Publication venue: **WSDM ’16**

Research Article
13 July 2015

Supplementary Material: Evolution of privacy loss in Wikipedia

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] *NICTA & University of New South Wales, 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

The purpose of this document is to detail information, which would be i) cumbersome for the main article and ii) of limited interest for the main scientific message presented in the main article. This document does not introduce new notions and the information presented here-after is not necessary for the comprehension of the main article. We present it for completeness and reproducibility reasons.

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, 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.

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 thematic of the articles. This categorization system defines main categories, such as Geography, History, Arts, which can be further divided into further subcategories. This forms a shallow hierarchy, *i.e.*, a hierarchy with a low number of levels. Loops are allowed inside the hierarchy, though discouraged, and the resulting category graph does not posses a strict tree structure. We select the 26 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 and revision

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

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

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.

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 proficient 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 reconstructed 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 1a), profession (Figure 1b) and religion (Figure 2). 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.*, christians, muslims, atheist and jews 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:

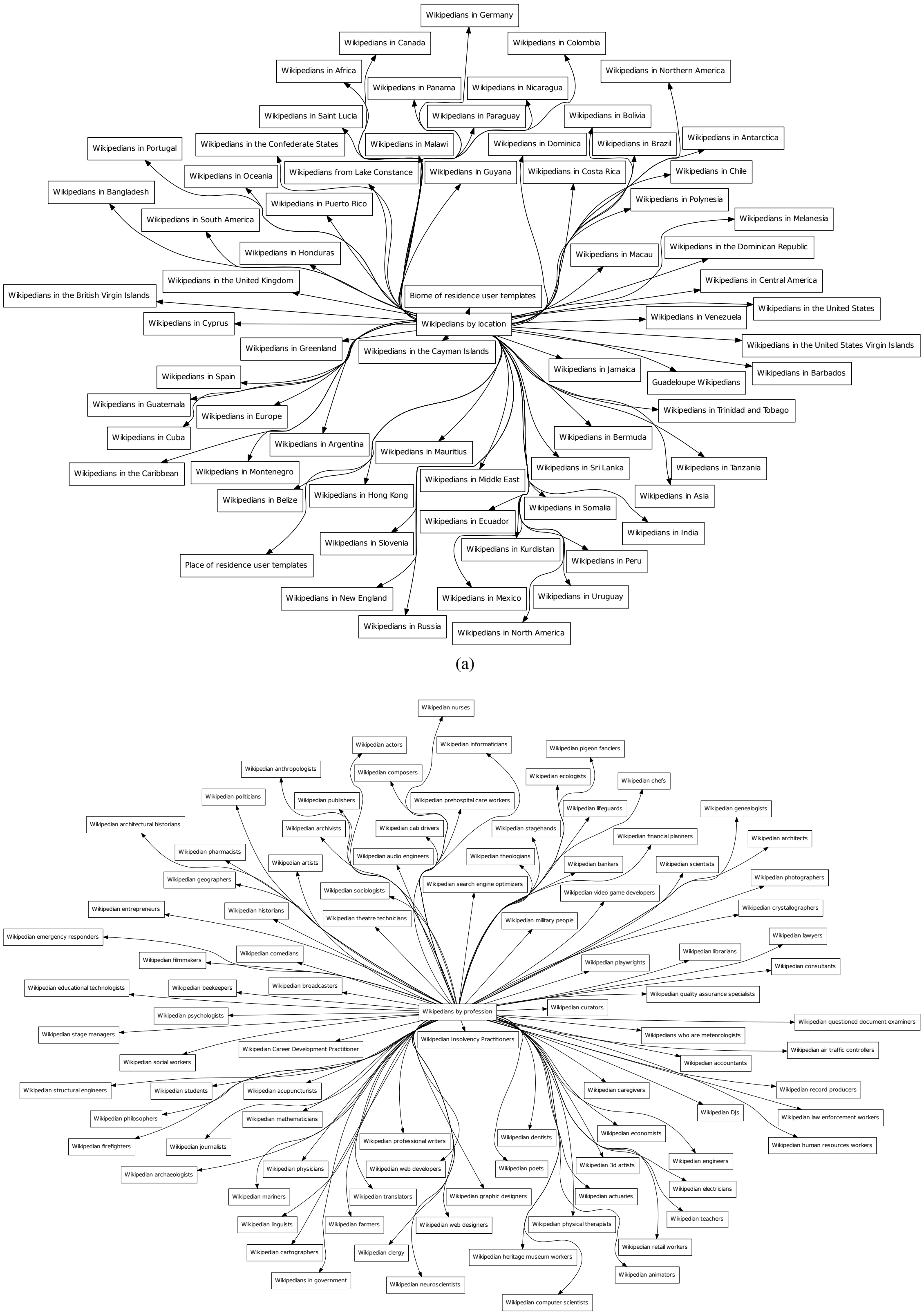
- a description of the user's editing activity, recorded down to revisions on individual pages, together with their timestamp;
- some private information, retrieved from their own user pages, under the form of user categories.

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.

REFERENCES

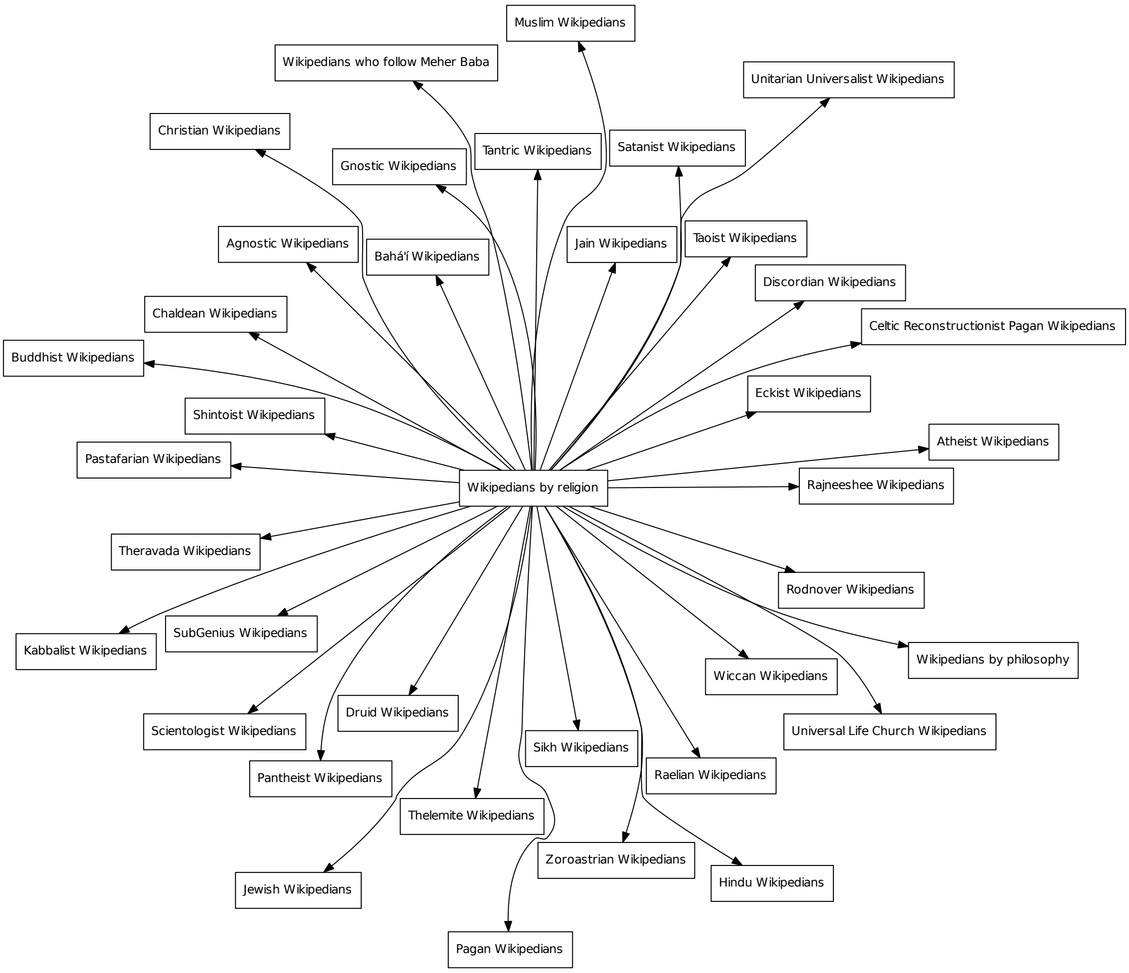
Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., and Kleinberg, J. (2012), Echoes of power: Language effects and power differences in social interaction, in World Wide Web, Proceedings of, WWW '12, 699–708



(a)

(b)

Supplementary Figure 1. 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)



Supplementary Figure 2. (cont. of Figure 1) Visual representation of user categories that can be extracted from the Wikipedia user pages, relating to editors' religion.