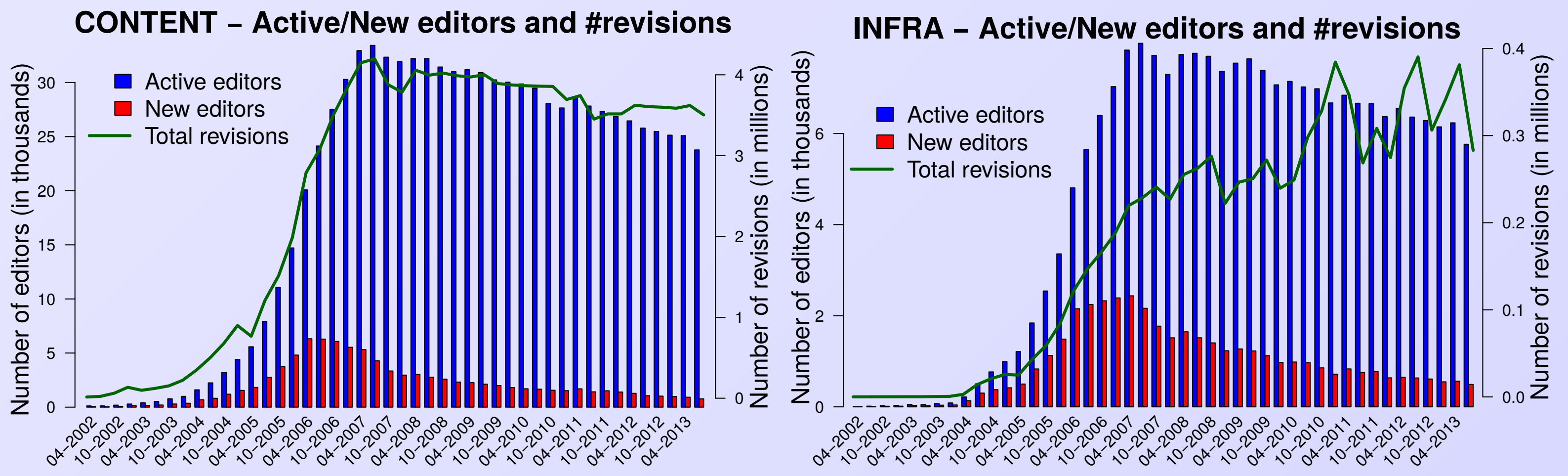


Figure 2: Left: The decrease of the number of active editors, new editors and the total number of revisions for CONTENT. **Right:** The maintenance effort (INFRA revisions) needed to internally handle the bulk of Wikipedia is increasing.

Mean editing behavior analysis shows regularities in the editing patterns, while unequal growth trends across editor demographics provide plausible explanations for the slowdown (*i.e.* editor specialization)

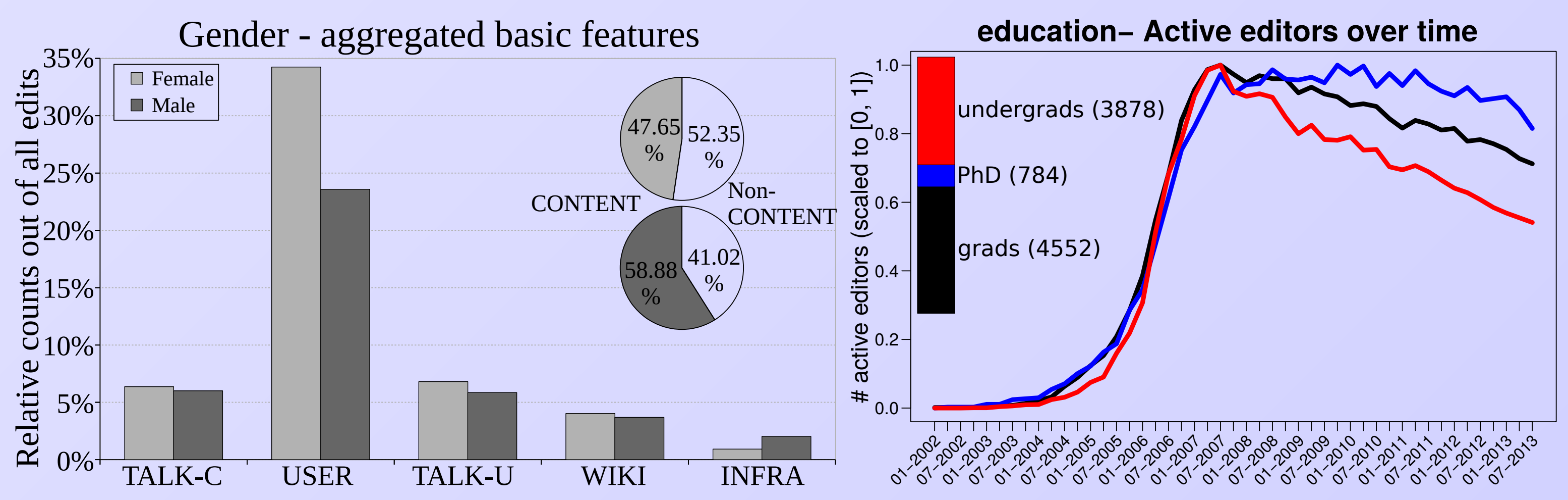
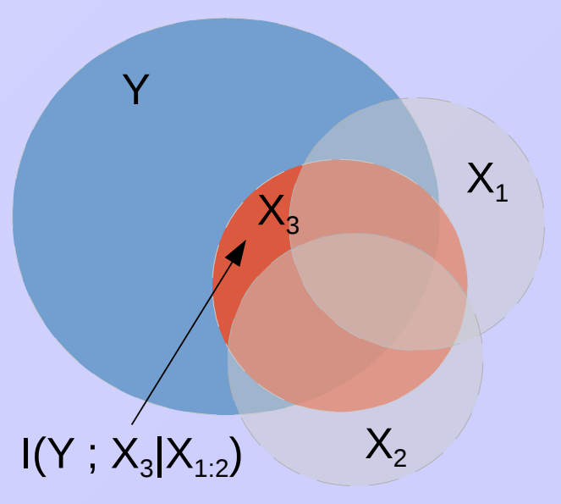


Figure 3: Left: aggregated descriptive features show differences in editing patterns, when tabulated per gender. **Right:** The population of active editors over time, broken down by education. Magnitudes for all classes are scaled from 0 to 1 (effectives in parenthesis).

Methods

Detecting Privacy Loss as a prediction task – private traits are inferred based on editing description. Setup: Logistic regression, L1 reg., hyperparam by cross-validation, 66%-33% stratified split, 20 inits, reporting mean and stdev.
Quantifying Privacy Loss using Information Theory measures – Uncertainty about private information: entropy of target variable Y $H(Y)$. Amount of information disclosed by a feature X: mutual information $I(Y;X)$. Amount of *new information* disclosed by a feature at time t: Information Transfer $I(Y;X_t|X_{1:t-1})$.



3 Trait prediction improves with time

Personal traits are inferred increasingly more accurately over time (differences of prediction performance are statistically highly significant). The steady increase of performance is interpreted as *privacy loss*.

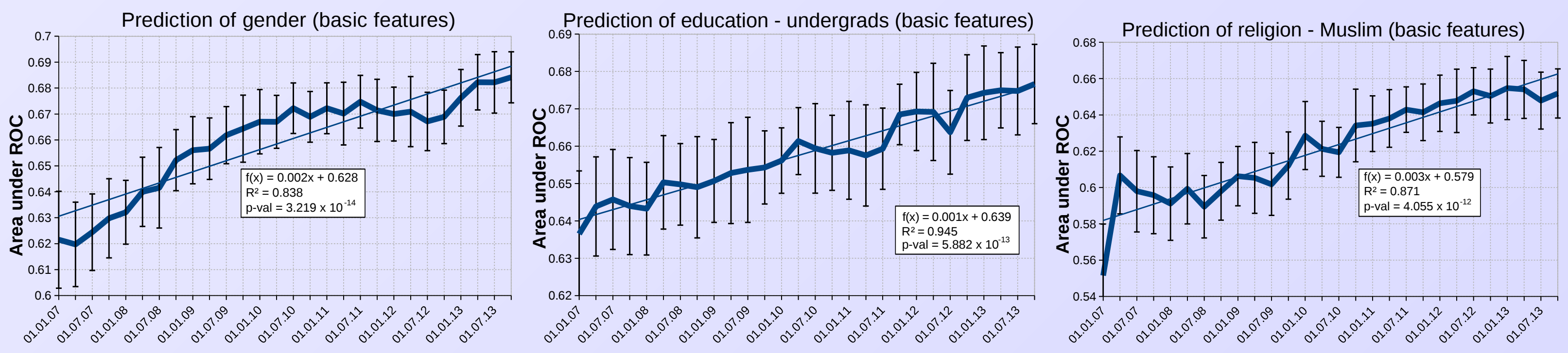


Figure 4: Personal trait prediction performance over time, measured using mean AUC value over 20 executions (error bars denote standard deviation). Result of inferring, using binary predictors on the basic feature set, of gender **(Left)**, education/undergrads **(Center** and religion/muslim **(Right)**.

4 What drives Privacy Loss?

Feature richness vs. editor population composition

Richer features (*i.e.* thematic profiling of editing behavior) improve prediction, but not Privacy Loss. **Newcomers**: information from the editors newly entered in the population hurts privacy.

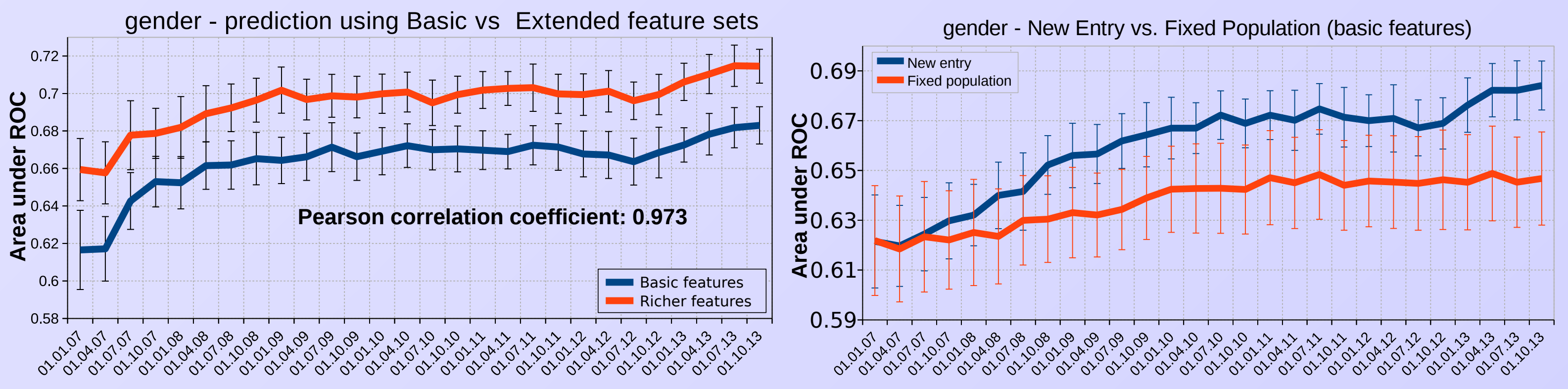
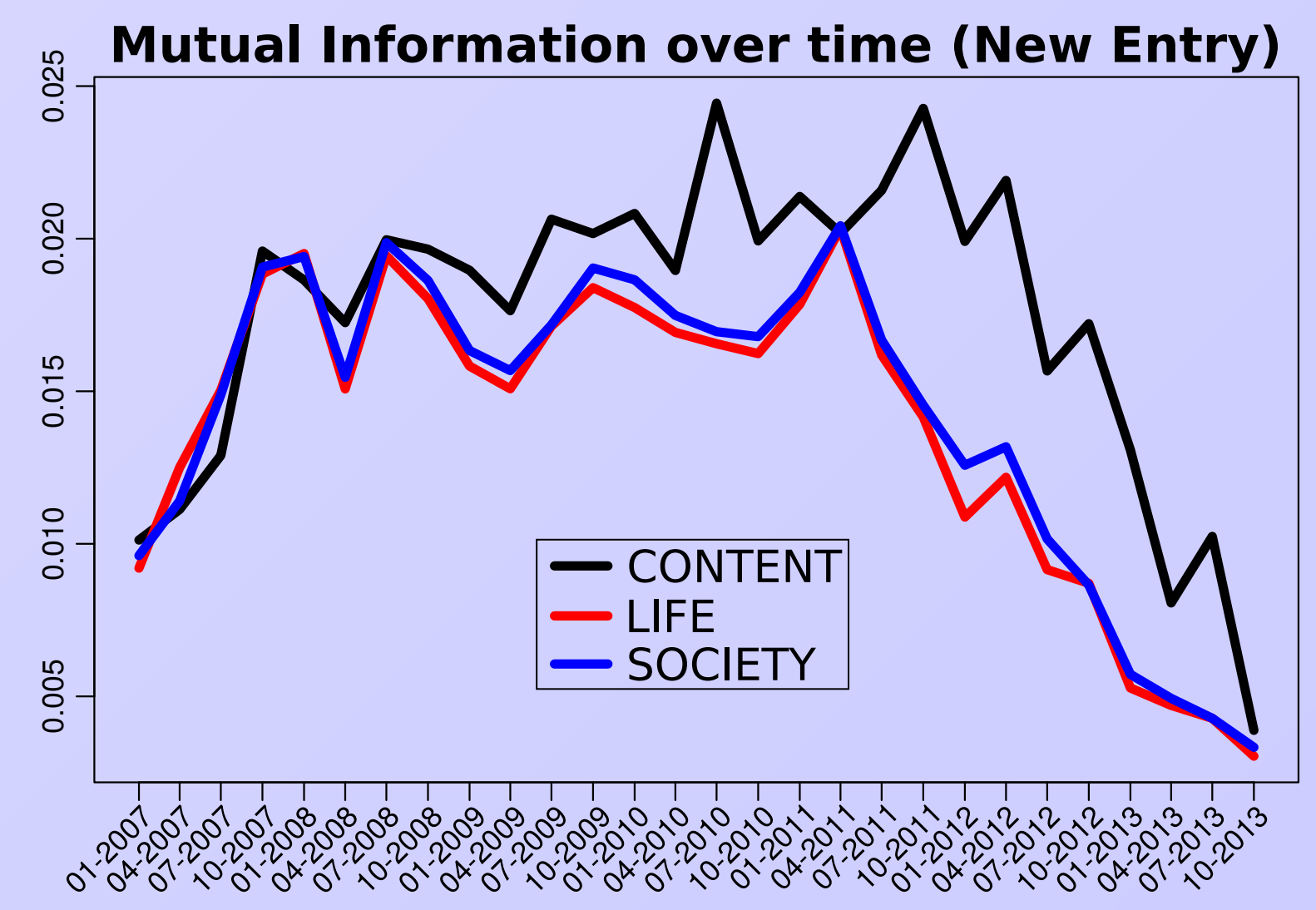


Figure 5: Left: the thematic feature set consistently provides better performances, while the AUC series of the predictors trained on the two feature sets are highly correlated and present the same trends. **Right:** evolution of privacy loss for the population fixed to its component in the first quarter of 2007 (*i.e.*, no newcomers) and a population in which new users can freely enter.

Quantifying privacy loss over time

Later edits contain just as much information as the earlier edits, however they tend to be less harmful since most of the information they bring was already known. The information inferred from newcomers is moderate, but consistent over time.



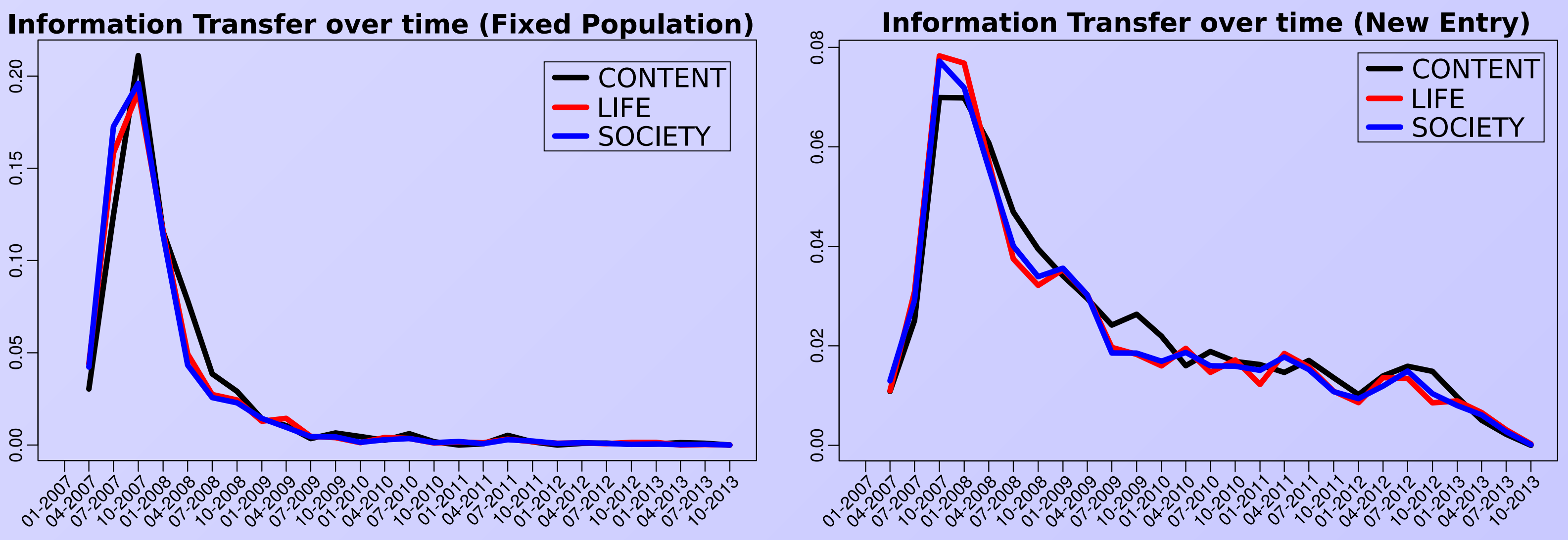


Figure 6: Top: Mutual Information between each feature at time t and gender. Information Transfer on Fixed Population and New Entry **(bottom row, left and right)**.

5 Privacy erodes even for *retired* users

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

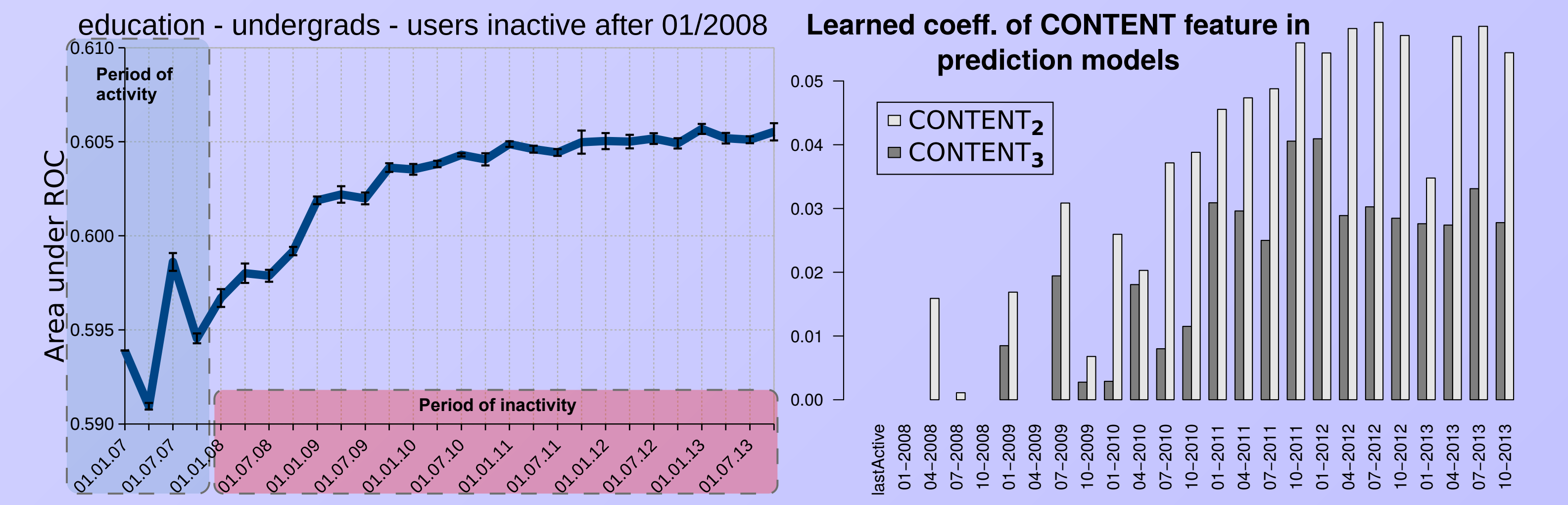


Figure 7: Left: Increase of prediction performance (undergrads) for editors retired after 01.2008. **Right:** CONTENT coeff. increase in importance for models in later timeframes.