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.

1. INTRODUCTION

Privacy is a relatively new concern of modern society [14]. Historically, compromising one's privacy was a difficult task, being mainly achieved by using constant physical surveillance, costly by nature, and easy to thwart. The advent of the online environment has changed the privacy landscape: users of social network, blogging, microblogging platforms 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 novel large-scale behavioral trace dataset, and by measuring the predictability of personal traits in a number of ways.

We construct a new dataset from all editing activities in and around Wikipedia – the largest encyclopedia to date collaboratively constructed by hundreds of thousands of users. We use as input each user aggregated editing activities in a number of broadly defined content and community categories, and the target output are personal traits from Wikipedia *badges*, i.e., what users choose to disclose on their personal pages. This problem and system setting allows us to make several key observations: (1) We show that Wikipedia editors' private traits can be inferred using off-the-shelf machine learning algorithms, and that the prediction performance consistently improves over our prediction period from 2007 to 2013. In particular, our results include predicting an individual's gender, educational status and religious views. Among the different personal attributes, a subset of showed high prediction accuracy as measured by the equal-precision-recall metric – namely, the editors' gender at 0.79, practicing muslim religion at 0.9 or being jewish at 0.91. (2) We quantify the effect of different features using a temporal measure called information transfer. We observe that while the marginal utility of newer features decreases over time, the new users consistently add additional information for the prediction tasks. (3) We show that the prediction of private attributes continues to improve for users who exited the system – or stopped editing after 2007. The continued loss of privacy for these users seems to be associated with two quantifiable factors: the information learned from a user's own activity (or *online breadcrumbs*) and the activity of other editors.

To the best of our knowledge, this is the first work to quantify longitudinal change in privacy loss, carried out on a large dataset and over more than 13 years. Not only do we show that private traits can be predicted increasingly well with time, we also provide several methods to quantify the value of new information over time, and the different source of information loss - from more activity or more users. Our

*Corresponding author: marian-andrei@rizoiu.eu

WSDM'16, February 22–25, 2016, San Francisco, CA, USA.

ACM ISBN 978-1-4503-3716-8/16/02..

DOI: http://dx.doi.org/10.1145/2835776.2835798

findings suggest privacy continues to erode for all users, even after one stops publishing data online. Our findings also can help design data storage and retention policies, can make users aware of the implications of seemingly harmless online activities, and adds to the very lively research topic about online privacy.

2. RELATED WORK

The value of privacy. Social scientists have been interested in how individuals perceive and values their privacy. Acquisiti et al. [2] revealed that the perceived value of privacy, while not entirely arbitrary, is highly malleable. For example, there is a large gap between the amount of money that individuals would accept to disclose private information and the amount of money they would pay to protect it. Furthermore, certain categories seem particularly vulnerable to the online privacy issue. Young adults have been shown [5] to be as worried about their privacy as older adults, especially in what concerns giving personal information to businesses, having photos of them uploaded to the internet or the legislation protecting privacy. On the contrary, Hoofnagle et al. [10] found that young adults tend to expose themselves more, especially on popular social networks, because they are less aware of the risks, less informed about the protection given by law and more prone to social peer pressure. Our work can contribute to the public understanding of privacy risks, by studying on public open data and with quantifiable outcomes.

Privacy Loss and inferring private traits. One definition of Personal Identifiable Information (PII) is private information relating to a person, which can be deductively identified, based on the person's public profile [19]. This is a source of concern particularly in the context of the Open Data effort of governments, in which anonymous datasets are released publicly, after removing private attributes such as name and contact information. The literature shows many applications in which private information, which was never intended to be publicly released, can be inferred from apparently harmless data. In one notorious example, the medical condition of an American politician was inferred starting from anonymized medical records released to the public [23]. Researchers found that de-anonymization can be carried out in large-scale. The 2011 IJCNN Social Network Challenge was won by de-anonymizing the identity of the Flikr users, including those in the test set [18]. More recently, De Montjoye et al. [7] showed that only four spatio-temporal points are enough to identify 95% of individual trajectories using mobile carrier's antenna information, while the same group [17] show that 90% of individuals can be re-identified using their credit card transactions trajectory.

Another salient source of private informations are patterns of online behavior. Kosinski, Stillwell and Graepel showed [12] how private traits like gender, sexual orientation, ethnic origin and even the fact that a user's parents have divorced before her twenty-first birthday can be accurately inferred from apparently harmless, naturally revealed public data, such as Facebook likes. In another data domain, the pattern of an individual's online or phone activity have been shown [21] to reveal precious information about her habits and preferences. Furthermore, computer-based evaluations of human personality have been shown more accurate than those of close friends [27].

We show that private information can be extracted not only from structured anonymized datasets (such as [23]) or datasets rich in social information (such as [12, 21]), but even from data traces left for the public good, such as Wikipedia. Ramachandran and Chaintreau [20] recently studies how the structure of locally connected individuals affects privacy loss. Our focus is in the time dimension – in quantifying the temporal evolution of privacy loss.

Editing behavior in Wikipedia. Wikipedia is the largest online collaborative encyclopedia. In its early years, Wikipedia showed rapid growth. Initial studies [3] explained the growth as driven by the rapidly increasing user base. The growth of the English Wikipedia slowed after 2007, with fewer new editors joining, and fewer new articles created. A few studies [8, 22] explain this dynamic as Wikipedia editors face increasingly limited opportunities to make novel contributions, with the *easy* articles already been created, leaving only more difficult topics to write about. To make useful contributions to the site, editors must also meet an increasingly high bar of expertise in their field. In addition, Halfaker and McNeil [9] consider that Wikipedia's mechanisms for managing quality and consistency deterred newcomers.

In our profile of Wikipedia's growth and decline (Sect. 5), we were surprised to see that the changes of activity over time are not uniform across content and personal demographic attributes. There is a rise in site maintenance, and some user groups showed slower decline (*PhD*), or even a rise (self-identified *muslism* users) in editing activities. Finally, while social interactions in Wikipedia have been studied before [6, 26], to the best of our knowledge, this is the first study of private traits that can be inferred from editing activities.

3. WIKIPEDIA ACTIVITIES AND USER TRAITS

Why Wikipedia? Wikipedia is an ideal data source for studying longitudinal predictability of private traits, due to the following three reasons. Firstly, it is an apparently harmless dataset, whose purpose is to be a reservoir of knowledge, with little or no focus on personal or social information. Unlike online social networks centered on users' profiles, Wikipedia is not intended to record any individual contributor's personal and social interactions. Secondly, Wikipedia's entire edit history is publicly accessible. Wikipedia provides the longitudinal editing history for individuals, spanning over a decade – 2001–2013 at the time of our snapshot. Such a unique long temporal extent allows the study of the effect of time in online privacy. Finally, Wikipedia contributors are from many geographic locations and numerous social, religious, educational and political backgrounds – providing a rich and diverse sample for activities and candidate personal traits. We show that as Wikipedia accumulates user data and editing activity, increasing amounts of private information can be inferred (see Sect. 6).

The Wikipedia dataset. Our dataset contains 13 years of edit history, from the beginning of Wikipedia in January 2001 to July 2013. 188,805,088 revisions are performed by 117,523 editors to 22,172,813 pages. A revision is a Wikipedia term referring to an atomic edit of a page by a user with an associated timestamp. All data used in this study are obtained from July 2013 public Wikipedia dump, more details about data processing are in the supplemental material (SI) [1]. For predicting private traits (Sect. 4 and 6.1) we limit the studied period between 01/2007 and 07/2013 in order to have sufficient numbers of users.

Table 1: Features to describe user editing patterns. (A) The *basic* feature set quantifying the number of edits to all Wikipedia articles and various community and user pages. (B) additional features in the *extended* feature set, encoding edits to the thematic categories within Wikipedia `CONTENT`.

A.	Feature name	Wikipedia namespace codes	Feature description
Basic feature set	`CONTENT`	0, 6	Revisions made to the body of the Wikipedia articles. Corresponds to the actual creation of information.
	`TALK-C`	1, 7	Discussions on the talk pages of the Wikipedia articles. Corresponds to the overhead around the creation of encyclopeadic information.
	`USER`	2	Revisions made to user pages, including a user's own page or another user's page. Similar to profile edits and posts in online social networks.
	`TALK-U`	3	Revisions on the talk page corresponding to user pages. This is similar to social discussions, e.g., writing to a user's wall in a social network.
	`WIKI`	4, 5	Revisions on community pages, help desk, village pump, and related talk pages.
	`INFRA`	8, 9, 10, 11, 12, 13, 14, 15, 100, 101	Revisions on pages that provide infrastructure for other tasks in Wikipedia; template, categories and portals.

B.	***Extended* feature set: the 23 thematic features added to the *basic* set**
	`AGRICULTURE`, `APPLIED-SCIENCES`, `ARTS`, `BELIEF`, `BUSINESS`, `CHRONOLOGY`, `CULTURE`, `EDUCATION`, `ENVIRONMENT`, `GEOGRAPHY`, `HEALTH`, `HISTORY`, `HUMANITIES`, `LANGUAGE`, `LAW`, `LIFE`, `MATHEMATICS`, `NATURE`, `PEOPLE`, `POLITICS`, `SCIENCE`, `SOCIETY`, `TECHNOLOGY`

User activity profiles. We encode a user's activity using two sets of features. In the *basic* set, we count the number of revisions performed in a given period of time, over six predefined categories of the edited pages. The intuition behind these features is to capture the intent of a user's editing effort. For example, the `CONTENT` feature captures edits made to main Wikipedia articles and can be associated with the knowledge creation effort. Similarly, the `WIKI` and `INFRA` features quantify the effort put into organizing the editing effort (*i.e.*, community pages, help desks), while `USER` and `TALK-U` captures the social components, such as constructing a personal page and talking to other users. Features in the *basic* set are based on the Wikipedia namespaces and a summary of their respective meanings is in Table 1. Wikipedia namespaces are organizational categories, to encode the intended purpose of a page (details in SI [1]). The second set of features, *i.e.* the *extended* set, is constructed by adding to the *basic* set 23 new categories based on Wikipedia's top-level category hierarchy. A Wikipedia page can be assigned by its editors to one or more of the 23 thematic categories such as History, Geography, *etc.* The *extended* set provides a more detailed profiling of a users' activity, by capturing their editing interests. Details of feature construction are described in Sect. 4.

The editors' personal information. Many Wikipedia users keep a user page (resembling a social network profile), on which they distribute information about themselves, their interests or the causes they support. Some distribute information typically considered as private, such as their gender, ethnic origin, religion, education, or even sexual preferences. We retrieve these records using public APIs[1], and use them as target personal traits. For the purpose of this study, we selected three private traits with sufficient user bases: gender (declared by 6936 users), education (*undergrads*, *grads* and *Phd*, declared by 9224 users) and religion (*christian*, *muslim*, *atheist* or *jewish*, declared by 7685 users).

4. MEASURING PRIVACY LOSS

We study the loss of privacy by modeling it as a prediction problem: how well can we predict a given a class variable Y (*i.e.*, gender, education or religion) using descriptive features X in the *basic* or *extended* set. By following a set of users through time, we observe the dynamics of the predictive performance. We define the temporal loss of privacy as better explaining a variable linked to a private trait, as we observe editing behaviors for longer periods of time. In this work we use two sets of tools. The first is a predictive approach: we predict the private traits of a hold-out set of users on increasingly longer activity history on Wikipedia, and we observe the change in prediction accuracy. The second approach uses *information transfer*, a measure from physics and economics, to quantify the uncertainty in Y explained by feature X over time.

4.1 Encoding activities over time

We denote a feature X_i^u as computed in the timeframe i for user u, with $X \in \{$`CONTENT`, `TALK-C`, `USER`, `TALK-U`, `WIKI`, `INFRA`$\}$ for the *basic* set, and encoded similarly for the *extended* set. We construct a series of temporal datasets, each having a 3-month period in addition to the previous. A user appears in a temporal dataset if she/he has performed at least one revision during or before the last 3-months. We construct two kinds of features over time, the instantaneous features f_i^u for user u in the i^{th} 3-month period alone, and the longitudinal feature $F_i^u = [f_{1:i}^u]$ for user u – containing the series up to (and including) timeframe i. Features are constructed by counting the number of revisions performed by u during the given timeframe, over the predefined categories, *e.g.*, $f_i^u = (\texttt{CONTENT}_i^u, \texttt{TALK-C}_i^u, \texttt{USER}_i^u, \texttt{TALK-U}_i^u, \texttt{WIKI}_i^u, \texttt{INFRA}_i^u)$ for the *basic* set. Naturally, features F_i^u describe both the past and the current activities, and contain temporally increasing quantities of information. In addition, for newly joined editors, we explicitly encode the missing values in previous timeframes. This is done by including a binary missing feature flags for each activity category, a value of 0 means an editor has joined Wikipedia (even if she is on a pause during timeframe i), and 1 means the editor has not joined Wikipedia, i.e. missing. For the *basic* set, this results in additional six binary features, i.e., $f_i^u = (\texttt{CONTENT}_i^u, \texttt{p\_C}_i^u, \texttt{TALK-C}_i^u, \texttt{p\_TC}_i^u, \texttt{USER}_i^u, \texttt{p\_U}_i^u, \texttt{TALK-U}_i^u, \texttt{p\_TU}_i^u, \texttt{WIKI}_i^u,$

[1] https://www.mediawiki.org/wiki/API:Users

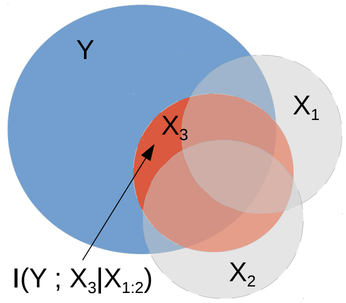


Figure 1: Venn diagram illustration of the Information Transfer measure between target variable Y and feature X, over three successive time points X_1, X_2 and X_3. Even if X_3 explains a large portion of the information in Y, most of Y was already explained by X_1 and X_2, and the new informationX_3 brings is given by the conditional mutual information $I(Y; X_3|X_{1:2})$.

$\texttt{p\_W}_i^u$, $\texttt{INFRA}_i^u$, $\texttt{p\_W}_i^u$), where features prefixed with `p_` are the missing flags of the preceding feature. For the *extended* set, additional features for the 26 categories are constructed in the same manner and appended to each f_i^u. We tried other schemes to encode user activity, such as cumulative features to encode activities from the beginning of until timeframe i, and found them to have lower performance. Therefore the rest of the paper presents only the incremental scheme.

4.2 Predicting personal attributes

We evaluate how well a private trait can be predicted by setting up a set of binary classification tasks. Multi-class target variables (*i.e.*, religion and education) are transformed to binary prediction tasks in a one-vs.-all fashion (*e.g.*, *christian* vs. non-*christian*). We use 2:1 stratified splits to construct the training and test user subsets – 66% of the editors are randomly selected to be part of the training set, and the remaining 33% are used for testing, with the random sampling preserving class priors. One model is learned for each class and each time period, using logistic regression classifier with L1 regularization that favors sparse feature weights, with the hyperparameter obtained by cross-validation [11]. We also tried the L2 regularizer and observed lower performances. The performance is evaluated using the AUC metric (the area under the ROC curve) [11], intuitively random guess classifiers have an AUC of 0.5, and perfect classification has 1.0. Accuracy and F-score for predicting each private trait are given in the SI [1]. We independently sample the train/test split 10 times, and record the mean and standard deviation of the AUC. While other classifiers can be used, we have not tried them in our experiments since our interest lies in the evolution of prediction performance and not its absolute value.

4.3 Information transfer over time

Information theory measures are useful for capturing feature relevance in prediction tasks [13]. One particular measure, Information Transfer, was recently used [24] to uncover hidden links in social media. Intuitively, we capture the uncertainty of private information with the entropy of the target variable Y. The quantity of private information explained by feature X is then given by the mutual information $I(Y; X)$. Since feature X takes different values for each 3-months timeframe, it is useful to quantify the additional information contained in time period X_t that were not already contained by earlier features $X_{1:t-1}$. The *Information Transfer* measure, $I(Y; X_t|X_{1:t-1})$, is designed for this purpose. Fig. 1 illustrated the intuition behind $I(Y; X_t|X_{1:t-1})$ using a Venn diagram. In this example, X_3 contains quite a lot of information about Y – expressed as the mutual information $I(Y; X_3)$, or the large intersection between red and blue circles. However, the *new* information that X_3 provides, in addition to $X_{1:2}$, is much smaller – expressed as $I(Y; X_3|X_{1:2})$, as $I(Y; X_3)$ minus the part already covered by $I(Y; X_1)$ and $I(Y; X_2)$. Intuitively, information transfer is the Conditional Mutual Information of Y and X_t given $X_{1:t-1}$, or the amount of uncertainty that will be reduced after observing X_t:

$$I(Y; X_t|X_{1:t-1}) = H(Y|X_{1:t-1}) - H(Y|X_{1:t}) \ .$$

The relationship above follows from the definition of mutual information and conditional entropy [13]. We implemented information transfer using the `infotheo` toolbox in R [15], which computes conditional entropies and in high-dimensional input spaces by quantizing the input. We use Information Transfer $I(Y; X_t|X_{1:t-1})$ and conditional entropies $H(Y|X_t)$ and $H(Y|X_{1:t})$ to answer two key questions: which features are most important in the disclosure of private information, and which time periods are critical to privacy loss.

5. EDITING BEHAVIOUR OVER TIME

In this section, we present a profile of wikipedia editing behaviour over time. While our profile concur with the slowdown of Wikipedia [8, 9, 22], our analysis detects the rise of maintenance effort and shows that different user groups contribute differently to various functional and topical sections of the encyclopedia.

The decline of editorship and rise of maintenance. It has been observed [8, 9, 22] that Wikipedia's growth slowed since 2007, with fewer new editors joining, and fewer new articles created. Our profiling shows the same phenomenon. Fig. 2 shows the number of active editors, new editors, and number of edits over time. A user is considered active in a time interval if she has submitted at least one revision in the given period. A user is considered a new users (or a newcomer) if she made her first revision in the given time interval. We can see that the slowdown started in 2007, with both the active population, and the total number of revisions decreasing steadily – as seen in the volume of `CONTENT` revisions in Fig. 2a, and revisions to the `Nature` section in Fig. 2b. We can also see that while the overall growth rate is slowing, the increasing amounts of accumulated information require an increasing effort to organize. Fig. 2c shows that the number of infrastructure-related revisions (`INFRA`) continues to increase, made by a decreasing number of users. To the best of our knowledge, this is the first work to detect and quantify this *rise of maintenance*.

Different growth trends across editor demographics. One explanation [8] for the slowdown of Wikipedia is that the easy articles have already been created. This means that in order to make a novel, useful contribution to the site, editors must meet an increasingly high bar of expertise in the field. Fig. 3a plots the active population size for users with a declared education level. The three curves corresponding to *undergrads*, *graduates* and *PhD* have been scaled in $[0, 1]$ to

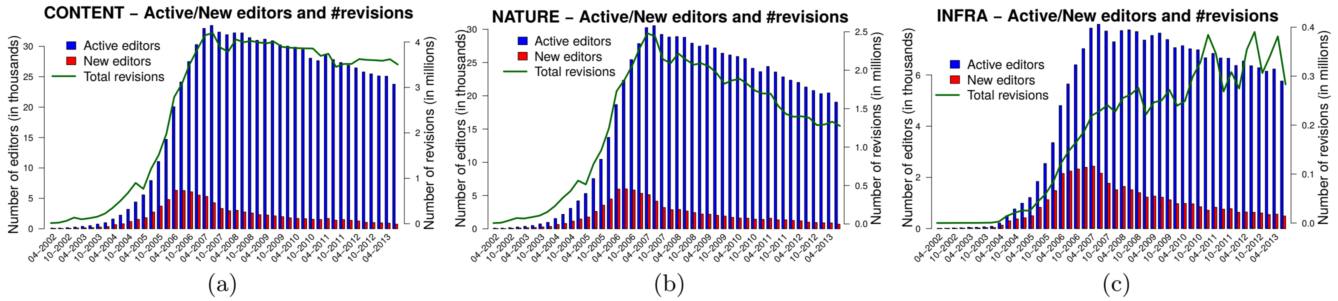


Figure 2: Wikipedia growth slowing down. (a) The decrease of the number of active editors, new editors and the total number of revisions for `CONTENT` (a) and thematic features (shown here `NATURE`, others in the SI [1]) (b). (c) The maintenance effort (INFRA revisions) needed to internally handle the bulk of Wikipedia is increasing.

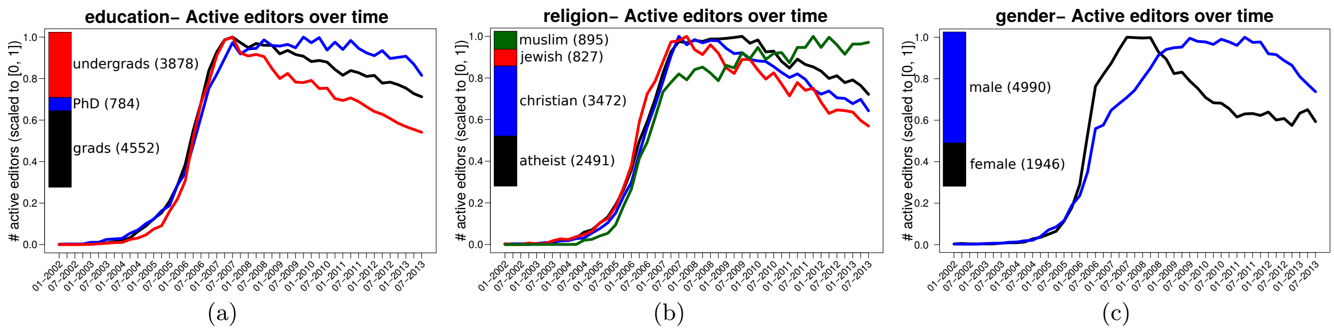


Figure 3: The population of active editors over time, broken down by (a) gender, (b) education and (c) religion. Magnitudes for all classes are scaled from 0 to 1. Barplots show the relative effectives of classes, absolute effectives are given in parenthesis.

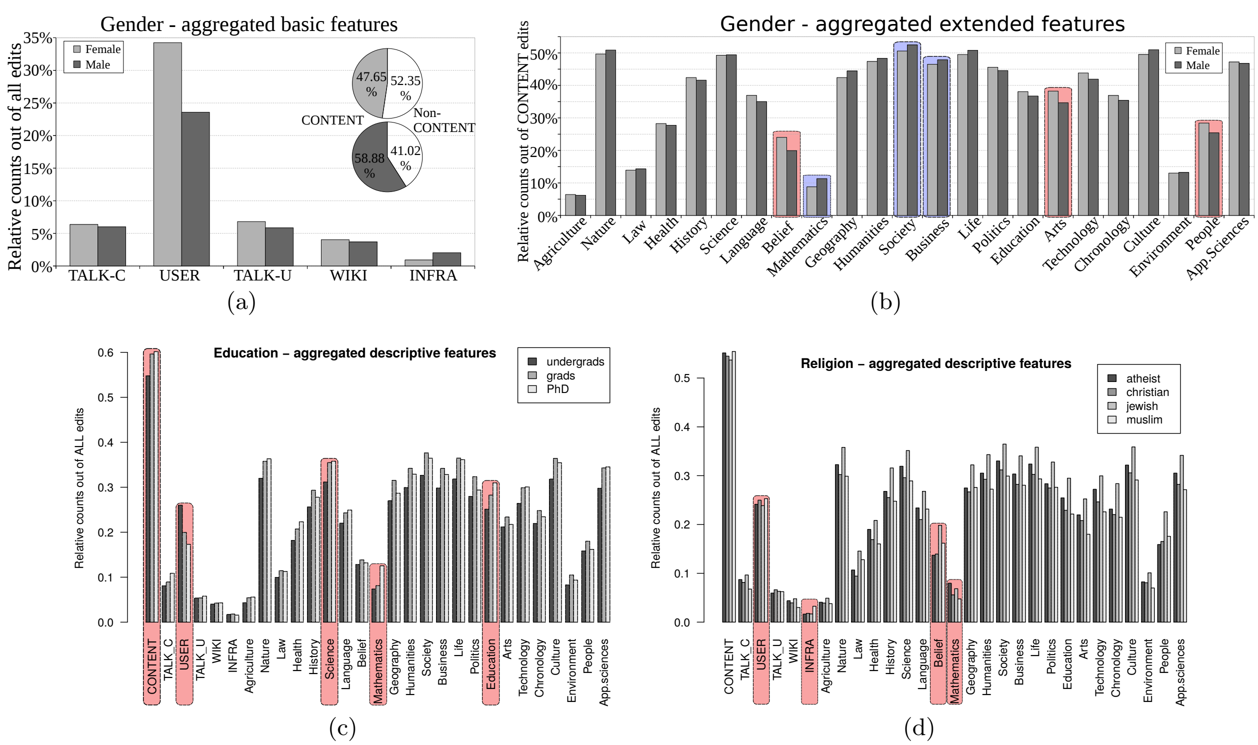


Figure 4: Aggregated descriptive features show differences in editing patterns, when tabulated per gender (a) and (b), education (c) and religion (d). Features were computed as percentages out of the total revision count Particularly interesting features (*i.e.*, features on which the separation is clearer or some patterns are inverted) are highlighted.

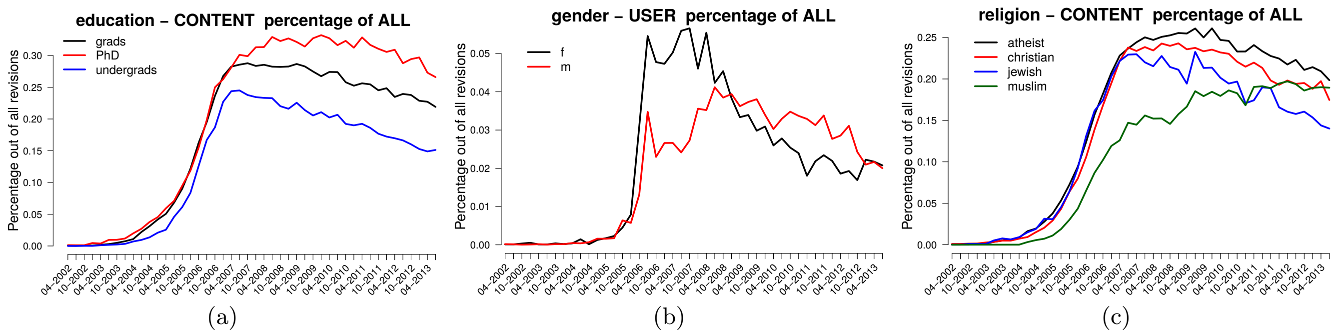


Figure 5: Temporal evolution of mean values of features, broken down by classes in each private trait. We present three selected examples of pairs (feature, trait), other graphics are in the SI [1]: (a) (`CONTENT`, education), (b) (`USER`, gender) and (c) (`CONTENT`, religion). All values are computed as the number of revisions on the given category during a timeframe and expressed as percentages of the total number of revisions. The mean value over all users is presented.

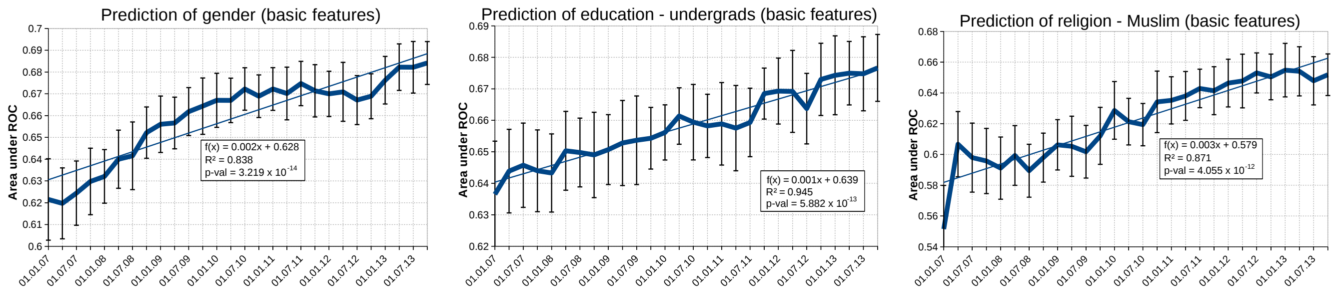
render them comparable. All three population show the expected initial rapid increase. The *undergrads* and *graduates* reach maximum at the same time as the general population The more specialized *PhD* population seems to peak much later, in early 2010, which seems to confirm the hypothesis that the required increase in the specialization of editors is one of the factors responsible for the slowdown of Wikipedia. We can also see differing demographic trends across different groups in editors' religion and gender. In Fig. 3b, the number of *christian*, *jewish* and *atheist* editors start to decrease around 2007. On the other hand, the self-declared *muslim* population seems to continuously increase, at a slower pace than during Wikipedia's initial growth 2001-2007. Fig. 3c plots the number of active editors by gender, we can see that the number of active female editors started decreasing earlier than that of male editors. While gender imbalance in Wikipedia has been previously studied [9], no other discussion of the evolution across time of gender, religion and education is present in prior literature.

Aggregated edit counts correlate with private traits. We describe a user's editing activity by aggregating her revision counts over a number of predefined categories. We conduct a exploratory analysis by presenting the averages of the features, with consideration for each of the private traits. Fig. 4a shows differences between the average male and female behavior: 59% of all the revisions performed by males are `CONTENT`, compared to 48% for females. Females tend to socially relate more, by writing more on `USER` pages (35% of all revisions for females, less than 25% for males). Similarly, Fig. 4b presents average male (highlighted in blue) and female (highlighted in red) behavior, over features in the *extended* set. Females edit more subjects like `Agriculture`, `Health`, `History`, `Language`, `Belief`, `Arts` and `People`, while males edit more `Mathematics`, `Society`, `Business`, `Geography` and `Culture`. While the absolute differences between average behavior on gender tend to be rather small, they indicate a separability of the two classes. We perform a similar analysis for education (Fig. 4c) and religion (Fig. 4d): *undergrads* create less `CONTENT` revisions and more `USER` revisions. *graduates* and *PhD* populations both dedicate a higher percentage of revisions to `CONTENT`. The *PhD* are more active on technical categories, such as `Science`, `Mathematics`, `Education`, `Technology`, `Health` and `Applied Sciences`, and the *graduates* edit more subjects like `People`, `Environment`, `Culture`, `Society` and `Life`. When aggregating per religion, Fig. 4d shows that the *jewish* editors are the most prolific in all thematic sections (the *extended* feature set), except `Mathematics`. *muslim* editors dedicate higher attention to `Belief`, `Language`, `Law` and `INFRA`, and lower attention to `Arts`, `Education` and `Society`. *Atheist* users dedicate more time editing `Mathematics`, `Science`, `Nature` and `Culture`, and less time to `People` and `Law`. This static analysis of mean behavior suggests that there are regularities in the editing patterns for each population. These could be exploited for training a classifier and predicting weather a new user belongs to any of these classes.

Evolution of editing patterns. We further study how editing patterns evolve over time. For each feature, we compute the mean number of revisions over each timeframe, broken down by class. This value is still an aggregate measure over an entire subpopulation, but it evolves temporally, therefore hinting changes in editing patterns. Fig. 5 shows examples of temporal evolution for three selected pairs (feature, private trait). More examples are presented in the SI [1]. Fig. 5a shows the feature `CONTENT` differentiated over levels of education. The static pattern shown in Fig. 4c is a result of change in editing patterns over time: editors with a PhD edit `CONTENT` more as Wikipedia matures, as also shown in the editor population breakdown in Figure 3a. For other features the editing pattern evolves over time. Contrasting Fig 5b and Fig 4a, we can see the differentiation of edits to the USER section by gender – female users edit more overall than male users, but this is only true until 2009. `TALK-C` and `BELIEF` also present the same temporal pattern shift, as shown in the SI [1]. Finally, some features present unexpected trends. Fig. 5c unveils that, unlike the general trend of decreasing number of contributions, `CONTENT` related edits increase for *muslim* editors. This temporal analysis reinforces the hypothesis that user editing patterns, as well as their evolution, are differentiated along different user traits.

6. PREDICTION RESULTS

We predict personal traits of editors using behavioral features described in Sec. 4.1. We report the prediction performance over time using different features. We also perform feature relevance analysis using the information trans-



(a) (b) (c)

Figure 6: Temporal evolution of privacy loss, measure 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 (a), education/*undergrads* (b) and religion/*muslim* (c). The results for all the other binary predictors are given in the SI [1].

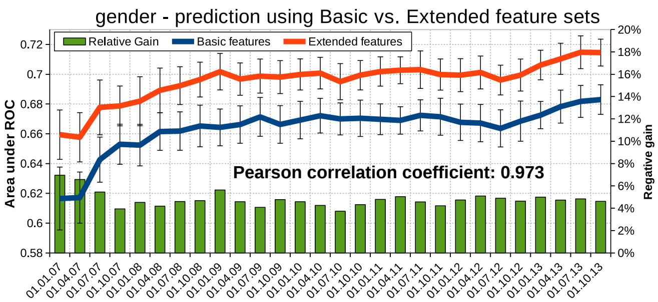


Figure 7: Comparison of the temporal evolution for the privacy loss on gender for the *basic* and *extended* feature sets. The *extended* feature set consistently provides better performances, while the AUC series of the predictors trained on the two feature sets are highly correlated and present the same trends.

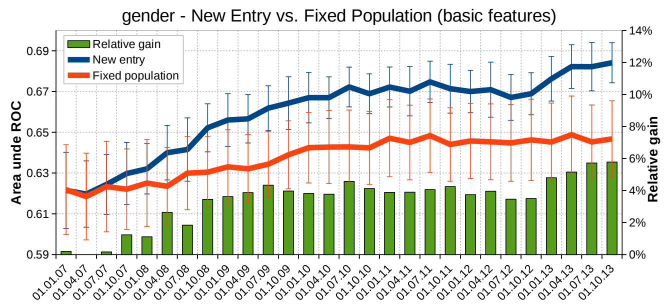


Figure 8: Evolution of privacy loss for the population fixed to its component in the first quarter of 2007 (*i.e.*, no newcomers) and a population in which new users can enter. We quantify the privacy loss due to newcomers as the relative gain of learning performance for the two populations.

fer metric to pin point the source of performance gain over time.

6.1 Predicting personal attributes over time

Predictability of private traits improves over time. We train binary predictors for every class of every private trait in our study, on datasets with increasing amounts of history (as shown in Sec. 4.2). The editors' activity is described using the *basic* feature set. Fig. 6 shows the AUC over time for three selected examples of private traits (all remaining classes are in the SI [1]). The graphics show the performance of predicting the gender of editors (Fig. 6a), whether they are *undergrads* (Fig. 6b) or of *muslim* religion (Fig. 6c). The AUC measure increases over time, roughly following a linear trend (coefficient of determination $R^2 > 0.83$ for all three examples). The AUC differences between the first and the last timeframes are statistically highly significant (t test $p < 0.001$, details and results in the SI [1]). We interpret this steady increase of performance over time as *loss of privacy*: as more historical information is available, the learning algorithm infers more accurately user traits which are potentially private.

We also report the model performance on an intuitive measure called *equal Precision and Recall* (ePR), defined as where a 45 degree line from the origin intersect with the precision-recall curve. For the three classes in Fig. 6, in the last timeframe, we obtain a mean ePR of 0.791 (for gender), 0.535 (for education/*undergrads*) and 0.9 (for religion/*muslim*). All the other classes and the evolution of ePR over time are found in the the SI [1]). The ePR measure allows to quantify how accurate are the predictions of certain private traits. We find that certain religious attributes have notably high prediction performance ($muslim\ ePR = 0.9$ and $jewish\ ePR = 0.913$) from behavioral features.

In a similar setup, we predict user gender using the *extended* feature set and we plot the AUC over time in Fig. 7. Alongside, we produce the results for the *basic* feature set and the relative gain between the two, for each timeframe. Predictions using the *extended* features consistently outperform those using the *basic* features, showing that knowledge about thematic editing patterns is informative about user private traits. The AUC over time series for the two types of features sets are highly correlated (Pearson correlation of 0.973). This shows that, while adding the thematic information improves the absolute value of the prediction accuracy, it seems to have little influence on the evolution of the privacy loss. We speculate that the evolution of privacy loss is not linked to the way the data is described, but it is rather intrinsic to the online social environment. To the best of our knowledge, this is the first study to highlight and quantify this *intrinsic cumulative effect* of time over privacy.

Sources of privacy loss: the *online breadcrumbs* and newcomers. We hypothesize that the temporal pri-

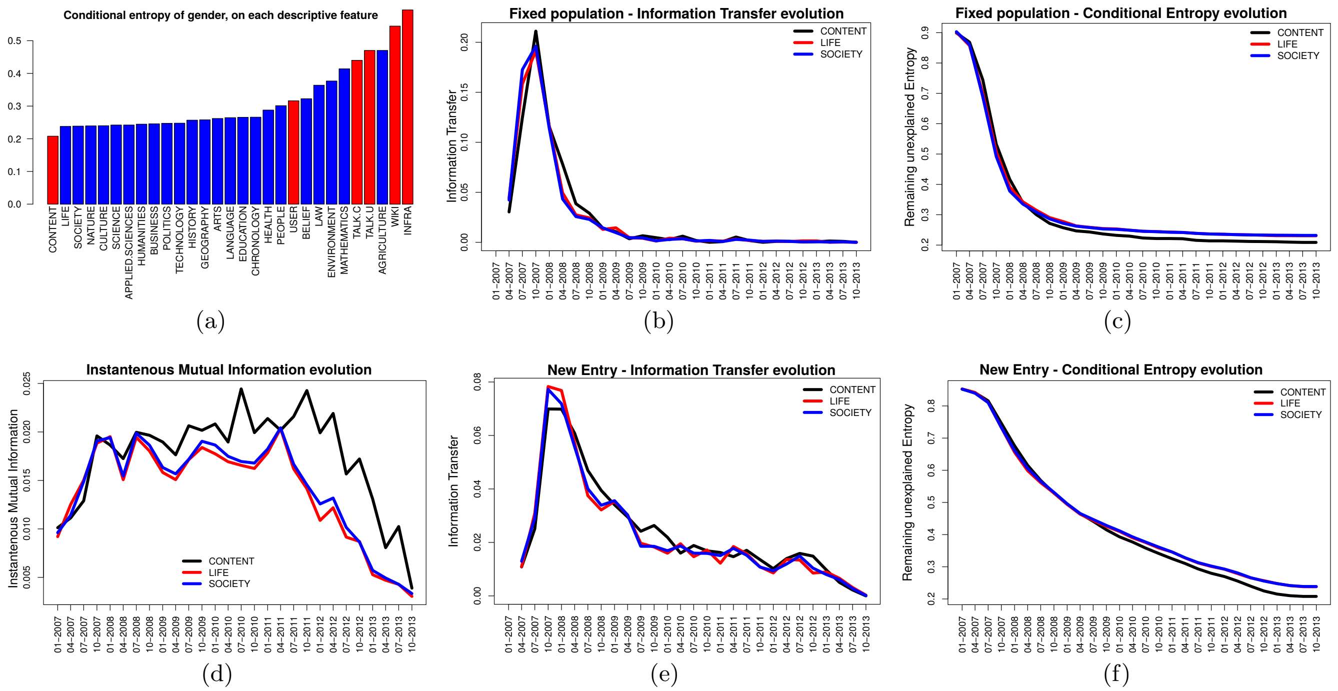


Figure 9: (a) Conditional entropy after conditioning on each feature (with New Entry NE). Feature are ordered by the conditional entropy $H(Y|X)$, here $Y = 0/1$ is the gender attribute, X is each feature. Lower is better. *Basic* features are shown in red, *extended* features in blue. (d) Mutual Information between each feature at time t and gender (on NE). *Information Transfer* on Fixed Population FP (b) and NE (e). Conditional entropy evolution on FP (c) and NE (f). We can see that while later edits contains just as much information about a user's privacy as the earlier edits, they contribute less to prediction gain, since most of the information they bring was already known.

vacy loss is caused by a joint effect of two factors: i) the information learned from new users who enter the population and ii) online breadcrumbs – a more accurate estimate of how the behavioral features correspond to personal traits. To separate these factors, we study two scenarios defined by subsets of the editor population: the *"New Entry"* (NE) in which new users can enter freely throughout time and *"Fixed population"* (FP), which is limited only to users active in the first timeframe (*i.e.*, first quarter of 2007). In Fig. 8, we plot the AUC over time when studied on FP and we compare it to NE. The curve corresponding to FP increases over time, though slower in later timeframes. The difference between the first and last timeframe is statistically very significant – t test $p < 0.01$. Intuitively, only the online breadcrumbs could cause this privacy decay for FP. By comparison, predictions on NE are constantly more accurate and continue to notably improve beyond the initial burst detected for FP. We attribute this additional improvement to information relating to new users entering the system.

Are later edits less harmful to one's privacy? Intuitively, the longer a user edits, the more she discloses about herself. We approach this issue using the Temporal Information Theory measure detailed in Sec. 4, on both FP and NE populations. Let T be the number of timeframes. Fig. 9a shows the conditional entropy of the class variable Y (here gender), after conditioning on all the temporal instantiations of a given feature X (*i.e.* $H(Y|X_{1:T})$), on the NE dataset. We can see which features which give out the most information about a user's gender. Consistent with the data profile in Sec. 5, CONTENT differentiates the most *males* from *females*. The thematic features, like Life, Society, Nature or Culture follow, having very similar scores. Surprisingly, USER distinguishes gender less than seen from the profiling analysis. An almost identical ordering of importance of features is obtained on FP. Fig. 9d presents the instantaneous mutual information over time between the gender variable and the three most important features. All three present very similar dynamics (both on NE and on FP). The information overlap between the temporal instantiation of features and the private trait remains almost constant until close to the end of the studied period. This answers the questions whether later edits are less harmful, by showing that later activity hurts privacy as much as the initial activity, as it discloses similar quantities of information. We further study the amount of *new* information introduced features in later timeframes. Fig. 9b and 9e show that the *Information Transfer* over time on respectively FP and NE, for the same three features. All series present an initial burst, after which they drop quickly. Note that, due to differences in the effectives of the FP and NE populations, the absolute values are not directly comparable and only their evolutions are meaningful. We can see that later features X_t bring few *new* information not already disclosed by the earlier features $X_{t-1}, X_{t-2} \ldots$. The take-home message for this subsection is: *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.*

The continuous impact of newcomers on privacy loss. Information Theory measures provide means for separately quantifying the privacy loss due "online breadcrumbs" and newcomers. For FP (Fig. 9b), the utility of later edits drops to virtually zero, whereas for NE (Fig. 9e) they decrease to a non-negligible score. The information inferred from newcomers seems to be moderate, but constant in time. This information is also responsible for the continuous increase of prediction performance detected in Fig. 8 for the NE population. Similar conclusions can be drawn by studying the conditional entropy over time for the two populations. For FP (Fig. 9c) it decreases rapidly and remains constant afterwards, showing that virtually no new information is learned after the initial burst. For NE (Fig. 9f), it continues to decay even after the initial burst, though at a slower pace.

6.2 Predicting the attributes of exited users

Privacy continues to erode even for retired users. We further analyze what happens to the privacy of users who left the system. After their retirement, no more user-originating information ("online breadcrumbs") is available to disclose private traits. We quantify the prediction performance on a user population who have edited prior to 01.01.2008, but stopped after this date. Therefore, any information introduced in the system by the users themselves is restricted to the timeframes before 2008. In Fig. 10a, we plot the AUC over time for the education/*undergrads* binary classifier. We observe an constant increase of prediction performance, even though it shows a saturation in later timeframes. An increase of prediction performance for for retired users is also observable for religion/*christian* (see the SI [1]), but not for any of the other binary classifiers. No longer being in activity, the loss of privacy after 01.2008 is not the result of the users' actions. A Temporal Information Theory analysis performed on this exited population shows *Information Transfer* values of zero after 01.2008 – i.e. no information originating with "online breadcrumbs". As far as we know, these are the first results to show that certain private traits could be predicted increasingly better even after the users exited the system.

Why does privacy degrade for exited users? Intuitively, user originating information is available only until the exit of the users, afterwards the source of new information are in the actions of other users. Predictions about unseen retired users can be made only using features relating to their period of activity (here 01/2007 - 12/2007). In the logistic regression models learned at each timeframe, the strength of the links between features and the class variable are given by the corresponding coefficients. We study the coefficients of features which encode the activity of users prior to their retirement (i.e. $X^u_{1:4}$). We show the coefficients relating to CONTENT (in Fig. 10b) and to USER (in Fig. 10c), in models learned for education/*undergrads* in each timeframe. In later timeframes, CONTENT features observe an increase in importance, with both $\texttt{CONTENT}_2$ (number of CONTENT revisions in the 2^{nd} quarter of 2007) and $\texttt{CONTENT}_3$ steadily increasing from being completely absent in the initial models. $\texttt{CONTENT}_4$ remains absent for all timeframes. Simultaneously, USER features decrease in importance, with $\texttt{USER}_4$ disappearing completely. We hypothesize that the AUC increase observed after 01.2008 originates with the currently active users *whose activity overlapped with the exited users*: classifier learns from users active both before and after 01.2008, by modifying the weights of features, including those before 01.2008. In this case, they learn that CONTENT features should have more importance, while USER features should have less. In summary, predicting of personal traits increase even for retired users, and the key factor for this improvement is better estimates on a subset of important features such as CONTENT.

7. DISCUSSION

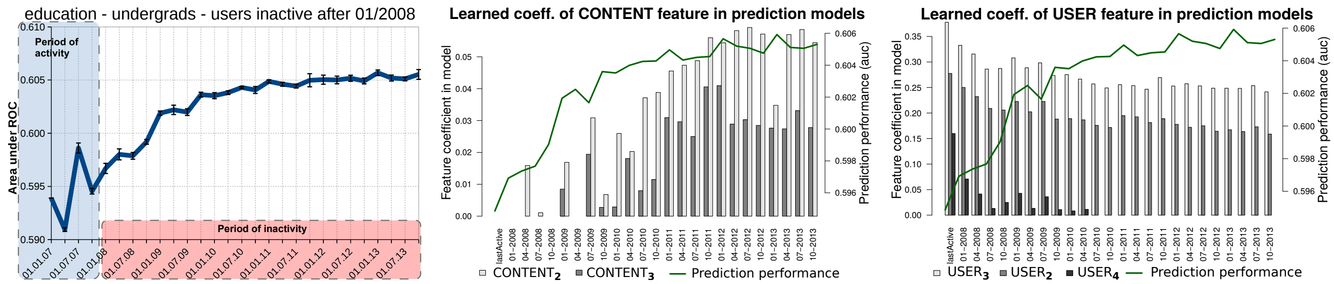
We present a first study to quantify the extent of gradual privacy erosion over six years. First, we set up a large scale evaluation using Wikipedia editing behavior to predict private traits over time. We analyze a 13 year history of Wikipedia edits made by more than 117 thousand users. Our descriptive analysis showed that as Wikipedia evolves, editors of different personal traits shows distinct patterns in their volume of edits and topical preferences. Second, we provide experimental evidence that time has an adverse effect on privacy. We show that prediction performance on private traits, such as gender, education and religion, increases with longer Wikipedia editing history. Third, we show that prediction of private traits improves even for users who have stopped editing Wikipedia. We further quantify the effect of predictability, and found that the improved performance can be attributed to two factors: new editors of Wikipedia, and better estimate of feature relevance - with the first having a larger effect. To the best of our knowledge, this is the first study to quantify the change of private traits over time, using Wikipedia, an open online dataset containing behavior breadcrumbs. This work shall raise awareness in the public on the privacy implications of online activity over long periods of time. The fact that the information of newcomers can be used to learn more about existing members has profound implications: users do not have complete control over the consequences of the information they release. Reflecting on this work, we would like to discuss a few of its limitations, practical implications and connections to other areas of research.

What does it really mean for Wikipedia users, should they be worried? To the best of our knowledge, no studies have shown that the real identity of Wikipedia users can be revealed based on their editing activity. However, this may not the case for other social networks.

User disclosure bias. This work uses self-disclosed personal traits on users' public profile. It is well-known that such a data source is prone to users' disclosure biase. The set of users who voluntarily disclose private information might be biased towards users less concerned with their privacy, who in turn has a distinct behavioral pattern. Validating the effect of such a bias would require an alternative source of groundtruth and maybe even behavioral data, and is beyond the scope of this study.

What about other online platforms? Although we predict predict private traits from Wikipedia, similar prediction results (on a static snapshop) was reported for other platforms such as Facebook [12]. Being a collaborative encyclopedia, Wikipedia records relatively small amount of information about its users. More detailed longitudinal analyses could be performed on platforms such as Facebook, and it is likely to also see an increasing trend for predicting personal traits.

A natural law of evolution of privacy loss. This study provides empirical analysis about privacy loss. A longer-term open challenge is a physical model for privacy loss, i.e., predict the de-anonymization rate of a given anonymized dataset.



(a) (b) (c)

Figure 10: (a) Increase of prediction performance for for users retired after 01.2008 – education/*undergrads* (other in SI [1]). Coefficients one model per timeframe: (b) `CONTENT` coefficient (absent in the model corresponding to the dataset where users were last active) increase in importance. (c) `USER` relating features decrease in importance.

Effective conditions for preserving privacy? There is a growing literature on characterizing which privacy guarantees can be obtained under a given information release protocol [25]. We hope our findings invite new empirical and theoretical investigation into the case in which data release is spatio-temporal and heterogeneous across different entities. In light of this study, we advocate that new means should be found to tackle the issue of online privacy. We argue one feasible means to preserving privacy is to construct laws which would enable erasing the recorded activity in the online environment.

Acknowledgments NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. This material is based on research sponsored by the Air Force Research Laboratory, under agreement number FA2386-15-1-4018.

8. REFERENCES

[1] Supplementary material: Evolution of privacy loss in Wikipedia, 2015. http://goo.gl/JT6WK7.

[2] A. Acquisti, L. K. John, and G. Loewenstein. What is privacy worth? *The Journal of Legal Studies*, 42(2):249–274, June 2013.

[3] R. Almeida, B. Mozafari, and J. Cho. On the evolution of Wikipedia. In *ICWSM '07*, 2007.

[4] D. Barth-Jones, K. E. Emam, J. Bambauer, a. Cavoukian, and B. Malin. Assessing data intrusion threats. *Science*, 348(6231):194–195, Apr. 2015.

[5] d. boyd and A. E. Marwick. Social Privacy in Networked Publics: Teens' Attitudes, Practices, and Strategies. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, sep 2011.

[6] C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *WWW*, page 699, 2012.

[7] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3, 2013.

[8] A. Gibbons, D. Vetrano, and S. Biancani. Wikipedia: Nowhere to grow. Tech. report, Standford, 2012.

[9] A. Halfaker, R. S. Geiger, J. T. Morgan, and J. Riedl. The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline. *American Behavioral Scientist*, 57(5):664–688, Dec. 2012.

[10] C. J. Hoofnagle, J. King, S. Li, and J. Turow. How Different are Young Adults from Older Adults When it Comes to Information Privacy Attitudes and Policies? Ssrn scholarly paper, Apr. 2010.

[11] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*. Springer New York, 2013.

[12] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 110(15):5802–5805, 2013.

[13] D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[14] A.-M. Meyer and D. Gotz. A new privacy debate. *Science*, 348(6231):194–194, Apr. 2015.

[15] P. E. Meyer. R package 'infotheo', 2014.

[16] Y.-a. D. Montjoye and a. S. Pentland. Assessing data intrusion threats–Response. *Science*, 348(6231):195–195, Apr. 2015.

[17] Y.-a. D. Montjoye, L. Radaelli, V. K. Singh, and a. S. Pentland. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539, Jan. 2015.

[18] A. Narayanan, E. Shi, and B. I. P. Rubinstein. Link prediction by de-anonymization: How we won the kaggle social network challenge. In *IJCNN*, pages 1825–1834, 2011.

[19] A. Narayanan and V. Shmatikov. Myths and fallacies of personally identifiable information. *Comm. of the ACM*, 53(6):24–26, June 2010.

[20] A. Ramachandran and A. Chaintreau. The Network Effect of Privacy Choices. In *Workshop EcoNet*, pages 1–4, 2015.

[21] J. Saramaki, E. A. Leicht, E. Lopez, S. G. B. Roberts, F. Reed-Tsochas, and R. I. M. Dunbar. The persistence of social signatures in human communication. *PNAS*, 2014.

[22] B. Suh, G. Convertino, E. H. Chi, and P. Pirolli. The Singularity is Not Near: Slowing Growth of Wikipedia. In *WikiSym '09*, pages 8:1–8:10. ACM, Oct. 2009.

[23] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[24] G. Ver Steeg and A. Galstyan. Information transfer in social media. In *WWW '12*, page 509, 2012.

[25] L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Jour. of American Stat. Assoc.*, 105(489):375–389, 2010.

[26] H. T. Welser, D. Cosley, G. Kossinets, A. Lin, F. Dokshin, G. Gay, and M. Smith. Finding social roles in wikipedia. In *iConference*, pages 122–129. ACM, 2011.

[27] W. Youyou, M. Kosinski, and D. Stillwell. Computer-based personality judgments are more accurate than those made by humans. *PNAS*, 2014.

Supplementary Material: Evolution of privacy loss in Wikipedia

Main article DOI: 10.1145/2835776.2835798

Marian-Andrei Rizoiu [1,*], Lexing Xie [2], Tiberio Caetano[3] and Manuel Cebrian [4]

[1]*NICTA & Australian National University, Canberra, Australia*
[2]*Australian National University & NICTA, Canberra, Australia*
[3]*Ambiata, ANU and UNSW, Sydney, Australia*
[4]*NICTA & University of Melbourne, Melbourne, Australia*

Correspondence*:
Marian-Andrei Rizoiu
NICTA Research Lab, 7 London Circuit, Canberra, Australia,
Marian-Andrei.Rizoiu@nicta.com.au

We provide in this document detailed information about:

- the construction of the Wikipedia dataset used in the Main Text. Additionally, we provide links for downloading the user dataset (useful for reproducing the prediction results) as well as the complete edit dataset;
- additional data profiling figures, completing the narrative in the Main Text and provided for completeness;
- additional prediction performance measuring.

CONTENTS

1 THE CONSTRUCTION OF THE WIKIPEDIA DATASET

Wikipedia uses internally a revision system, which records every edit or modification made to a page by a user. Such atomic operations are called "revisions" in Wikipedia's vocabulary. The version of a page at any given moment in time can be obtained by overimposing all the revisions made to the page from the beginning of time until the given moment. The entire history of revisions is publicly available for download [1]. The user descriptive features introduced in the main article were constructed starting from the July 2013 English Wikipedia stub dump. This dump contains all the history of Wikipedia,

[1] Download Wikipedia dumps: `https://dumps.wikimedia.org/enwiki/latest/`

starting from its creation in January 2001 until July 2013. We record for each individual revision: its editor, target page and timestamp. For purposes of internal organization, Wikipedia page are assigned into namespaces, which are categories based the intended purposes of pages: main articles, talks around article, user pages, user talks, community pages, project pages *etc.* The basic descriptive features for users are constructed based on the Wikipedia namespaces of edited pages, as shown in the main article.

1.1 THE WIKIPEDIA CATEGORIZATION SYSTEM AND THE EXTENDED FEATURES.

We construct the extended descriptive features, based on the Wikipedia categories[2], which are a categorization system, based on the theme of the articles. This categorization system defines main categories, such as Geography, History, Arts, which can be further divided into finer subcategories. This forms a shallow hierarchy, *i.e.*, a hierarchy with a low number of levels. Loops are allowed inside the hierarchy – a given category can be a sub-category of multiple parent categories –, though discouraged, and the resulting category graph does not posses a strict tree structure. We select the 23 main categories of Wikipedia to serve as thematic features in a user's editing description. Each Wikipedia article can be placed under one or multiple (sub-)categories. Every time a user edits a page, we propagate the resulted revision through the category graph by following the category parent relation. The user's revision counts for all reachable main topics are incremented. We avoid the infinite propagation through the loops using a propagation threshold, equal to the average "height" of the hierarchy. While not a proper hierarchy, we define the "up" propagation direction as towards the main categories.

1.2 CONSTRUCTING THE EDITOR'S PRIVATE DATA: THE PRIVATE TRAITS.

User pages are similar to regular Wikipedia pages, with the difference that they are dedicated to users. Each user has, by default, a user page in Wikipedia, which serves the role of a "social profile" as in an online social network. The associated talk pages are employed for private discussions. The user pages and the user talk pages form Wikipedia's social aspect, which has been shown to be a prolific environment for activities such as campaign for internal elections (**Danescu-Niculescu-Mizil et al.**, 2012). All information that we use as private information is provided by some of the editors themselves, on their personal user pages. By adding labels to their user pages, editors give information about their geographic location, nationality, religion, profession, education level, philosophy or even sexual preferences. Similar to the aforementioned page categories, user categories are re-constructed bottom-up, based on the label information scrapped from the user pages using the Wikipedia API. As an example about the type of information that we can obtain about editors, we show the visualization of the first level child nodes of the categories providing information about the editor's location (Figure 2a), profession (Figure 2b) and religion (Figure 3). Some of the categories are further divided into subcategories, which provide increasingly fine-grained information. The three categories selected as private traits in the main article (*i.e.*, gender, religion and education) were chosen for being closer to what humans perceive as private information, as opposed to spoken languages or geographic location. We selected only a subset of subcategories out of all available subcategories (*e.g.*, *christian*, *muslim*, *atheist* and *jewish* for religion) in order to have a reasonable number of editors to perform the analysis on.

We restrict our dataset to only registered users for which at least one private information is retrieved from their user page. Summarizing, the dataset includes for each user:

- a description of the user's editing activity, recorded down to revisions on individual pages, together with their timestamp;

[2] Wikipedia categories descriptions: `https://en.wikipedia.org/wiki/Portal:Contents/Categories`

- some private information, retrieved from their own user pages, under the form of user categories.

The dataset contains:

- 188,805,088 revisions
- 117,523 users
- 8,679 user categories (and their hierarchical relations)
- 22,172,813 edited pages
- 430,410 page categories (and their hierarchical relations)
- extent of time: beginning of Wikipedia (January 2001) until July 2013.

1.3 DOWNLOADING THE WIKIPEDIA DATASET

We make available for the public to download the Wikipedia dataset used in this work. We provide two versions of the dataset: the *temporal user editing behavior dataset* and *complete editing dataset*.

Temporal user editing behavior dataset
Download (~82 MB): `http://goo.gl/Tx5SoI`
Description: Dataset containing user activity over time and the three private traits used in this work: gender, education and religion.
Usage: This dataset is provided so that any third party can reproduce the experiments described in this work.
Technical description: This dataset captures the editing activity of Wikipedia users from 01.2002 until 07.2013. It is comprised of 47 CSV ("Comma Separated Values" format) files, one for each quarterly timeframe. Each file contains 117,523 lines (plus a header line), describing the editing activity of each of the users in our dataset, from the beginning of Wikipedia until the beginning of the timeframe corresponding to the given file. The columns in the CSV file (starting from the 6^{th} column onwards) correspond to the *basic* and *extended* features, described in the Main Text, Sec. 4.1. For example, column `AGRICULTURE` in file "2003_07_dataset.csv" gives, for each editor, the total number of revisions relating to `AGRICULTURE`, performed from the beginning of Wikipedia until June 31^{st} 2003. Columns `ALL` gives the total number of revisions performed by an editor. Columns `gender`, `education` and `religion` correspond to the private traits of the editors. The content of these columns does not vary between the different temporal datasets. Fields in these columns take values when information is known about the users, otherwise they take the value `NA` (not available).

Complete editing dataset
Download:
  1% Sample (~495 MB): `http://goo.gl/T47UVj`
  Complete (~3.6 GB): `http://goo.gl/2iLH7A`
Description: Dataset containing information about Wikipedia pages, categories, user, user categories and revisions, under a relational database format.
Usage: This dataset is provided to allow extensions to current work and/or to learn new insights for this data.
Technical description: The *complete editing dataset* is provided as a SQL relational database, with the table schema presented in Fig. 1. It contains three main tables:

- `user` – gives the names and total number of edits for each used in the dataset;
- `page` – provides information about the Wikipedia pages in the dataset (title, namespace, creation date *etc.*);

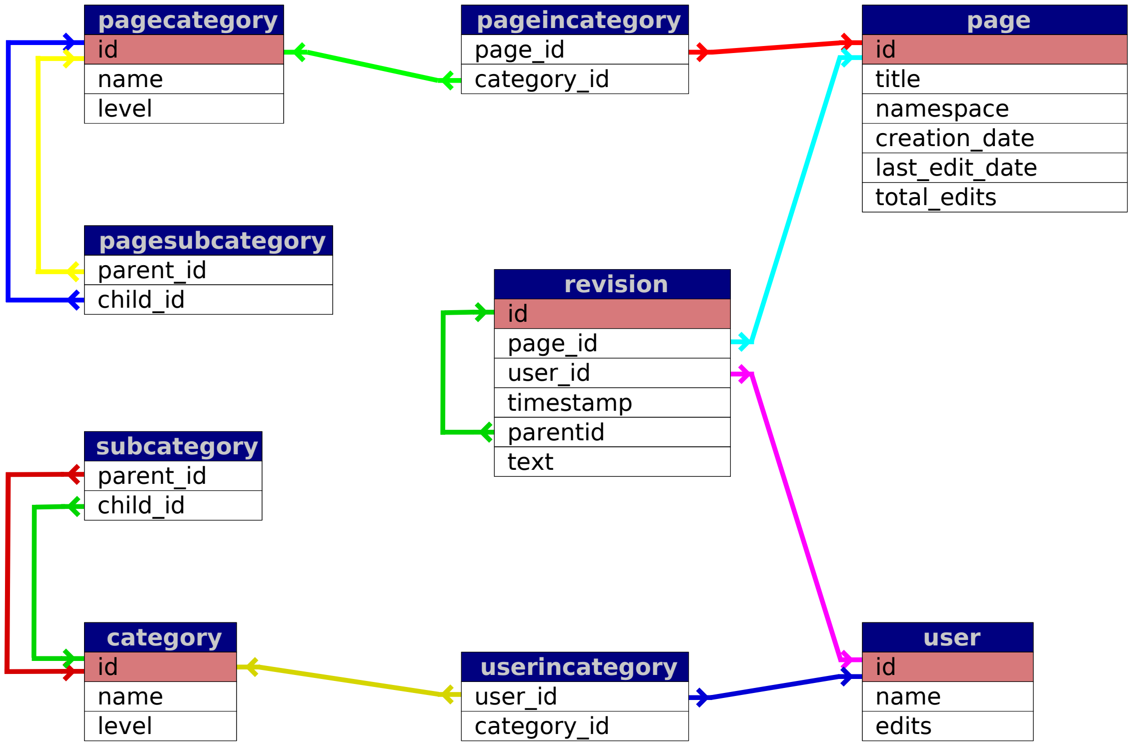


Figure 1. SQL table schema, showing the relations between the tables in the "complete editing dataset". Table names are shown on blue background, primary keys are shown on red background. Arrows indicate foreign key relations, with the direction from a given field to its corresponding foreign key.

- `revision` – records atomic edits (revisions) made on page *page_id* by user *user_id* at time *timestamp*. The field *text* is always empty since we work with the structure of Wikipedia only and not with the text. This field exists here to allow future enhancements. With more than 188 million revisions, this table is the bulk of the dataset. For easier handling and speed of prototyping, we provide the "1% Sample" dataset, which contains only 1% of the revisions in the "Complete" dataset (the content of all the other tables is identical).

The remaining 6 tables describe the page/user categories and their relations. Table `pagecategory` gives information about page categories, which are constructed based on the category system described in Sec. 1.1. Similarly, table `category` provides information about user categories (described in Sec. 1.2), which serve as private traits. Table `pageincategory` (`userincategory`) gives the n-to-m relations between pages (users) and their respective categories. Table `pagesubcategory` (`subcategory`) encodes the hierarchic relation between categories.

2 ADDITIONAL DATA PROFILING GRAPHS

As indicated in the main article, we present hereafter additional graphs showing:

- Fig. 4 and 5: The decline of editorship and the rise of maintenance, number of total revisions, new and Thematic features all present very similar evolutions to those presented here and in the main article.
- Evolution of editing patterns for gender (Fig. 6 and 7), education (Fig. 8 and 9) and religion (Fig. 10 and 11). Pairs (feature, private trait) for all the *basic* and *extended* features. The number of revisions performed by editors during each each quarterly timeframe on each feature are computed as percentages of the total number of revisions (`ALL`), and the mean value over all users is presented.

3 STATISTICAL TESTING OF INCREASING PRIVACY LOSS

In the Main Text Sec. 6.1, we show that Privacy Loss follows an increasing trend. The prediction performances, measured using the AUC ROC indicator **James et al.** (2013), obtained for later timeframes are higher compared to those of earlier timeframes. This evolution is shown in Main Text Fig. 6 and completed in SI Fig. 13. In order to rigorously test this increasing trend, we perform statistical testing analysis. For each timeframe, we train the Logistic Regression 20 times (cf. testing protocol described in Main Text Sec 4.2). We perform a *2-sample two-sided t test* between the AUC ROC results obtained for the first and the last timeframe, for each binary predictor. The null hypothesis states that the measure for both timeframes has the same mean value (i.e. no Privacy Loss), whereas the alternative hypothesis states that the mean values for the two timeframes are different. Table 1 shows the p-values of the performed t tests, as well as the significance level: * significant ($p < 0.05$), ** very significant ($p < 0.01$) and *** highly significant ($p < 0.001$). With the exception of religion/*jewish*, the increase of prediction performance for all binary predictors is statistically significant. Furthermore, we study how significant is the increase of prediction performance of the Fixed Population (`FP`) described in Main Text Sec. 6.1. The evolution of the AUC ROC measure, depicted in Main Text Fig. 8, is increasing in early timeframes and plateaus afterwards. By performing statistic testing, we show in Table 1 that even the loss of privacy due to the users' own behavior is statistically highly significant — line "gender `FP`".

Table 1. Hypothesis testing of the statistical significance of the increase of learning performance. For each trained predictor, we show the p-value and the significance level.

	class	p-value	Significance level
	gender	3.219×10^{-14}	***
	gender `FP`	1.262×10^{-4}	***
educat.	*undergrads*	5.882×10^{-13}	***
	grads	2.016×10^{-6}	***
	PhD	1.496×10^{-2}	*
religion	*atheist*	1.018×10^{-3}	**
	christian	6.916×10^{-12}	***
	jewish	1.064×10^{-1}	
	muslim	4.055×10^{-12}	***

4 COMPUTING OTHER PERFORMANCE MEASURES: FSCORE AND "EQUAL PRECISION AND RECALL"

In statistics, the receiver operating characteristic (ROC curve) is a graphical plot that illustrates the performance of a binary classifier system, while varying its discrimination threshold. When using normalized units, the area under the curve (the AUC measure) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative') (**Fawcett**, 2006). One of the advantages of the AUC measure, used in the Main Text to evaluate learning performances over each private trait, is that it does not require setting a decision threshold. Other measures, such as Accuracy, Precision, Recall or F-Measure, require establishing a decision boundary in order to decide if a given example belongs (or not) to a given class.

We devise the following protocol: for the predictions of a given model, we perform a line search for the decision threshold. For F-Measure, we search for the point which maximizes the measure. For the *point of equal Precision and Recall* (ePR), we choose the boundary for which Precision equals Recall. This measure is also known as the Precision-recall breakeven point. Once a boundary has been chosen, the given prediction is scored using the value of the measure at that point (i.e. the F-Measure or Precision/Recall). According to the testing protocol described in the Main Text Sec 4.2, each classifier is trained 20 times. We calculate and report the mean value and the standard deviation.

Table 2 presents these statistics, computed for the latest available snapshot (July 2013), alongside with each class prior (for gender, we present the prior of *male*). Three traits seem to be inferred particularly accurate: gender, religion/*jewish* and religion/*muslim*. The evolution over time of the equal Precision/Recall of each attribute is presented in Fig. 12. Two attributes present a noteworthy evolution (clearly outside the standard deviation denoted by error bars): gender can be predicted increasingly accurate overtime, while religion/*muslim* decreases slightly, probably due to the increase of active self-declared muslim editors.

Table 2. Performance of prediction of each private traits in the last timeframe (10/2013). Mean and standard deviation (in *italic fontface*) of F-Measure and "point of equal Precision and Recall" are calculated over 20 random stratified splits (each class prior is given in the "Class prior" column).

	class	Class prior	F-Measure		ePR	
	gender (*m*)	0.719	0.840	±*0.001*	0.791	±*0.006*
educat.	*undergrads*	0.420	0.605	±*0.007*	0.535	±*0.012*
	grads	0.493	0.701	±*0.001*	0.607	±*0.008*
	PhD	0.086	0.239	±*0.018*	0.175	±*0.030*
religion	*atheist*	0.324	0.796	±*0.001*	0.693	±*0.008*
	christian	0.451	0.704	±*0.000*	0.555	±*0.008*
	jewish	0.107	0.952	±*0.000*	0.913	±*0.002*
	muslim	0.116	0.937	±*0.000*	0.900	±*0.002*

5 ADDITIONAL PREDICTION PERFORMANCE GRAPHS

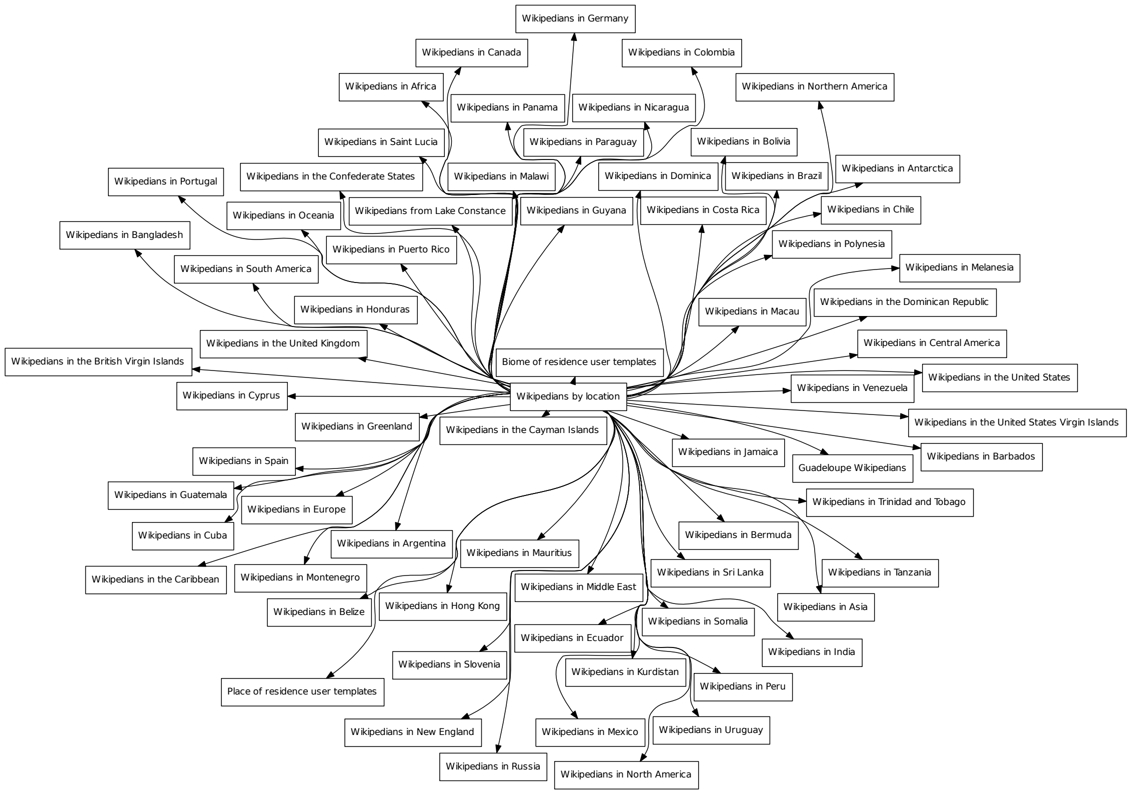
- Figure 13: Editing Wikipedia discloses private information. AUC evolution on the remaining classes of private traits, apart from those presented in the Main Text.
- Figure 14: In addition to education/*undergrads* information revealed about editors retired after 01.2008 (shown in Main Text Fig. 10a), we present also the increase of prediction performance for religion/*christian*. No other binary classifier presents an increase of AUC score over time.

REFERENCES

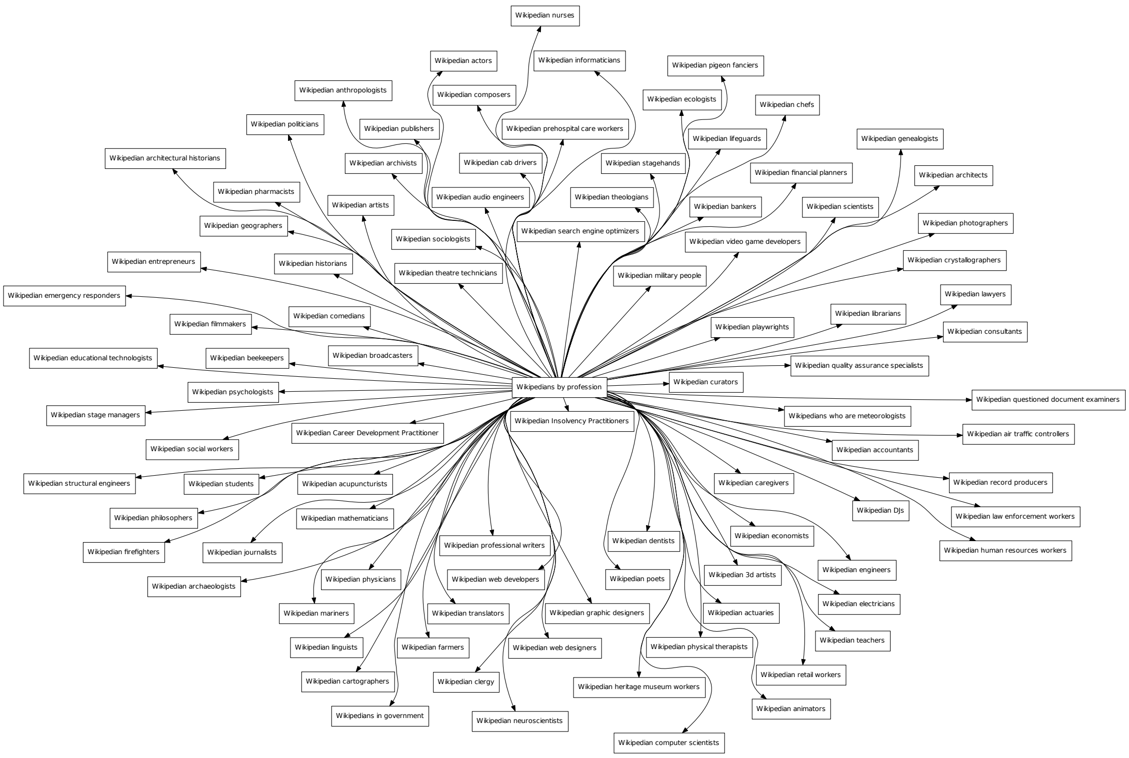
Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., and Kleinberg, J. (2012), Echoes of power: Language effects and power differences in social interaction, in World Wide Web, Proceedings of, WWW '12, 699–708

Fawcett, T. (2006), An introduction to roc analysis, *Pattern recognition letters*, 27, 8, 861–874

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), An Introduction to Statistical Learning, volume 103 of *Springer Texts in Statistics* (Springer New York), doi:10.1007/978-1-4614-7138-7



(a)



(b)

Figure 2. Visual representation of first level user categories that can be extracted from the Wikipedia user pages, relating to geographic location (a) and profession (b). Some categories are further subdivided into subcategories (not shown in this figure)

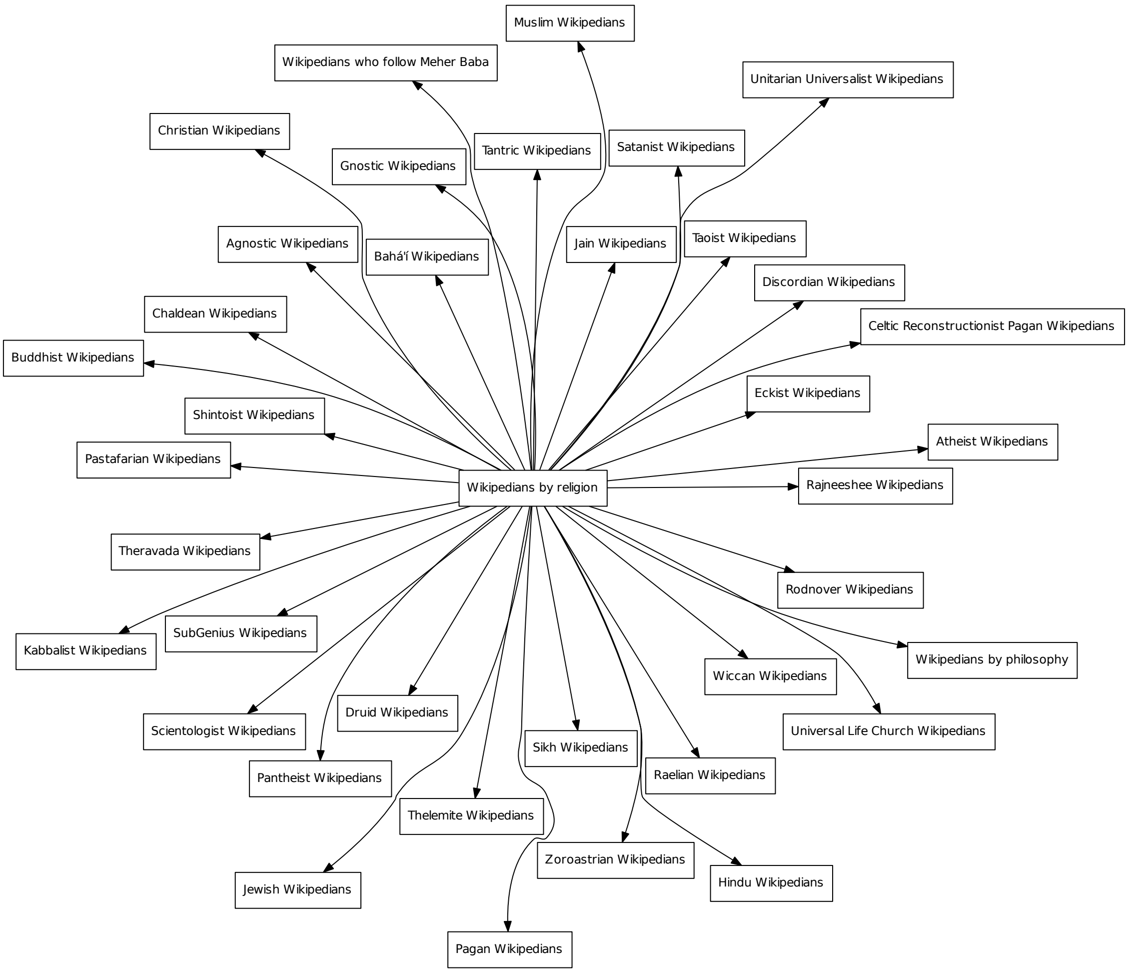


Figure 3. (cont. of Figure 2) Visual representation of user categories that can be extracted from the Wikipedia user pages, relating to editors' religion.

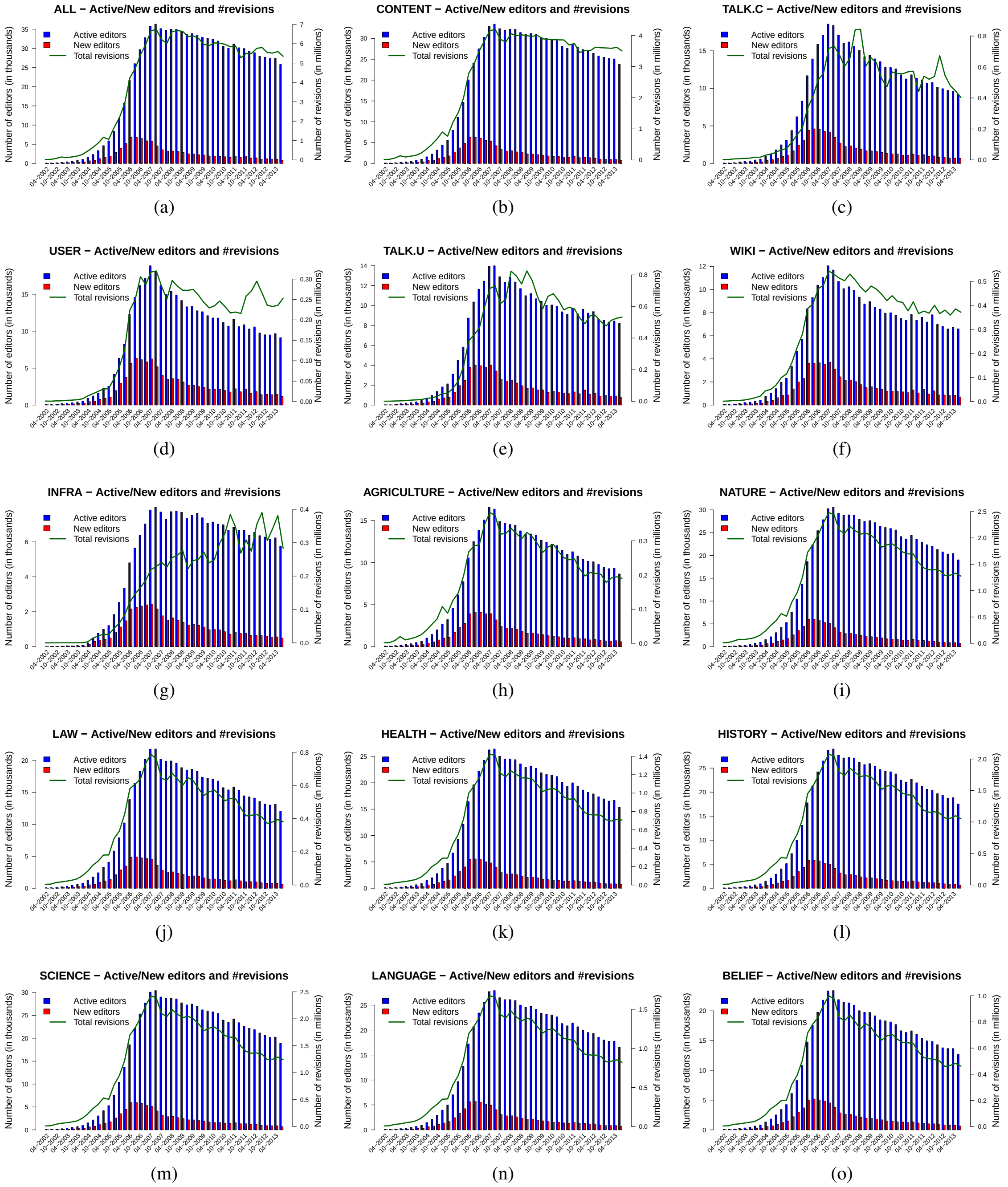


Figure 4. The number of revisions added to Wikipedia is constantly descending, as well as the number of active editors and new editors. Showing the evolution, broken down by feature.

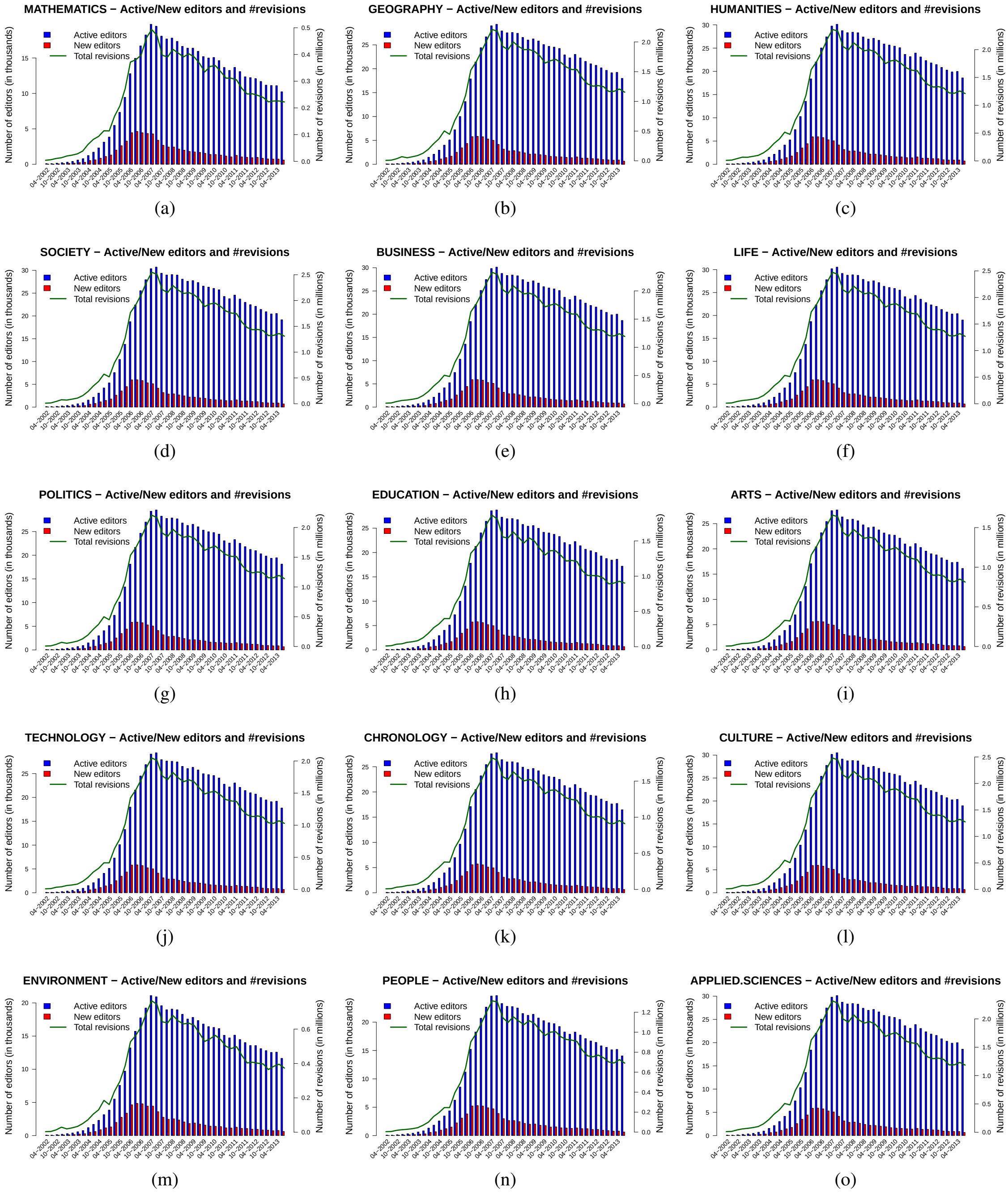


Figure 5. (cont. of Fig. 4) The number of revisions added to Wikipedia is constantly descending, as well as the number of active editors and new editors. Showing the evolution, broken down by feature.

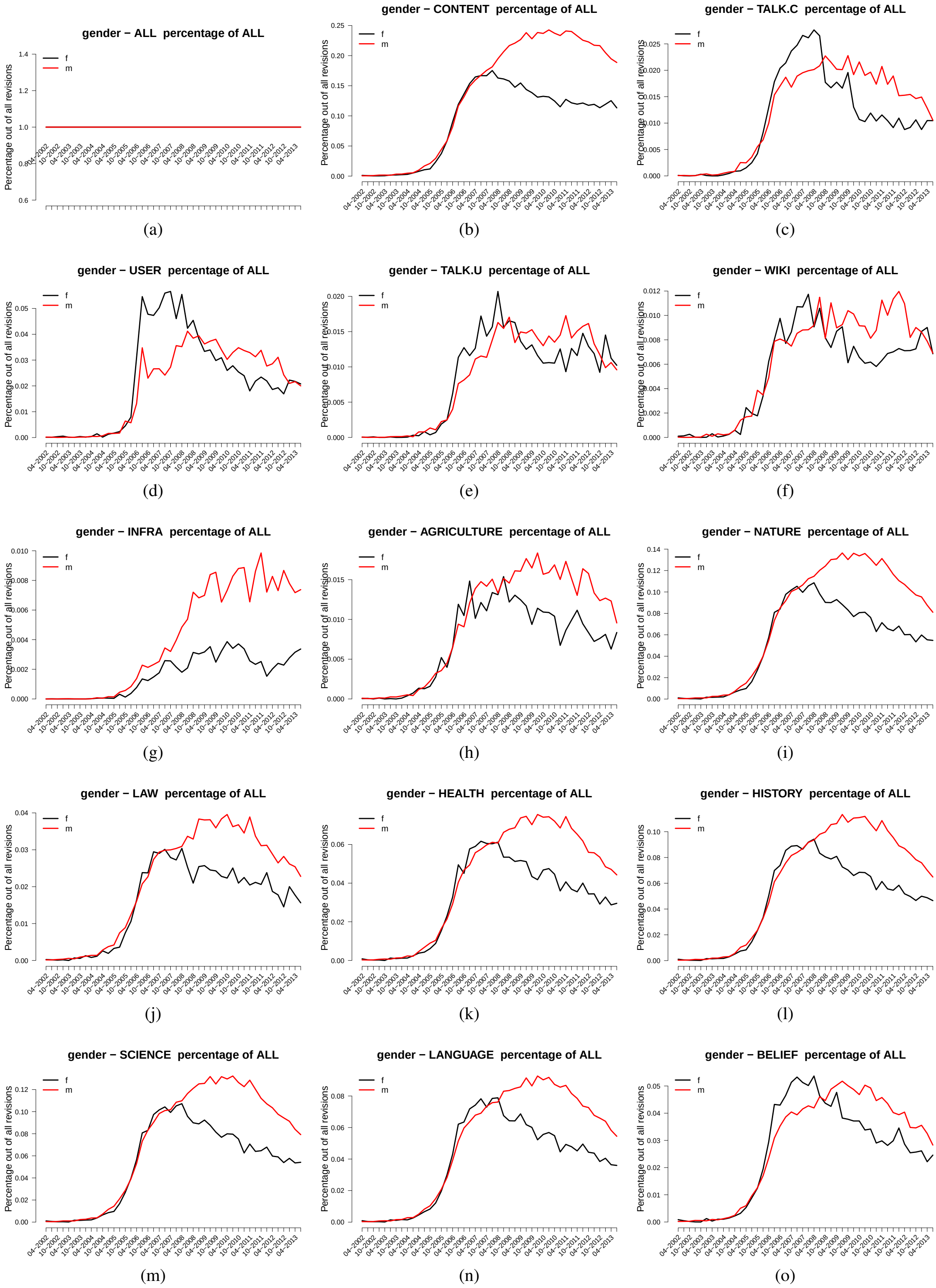


Figure 6. Temporal evolution of mean *basic* and *extended* values of features, broken down by gender. All features are expressed as percentages of ALL.

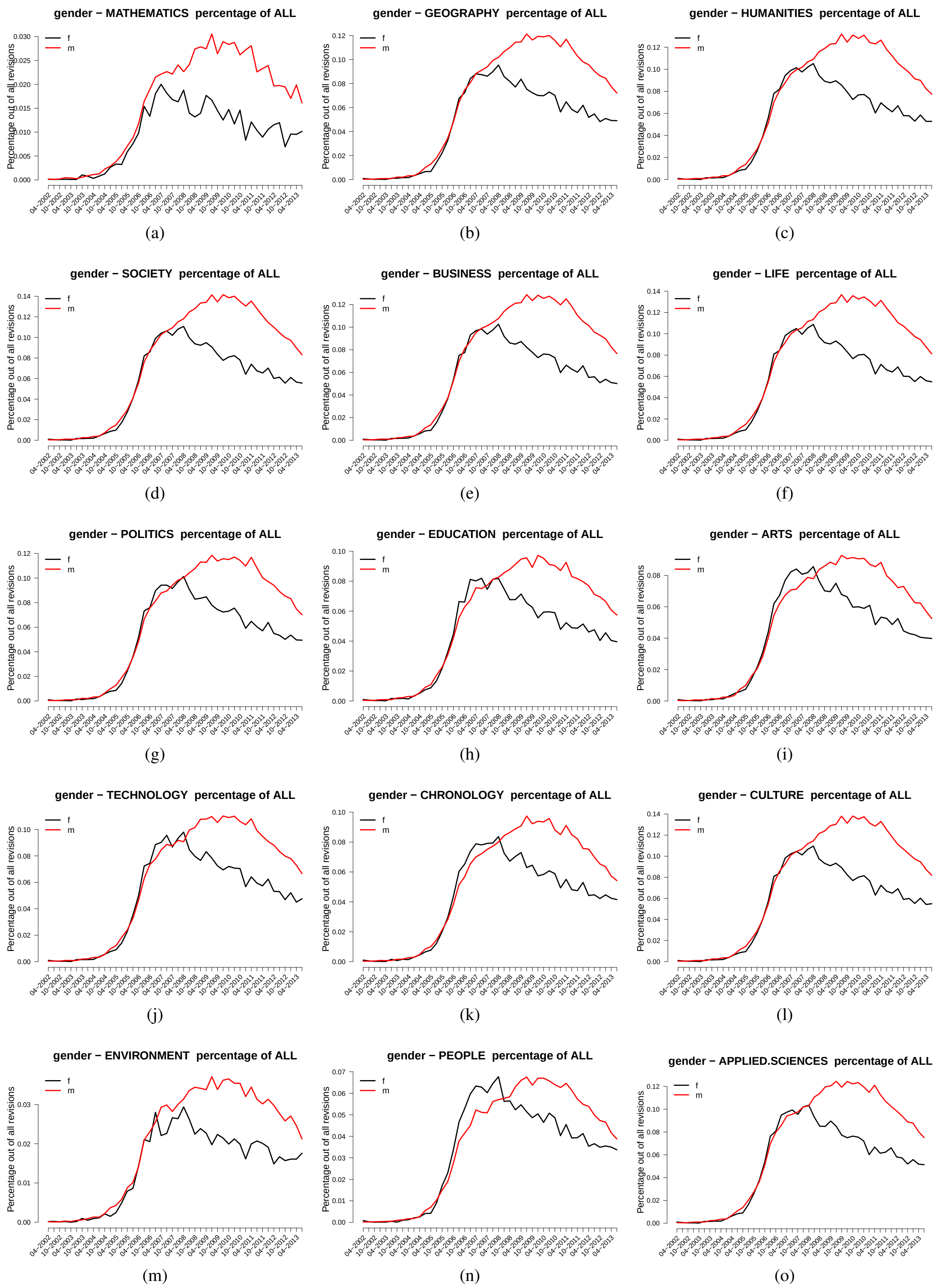


Figure 7. (cont. Fig 6) Temporal evolution of mean *basic* and *extended* values of features, broken down by gender. All features are expressed as percentages of `ALL`.

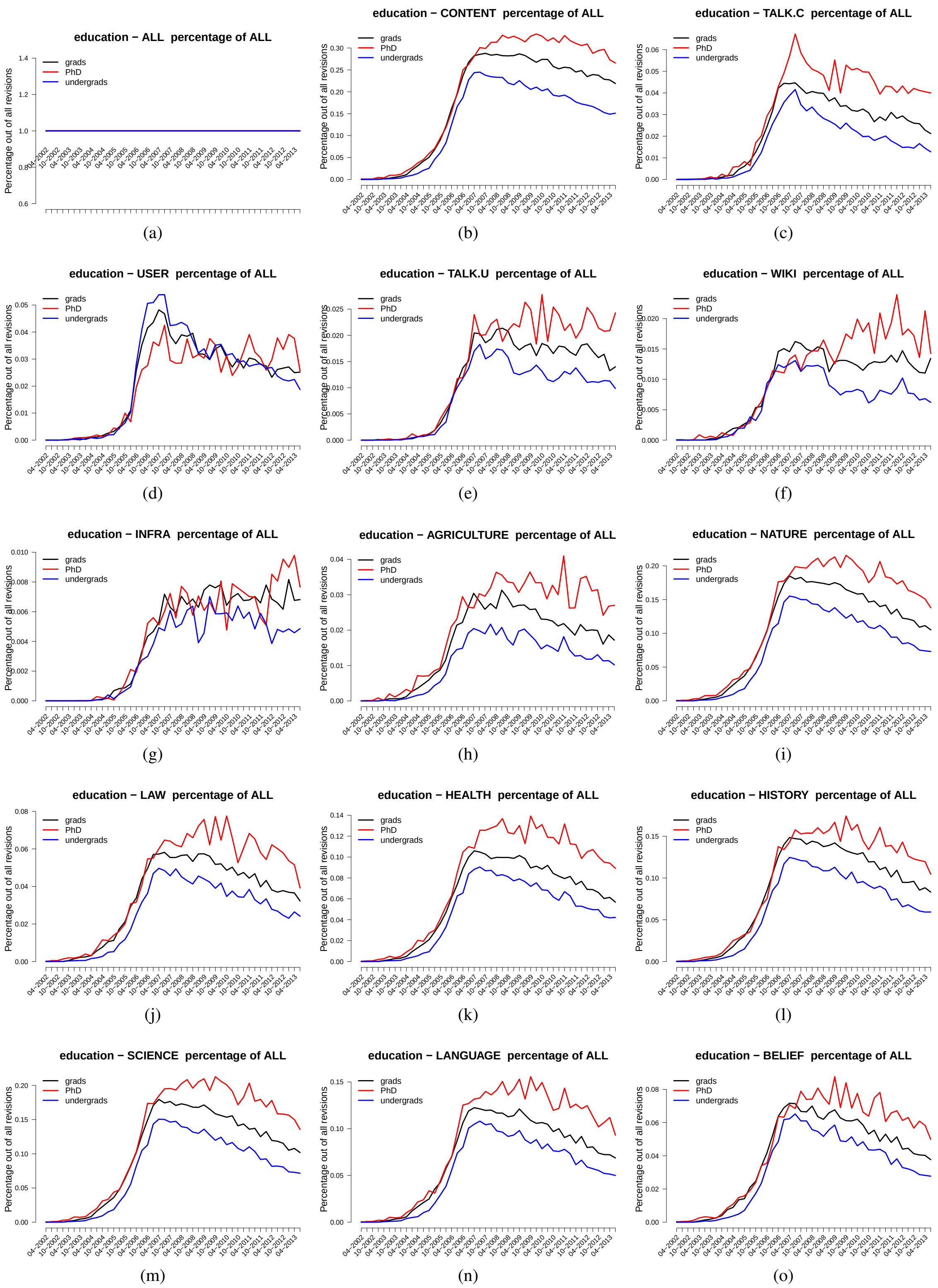


Figure 8. Temporal evolution of mean *basic* and *extended* values of features, broken down by **education**. All features are expressed as percentages of `ALL`.

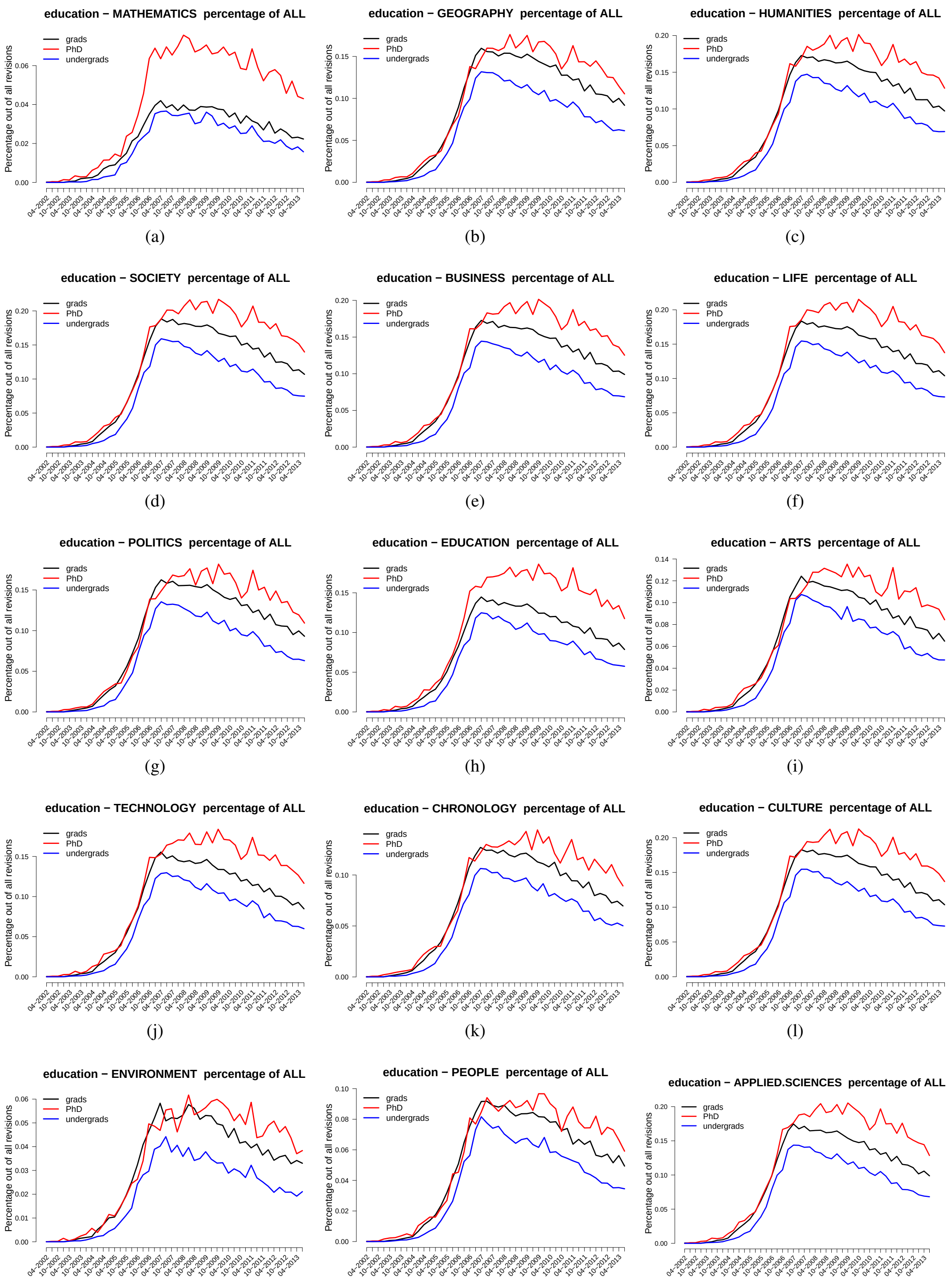


Figure 9. (cont. Fig 8) Temporal evolution of mean *basic* and *extended* values of features, broken down by **education**. All features are expressed as percentages of `ALL`.

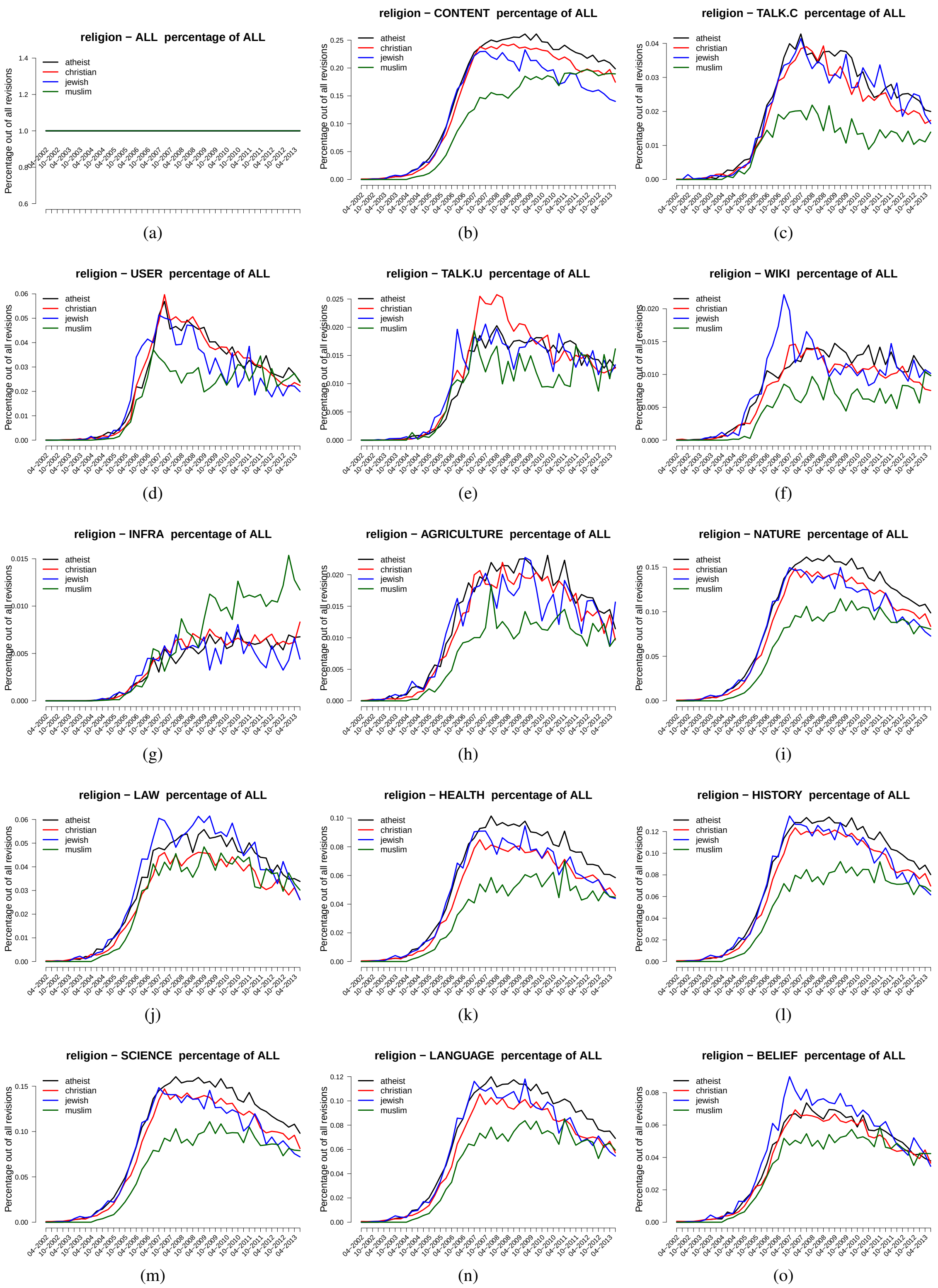


Figure 10. Temporal evolution of mean *basic* and *extended* values of features, broken down by **religion**. All features are expressed as percentages of `ALL`.

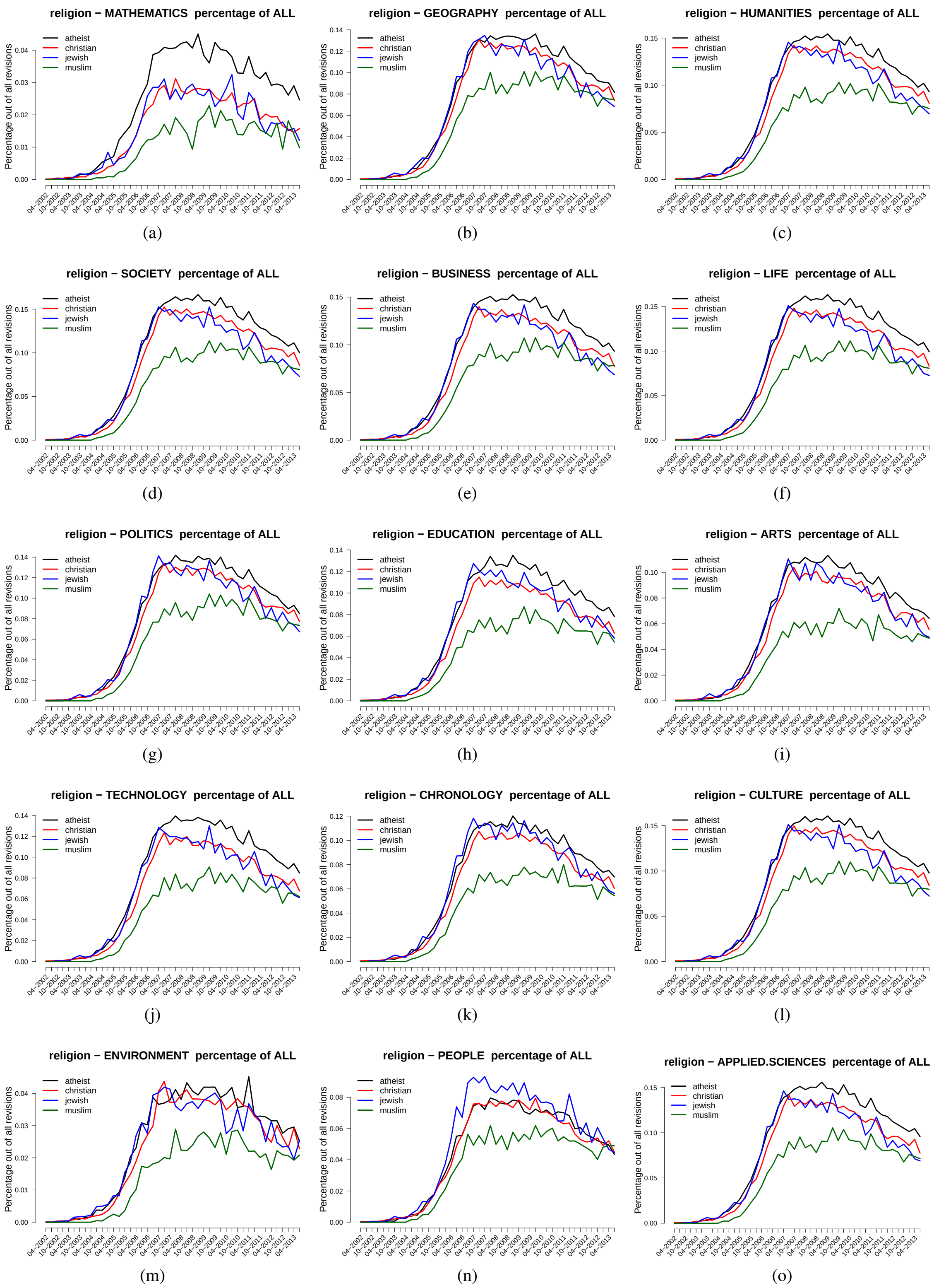
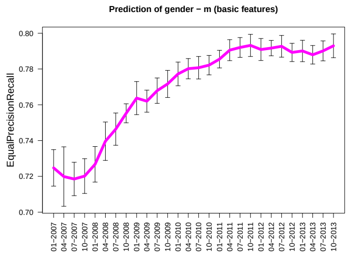
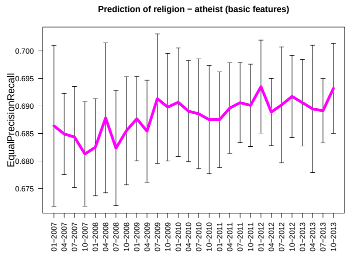


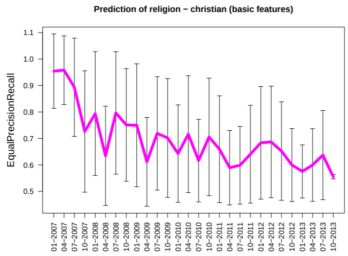
Figure 11. (cont. Fig 10) Temporal evolution of mean *basic* and *extended* values of features, broken down by **religion**. All features are expressed as percentages of ALL.



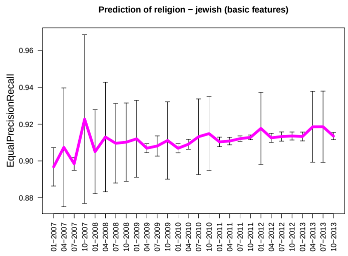
(a)



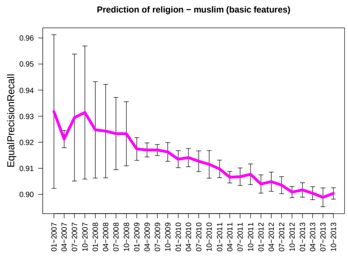
(b)



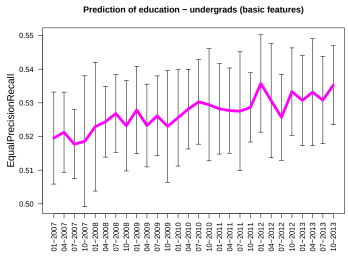
(c)



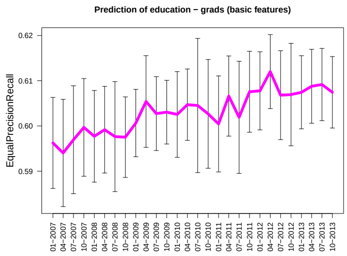
(d)



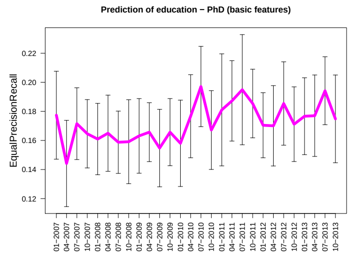
(e)



(f)



(g)



(h)

Figure 12. Evolution of prediction over time, measure using the *Equal Precision Recall point*. The lines represents the mean value and the error bars represent the standard deviation over 20 random split. In order from (a) to (h), the graphics present respectively gender, religion/*atheist*, religion/*christian*, religion/*jewish*, religion/*muslim*, education/*undergrads*, education/*grads* and education/*PhD*. The values for the rightmost timeframe are given in Table 2.

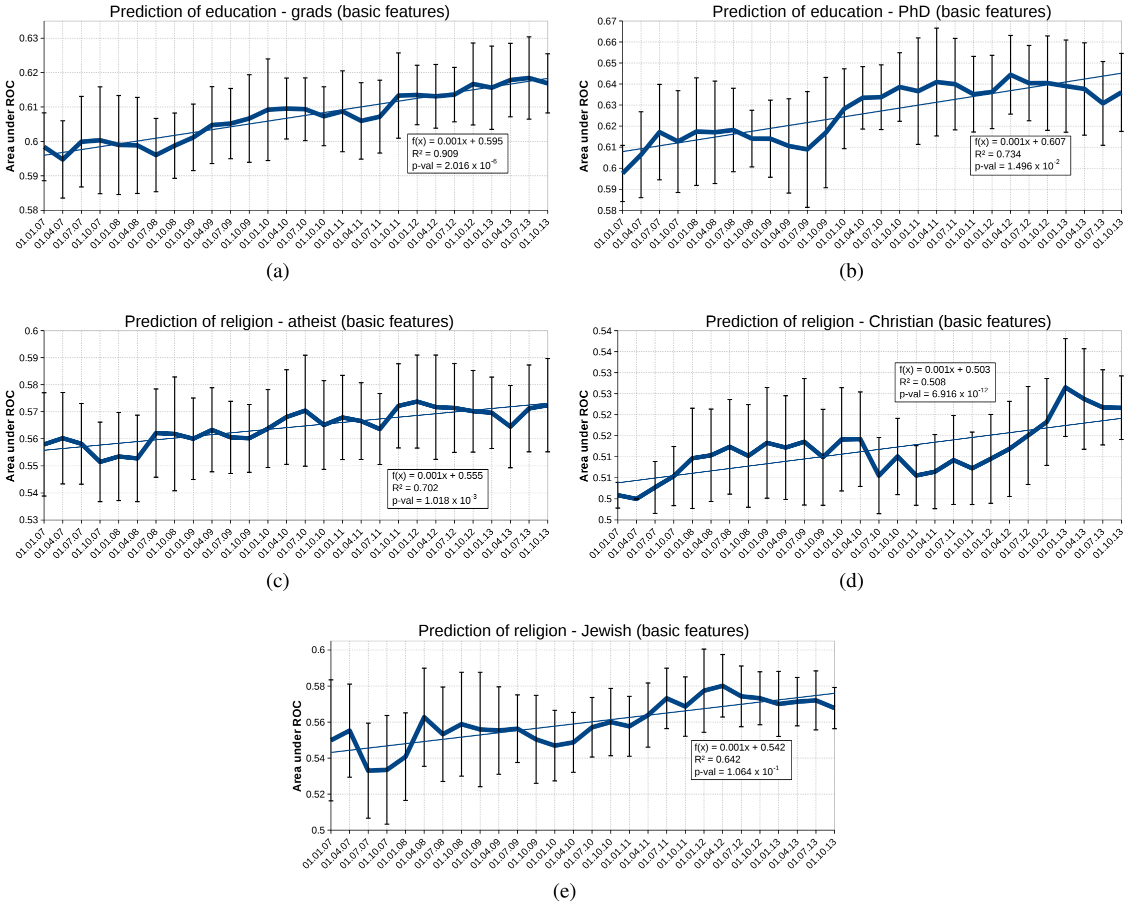


Figure 13. Results on the remaining binary predictors: Temporal evolution of the AUC, measuring privacy loss. Result of inferring, using binary predictors on the basic feature set, of education (*graduates* and *PhD*) and religion (*atheist*, *christian* and *jewish*).

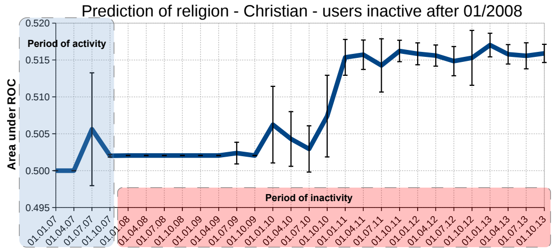


Figure 14. Increase of prediction performance for for users retired after 01.2008 – religion/*christian*.