Article:
Sarasua, C., Checco, A. orcid.org/0000-0002-0981-3409, Demartini, G. et al. (3 more authors) (2018) The evolution of power and standard Wikidata editors: comparing editing behavior over time to predict lifespan and volume of edits. Computer Supported Cooperative Work. ISSN 0925-9724

https://doi.org/10.1007/s10606-018-9344-y

Noname manuscript No.
(will be inserted by the editor)

The Evolution of Power and Standard Wikidata Editors: Comparing Editing Behavior over Time to Predict Lifespan and Volume of Edits

Cristina Sarasua · Alessandro Checco · Gianluca Demartini · Djellel Difallah · Michael Feldman · Lydia Pintscher

the date of receipt and acceptance should be inserted later

Abstract Knowledge bases are becoming a key asset leveraged for various types of applications on the Web, from search engines presenting 'entity cards' as the result of a query, to the use of structured data of knowledge bases to empower virtual personal assistants. Wikidata is an open general-interest knowledge base that is collaboratively developed and maintained by a community of thousands of volunteers. One of the major challenges faced in such a crowdsourcing project is to attain a high level of editor engagement. In order to intervene and encourage editors to be more committed to editing Wikidata, it is important to be able to predict at an early stage, whether an editor will or not become an engaged editor. In this paper, we investigate this problem and study the evolution that editors with different levels of engagement exhibit in their editing behaviour over time. We measure an editor's

Cristina Sarasua
University of Zurich
Switzerland
E-mail: sarasua@ifi.uzh.ch

Alessandro Checco
University of Sheffield
United Kingdom
E-mail: a.checco@sheffield.ac.uk

Gianluca Demartini
University of Queensland
Australia
E-mail: g.demartini@uq.edu.au

Djellel Difallah
New York University
United States
E-mail: djellel@nyu.edu

Michael Feldman
University of Zurich
Switzerland
E-mail: feldman@ifi.uzh.ch

Lydia Pintscher
Wikimedia Deutschland
Germany
E-mail: lydia.pintscher@wikimedia.de

engagement in terms of (i) the volume of edits provided by the editor and (ii) their lifespan (i. e. the length of time for which an editor is present at Wikidata). The large-scale longitudinal data analysis that we perform covers Wikidata edits over almost 4 years. We monitor evolution in a session-by-session- and monthly-basis, observing the way the participation, the volume and the diversity of edits done by Wikidata editors change. Using the findings in our exploratory analysis, we define and implement prediction models that use the multiple evolution indicators.

1 Introduction

Knowledge Bases have become key to enabling semantic search and exploration functionalities in a wide range of applications on the Web. Besides the general interest knowledge bases owned and managed by companies (e. g. Googles Knowledge Graph), there exist openly available knowledge bases as a result of open data initiatives. For example, the Linked Open Data (Schmachtenberg et al., 2014) consists of over one thousand structured data sets describing various topical domains, and published by different agents including universities, private companies and other organisations. Wikidata is an open, free, multilingual knowledge base (Vrandečić and Krötzsch, 2014) started by Wikimedia Deutschland, containing as of October 2017 more than 37 million data items. While the creation of Wikidata was motivated by the need of more efficient data management methods within Wikipedia, Wikidata has become an important knowledge base for many other systems and applications that reuse Wikidata's item descriptions. Moreover, hundreds of other external data sets such as VIAF, the Library of Congress, Europeana or Facebook Places created by libraries, governmental and private organisations have been integrated with Wikidata, making Wikidata the hub to explore a network of open knowledge spread over the Web.

Wikidata, in contrast to the vast majority of knowledge bases on the Web of Data, has a strong focus on human intervention across its complete data management process. Tools and bots operate on the data to some extent, but the data is primarily curated and maintained by a community of volunteers. The community – editors, developers, data providers and researchers – discusses and collaborates to decide how to model, ingest and patrol information. This human-driven process enables Wikidata to diminish the data quality problems that appear in (semi-)automatically generated knowledge bases, including entity misclassification, inconsistencies, semantically inaccurate links, and outdated data (Zaveri et al., 2016).

Weaving a community of devoted people who are able to contribute duly of their own free will is one of the main challenges in Wikidata. Since Wikidata's mission is to represent human knowledge in a structured way, the project needs the help of a large number of people. Despite the positive response from thousands of volunteers, there is a clear need for attracting new contributors and growing the community, because there is still much data and many tasks to work on [1]. At the same time, it is important to retain contributors who already approached Wikidata by encouraging them to contribute. To address this challenge of acquiring new people and maintaining the activity of existing community members, Wikimedia organises events to advocate for free knowledge, disseminate the project, offer technical training for newcomers, establish social tights between members, as well as edit data and develop software together. There are also resources that aim to facilitate the editors' contribution. For

[1] Wikidata's Phabricator Ticketing System `https://phabricator.wikimedia.org/tag/wikidata/`

example, the Wikidata Games [2] present editors with very simple edit requests (e. g. add the occupation or the gender of a person) that help to improve the completeness of the knowledge base. Still, Wikimedia's official reports indicate that there is a big number of the people who one day edited Wikidata but are currently inactive. The latest statistics [3] show that from August 2016 until August 2017 the number of active editors [4] was between 7.8K and 8.7K. Taking into account that Wikidata registered more than three hundred thousand contributors during its history, the inactivity ratio is large. This fact could be detrimental for Wikidata, especially when we consider that many of these inactive contributors did not edit for a very long time and might have, presumably, abandoned the project. This level of abandonment rate suggests that Wikidata may not be maintained nor extended to its full potential. One may think that increasing editors retention in Wikidata could lead to more item descriptions and more complete and up-to-date data.

In order to improve the situation and reduce editor attrition, the community (and especially Wikimedia as the main manager) needs to design and implement methods that stimulate a change in the behavior of these contributors who become non-active. The literature in marketing research has extensively studied the problem of customer churn (or customer turnover) and designed churn management strategies that aim at increasing customer retention (Rosenberg and Czepiel, 1984; Ang and Buttle, 2006), because losing customers can endanger a business – having fewer customers often means obtaining a lower revenue. While in Wikidata (and in general in any Wikimedia project) the motivation to keep editors contributing to the knowledge base is not economic, there is attrition and participation inequality. Therefore, even though the concrete actions to engage further contributors are different from scenarios with economic transactions and customers, it makes sense to use assumptions and techniques from this field of research.

Marketing analysts and researchers highlight the importance of targeting customers to improve retention (Verhoef, 2003). However, it is important to provide tailored solutions based on the customers' behavior, especially since loyal customers need a different attention than likely-to-drop-out customers. That is why traditional retention strategies first tend to predict whether a customer will be a *churner* or not [5], and then implement actions to convert likely-to-be churners into non-churners (Gordini and Veglio, 2017; Difallah et al., 2014). For these actions to work effectively, the prediction needs to be done as early as possible. The subfield of survival analysis (Cox, 1992), often used in the context of retention management, estimates the "time left" until an event (e. g. time until death in the medical domain, time until drop out in the context of online shopping). This is useful to act against retention and intervene before customers leave. Hence, having a method to estimate the time that customers will spend in the system is a prerequisite to design a retention management solution.

In this article, we work towards developing such a method in the context of Wikidata. We aim at predicting the time that contributors will be in Wikidata – lifespan, as well as the volume of edits they will contribute with, to understand the magnitude of their action in the knowledge base. Therefore, our goal in this work is not to provide the solution to the retention problem designing a retention campaign but to build the prerequisites, pro-

[2] Wikidata Game https://tools.wmflabs.org/wikidata-game/distributed/#

[3] Wikidata Revolution presentation at Wikimania 2017, in August 2017 https://wikimania2017.wikimedia.org/wiki/Submissions/The_(Wiki)Data_(R)Evolution

[4] According to Wikimedia Foundation, an active user is "A user with 5+ edits in the main namespace of a given project over the last 30 days" https://meta.wikimedia.org/wiki/Research:Metrics

[5] Retention Science https://go.retentionscience.com/hubfs/Documents/Retention_Science_Predicting_Customer_Churn_Guide.pdf?t=1507690636756

viding a data-driven approach to the prediction task. In order to do so, first we carry out an exploratory analysis to understand differences between power and standard editors, and second, we provide and evaluate a predictive model that uses the insights of the exploratory data analysis.

Given that we look at lifespan and volume of edits, we consider 4 types of contributors: (a) contributors with *long lifespan* and *high volume* of edits, (b) contributors with *short lifespan* and *high volume* of edits, (c) contributors with *long lifespan* and *low volume* of edits and (d) contributors with *short lifespan* and *low volume* of edits. The contributors with the highest impact on the system are contributors who contribute extensively, and for a long time (group a). Groups (b) and (c) are also valuable contributors. For example, a contributor supervising recent changes to revert and correct malicious edits, a few times a month; she might do only a couple of edits a month, but if she does it for a long period of time, her contribution can help Wikidata in terms of data quality. Yet, ideally one would like to have as many contributors as possible in group (a). Predicting the lifespan and volume of the contribution of editors, we are able to classify existing contributors into one of these groups, and we can consequently, decide whom to address and how to do it.

In order to make such a prediction, we analyze the evolution of editing behaviour. We analyze this information from two different perspectives: we run a **(i) session-based analysis**, and we also study the editing progress **(ii) on a monthly basis**. The two perspectives are complementary: with the former, we aim at understanding the extent to which editors change their behaviour as they gain more experience and do more edits in each session they spend editing; while with the latter, we perform a time-sensitive analysis. When we analyze the behavioural evolution throughout the editors' lifetime (in sessions and in months) we measure indicators related to the editors' productivity, editors' participation and the diversity of the types of edits (cf Section 7).

To the best of our knowledge, there is no previous work analyzing the evolution of editing behaviour in these terms as a predictor of the volume of edits and lifespan in Wikidata. The research around Wikipedia, older than Wikidata, has examined the edit history in terms of edit quality, editor interaction, editor participation, as well as emerging information cascades (see Section 3 for a detailed description of the Related Work). Even if both systems share commonalities, Wikidata has features that could, in theory, encourage people to work with different patterns than in Wikipedia. Besides that, there is no published work about the intersection of both communities; so, it should not be assumed that contributors in Wikipedia and Wikidata work exactly in the same way. It is, thus, important to collect empirical evidence and study the behaviour of contributors in the Wikidata environment, too. Anyhow, other works have not used the trend in the evolution of contribution, participation and diversity as predictive factors of lifespan and volume of edits. Therefore, not only do we contribute to the state of the art by studying contributors in Wikidata, but we also contribute with a method. In the context of Wikidata, Mueller-Birn et al. (Müller-Birn et al., 2015) grouped editors based on the types of tasks they focus on, and as an extension (Cuong and Müller-Birn, 2016) observed the extent to which Wikidata editors change between roles. Piscopo et al. (Piscopo et al., 2016) surveyed Wikidata editors to understand differences between novice and expert users, examining how motivations, goals, usage of interfaces and type of actions differ. While these works help to understand the fundamental differences between some groups of editors, and the way editors change the type of actions they work on as they become more experienced, these works fail to (i) provide a method that helps to predict the extent to which editors will be engaged in terms of time and contribution, and they (ii) do not explain how power and standard editors change or maintain their contribution, participation, and diversity of type of edits.

The main contributions of this paper are:

- We run a quantitative analysis about the volume of edits and the lifespan of editors (Section 6).
- We perform a longitudinal study over the Wikidata history and identify the trends in the evolution of editing behavior of different groups of editors, mainly standard and power editors (Section 7).
- We define supervised classification methods to predict the range of months that an editor will be contributing to Wikidata, and the range of edits that an editor will do in Wikidata (Section 8).
- We highlight a set of implications that our findings may have in the Wikidata community Section 9.

2 Wikidata: A Crowdsourced Knowledge Base

Wikidata is a freely-available openly-editable knowledge base. It is the result of a continuous community effort started by Wikimedia Germany in 2012. More than just human contributions, Wikidata serves as a data integration hub where other knowledge bases (e. g. VIAF, Europeana, DBpedia) link to or are imported in[6].

Wikidata is set apart from many of the other available knowledge bases in several ways:

- *Community curated*: Wikidata's initial primary goal was to support Wikipedia editors by providing them with a central knowledge base that holds data to be shared between all Wikimedia projects. In order for this to happen and in the spirit of its sister projects, Wikidata is open for editing by anyone and maintained by an open community of editors.
- *Multilingual*: In order to fulfil its initial primary goal of supporting Wikipedia editors, Wikidata needs to provide one central place to collect and maintain the same data that is then shared between all Wikimedia projects. All editors must work on the same data independent of their language. Wikidata achieves this goal by identifying its items and properties (the basic building blocks of Wikidata used to describe concepts in the real world) with language-neutral identifiers. For example, the property "instance of" is identified by "P21" and the concept "Earth" is identified by "Q2".
- *Knowledge diversity*: Wikidata is built as a secondary database. This means it is not meant to record raw facts. Instead, it collects statements from other sources and references them. With this model, different views can be recorded on controversial topics and the consumer of the data can investigate further and judge which of the sources they accept. This is crucial for Wikidata in order to cater to the many different cultures in Wikimedia projects. It also aligns with Wikipedia's ethos of referencing information and making it possible for the reader to dig deeper into a topic.

Wikidata is set apart from its sister project Wikipedia in several ways as well:

- *Language and culture*: Wikipedia is divided by language and by extension culture. In Wikidata editors from all these languages and cultures come together to work together on the same data.
- *Notability*: Wikidata serves all Wikimedia projects and therefore needs to cover more concepts than any of the individual Wikipedias.

[6] In 2015 Google announced the port of the Freebase knowledge base to Wikidata.

- *Large scale editing*: Wikidata, as a result of its virtue of being machine-readable and editable, is seeing considerably more edits done with the help of tools and bots than the Wikipedias. Indeed, it currently accumulates one third of all edits across Wikimedia projects.
- *Editing interface*: Wikipedia offers a text-editor like interface as well as a WYSIWYG editor. Wikidata offers a form-based interface as well as a large amount of special-purpose tools (e g. WikiShootMe, Wikidata Game).

3 Related Work

The consolidation of social computing has led to multiple studies that examine human behaviour in various systems, including major volunteer crowdsourcing projects like Wikipedia and Open Street Maps. We review related work focusing on methods and findings about the contribution of Wikidata's and Wikipedia's volunteers, user attrition in Web systems and the evolution of user behavior in volunteer systems.

3.1 General Knowledge about Volunteers' Contribution in Wikidata

One of the first works looking at Wikidata, by Mueller-Birn et al. (Müller-Birn et al., 2015) analyzed the emerging roles of editors, in terms of the possible operations in Wikidata, including the creation of items and ontology elements, as well as the addition of references. The results showed that a majority of editors have specialised contributions, and only a small active group contribute across many areas of the project. As a continuation of that work, the authors in (Cuong and Müller-Birn, 2016) looked at 2 years of edit history, to analyse the transitions between editorial roles and found that "users who joined earlier are persistent contributors even though they take part in different roles, whereas users who join late are quite stable in their behavior". We divide contributors in a different way, based on lifespan and volume of edits instead, and we run a temporal-based analysis.

A more recent work by Piscopo et al. (Piscopo et al., 2016) looked at the way editors grow from novice to proficient contributors. The authors performed a qualitative analysis surveying Wikidata editors, and found that editors become more responsible for their work over time, and as time goes by they participate more with the community, carry out more advanced tasks and use more different tools. Compared to this work, our differences are that we have a different focus (prediction using the evolution), we look at the progress of different dimensions (contribution, participation and diversity of the type of edits), and we run a data-driven quantitative analysis considering edit sessions – while (Piscopo et al., 2016) surveyed editors.

3.2 General Knowledge about Contributions in Wikipedia and Other Knowledge Bases

In (West et al., 2012) West et al. look at who the Wikipedia editors are and how their Web usage patterns differ from non-editors concluding that editors are typically more expert on certain topics and get informed on the Web before starting to edit Wikipedia. Another work looking at Wikipedia edit patterns (Yasseri et al., 2012) performed a large scale analysis of Wikipedia edit data across different languages by geo-locating editors and looking at temporal edit patterns. The authors identified two main categories of editors: those more

active during week days and those more active during weekends. In (Iba et al., 2010) the authors leveraged Wikipedia edit data to build a social network across editors. They observed a differentiation between two main types of articles: those topic-focused having few expert editors and those of general interest involving many casual editors.

Previous work on Wikipedia has also focused on measuring and predicting the quality of contributions. For example, in (Druck et al., 2008) the authors looked at contribution quality prediction using the revert time of an edit (*expected longevity*) as quality measure. The prediction methods used are based on features such as change type, words used, article, and user. The authors concluded that the prediction of the edit quality depends very much content-dependent, and not only user-dependent. Another work looking at quality, by Halfaker et al. focused on how editors perceive revert actions on their contributions (Halfaker et al., 2011). The authors showed that reverts are good to improve the overall quality of Wikipedia but also affect user motivation and future engagement. They also highlighted the 'newcomer retention problem' which we look at in our paper. Finally, there are works like (Walk et al., 2015) by Walk et al. that studied the way the content of the knowledge base impacts the editing behaviour. The authors studied the sequences of edits that users in an ontology editing environment did over several ontologies, and compared different hypotheses using the HypTrails (Singer et al., 2015) method. They observed that the hierarchical structure of the ontology and the entity similarity are the dimensions that have the strongest influence on the behaviour of editors.

The focus of our work is primarily on the editors' engagement patterns, independently from Wikidata's content because we would like to first understand if there are editors with different levels of commitment and different habits. This is high importance to the community, because a project like Wikidata greatly benefits from unconditional contributors who provide knowledge, no matter the status of the knowledge base.

3.3 User Engagement and Attrition in Volunteer Communities

Measuring user engagement in volunteer projects is key to understand who the most valuable users are and to design mechanisms that decrease attrition. A work related to the effect of Wikipedia edit activity on engagement is (Gandica et al., 2015), which defines a function of edit probability where the more a user has edited the more likely she is to edit in the future. Danescu et al. (Danescu-Niculescu-Mizil et al., 2013) study how users join and leave an online community focusing on the linguistic aspects of their contributions. They provided a method for predicting the range of time when users would drop out of the community based on their use of words over time when writing posts in a beer-related community. Their empirical research showed that the linguistic evolution stabilises after a while staying stable until drop out. As the authors showed, in the data set they analysed, this phenomenon is relative to the lifetime of the user, and not to absolute to a biological frame. Ponciano et al. (Ponciano and Brasileiro, 2014) measured the activity ratio and the daily devoted time by users in the citizen projects of Galaxy Zoo and The Milky Way Project, grouping people into profiles like the hardworking, spasmodic, persistent, lasting and moderate users. In both data sets, they found that the majority of users are classified as moderate, and only a few are persistent users.

In our work we also look at contribution and participation, but using different measures (i. e. number of edits per month / session, number of edits per item, number of items and seconds invested in the session) and observing the trend –increasing, decreasing or constant–

of the measurement over time. Furthermore, we compare power and standard users, without clustering them.

3.4 Evolution of User Behaviour in Volunteer Communities

The work of Geiger et al. (Geiger and Halfaker, 2013) was one of the first works applying the common technique in Information Retrieval of grouping edit activities into edit sessions in which an editor performs a number of related edits and then stops for a certain period of time (e.g., few days or weeks). Geiger et al. empirically defined Wikipedia sessions as of a one hour inter-edit time and then looked at time elapsed between sessions and at the evolution of sessions over years. In our work we follow the methodology they used to define edit sessions in Wikidata, and use the inter-session time to observe participation. As a difference, we look at the progress over time to use it in the prediction.

The study done by Panciera et al. on Wikipedia editors (Panciera et al., 2009) is extremely relevant for our work. The authors discovered that "Wikipedians are born, not made", which means that editors do not contribute more, more frequently or with higher quality over time, and rather maintain high levels of contribution. As the authors explain, these findings suggest that the system does not encourage further engagement and people who are truly committed are devoted to the cause of free knowledge from the beginning. In our work we test this hypothesis in the context of Wikidata, adding hypotheses and observations about participation and task diversity over time, as well as defining evolution in two different ways (i. e based on months and sessions). Moreover, we complement the exploratory research about the evolution of editing behavior with a prediction problem in terms of lifespan and volume of edits, as prediction is a central component in churn management and in computational social science in general (Alvarez, 2016; Strohmaier and Wagner, 2014).

Walk et al. provided a model for activity decay in collaboration networks (Walk et al., 2016) that captures activity decay rate and peer influence growth rate. They evaluated their approach with Semantic MediaWiki data sets, to prove that the activity dynamics they simulated is close to reality. While the goal of the paper is not to describe the data sets, the empirical evaluation showcased the activity decay present in several communities. In our works, we do not observe the interaction between editors.

In the context of OpenStreetMap and humanitarian mapping, Dittus et al. (Dittus et al., 2016) found that different contributor cohorts (working around different initiatives) show different retention patterns. Initiatives that were designed with tight coordination practices (e. g. including mapathons) had contributors with higher retention. Interestingly, the authors also found that early abandonment was related to higher contribution, suggesting that some standard contributors had a burnout effect. In the descriptive part of our research we measure the relation between lifespan and volume of edits.

4 Research Hypotheses

The volunteering-based design in Wikidata, akin to any other Wikimedia project, encourages the contribution of intrinsically motivated people who believe in free knowledge and are eager to help with their expertise and cooperation. The system provides mechanisms for accountability and transparency (i. e. anyone can see who did what, and discussions are

public), and strong contributions (and contributors) are openly acknowledged and recognised. The feeling of belonging to the community also drives volunteers in Wikidata, like in Wikipedia (Nov, 2007).

Because our goal is to understand and predict who will and who will not thrive as volunteer, we support our research in past studies and related theories that highlight differences in the behavioural evolution of effective and non effective people –related to the volume of edits – and committed/persistent and uncommitted people – related to lifespan.

The work by Panciera et al. (Panciera et al., 2009) suggests that Wikipedians maintain a constant level of contribution. The fact that the Wikidata and Wikipedia communities share some commonalities, makes us hypothesise that this is a key difference between power and standard editors in Wikidata, too. Note that it is still worthwhile testing the hypothesis empirically, because Wikidata has many more ways to contribute than Wikipedia, and it has a unique feature regarding knowledge curation, as compared to Wikipedia, that lays in the intrinsic structured nature of its content (i. e. each item is formed by a collection of structured factual statements rather than encyclopaedic articles written in natural language). These two differences could potentially influence editing behaviors. If editors show a constant editing behavior, it suggests that they have habits. Habits are related to commitment and effectiveness (Duhigg, 2012), and often the users who do not develop habits are related to churn. We set up three hypotheses, focusing on contribution (e. g. number of edits per session / month), participation (e. g. time invested in a session) and diversity (e. g. type of task variability).

Hypothesis 1: *A constant contribution over time is a signal of power editors but not of standard editors.*

If editors develop their editing habits, they are likely to schedule them regularly in their agendas, and the longer a habit runs for, the more established it becomes (Duhigg, 2012). So, in terms of the time spent while contributing, we hypothesise:

Hypothesis 2: *A constant participation over time is a signal of power editors but not of standard editors*

The fact that Piscopo et al. (Piscopo et al., 2016) found surveying editors, that indicates that editors take more responsibility and do different tasks over time (when they grow from novice to proficient), makes us hypothesise that an increasing trend in the diversity of tasks over time differentiates power from standard editors, assuming that a standard editor will have a lower probability of crossing the line from novice to proficient editor. Hence, we formulate the third hypothesis as:

Hypothesis 3: *An increasing diversity in the types of tasks done is a signal of power users but not of standard users*

If these hypotheses are confirmed, these dimensions will help us predict the class to which contributors will belong (i. e. power or standard contributors in terms of lifespan and volume of edits).

5 The Wikidata Edit History Dataset

We obtained the XML data dump (as of 01.07.2016 [7]) provided by Wikimedia containing information about each of the editing actions done by contributors. We parsed the data, transformed it into CSV data, and imported into a memory-based database (MemSQL). For each edit, we kept information about the editor who did the edit, the timestamp when the edit was completed in Wikidata's database, the item where the edit was done, and the comment

[7] Wikidata Wiki dump `https://dumps.wikimedia.org/other/incr/wikidatawiki/`

that MediaWiki automatically generates to annotate the changes in the database. We then classified the edits based on the (i) type of editor, (ii) the type of thing edited, (iii) the means used to edit, and (iv) the type of edit carried out.

- *Type of editors*: We distinguish between users who are registered users (i.e. have a username and edit Wikidata while being logged in) and users who are anonymous (and from whom we only know an IP address). Registered users can be humans or bots. We identify bots by looking up the public list of registered bots and discard the edits done by this set of users, because we are primarily interested in understanding human behaviors of Wikidata editors. It is important to distinguish between registered and anonymous users, not only because people might behave differently when they reveal their identity, as Shih-Wen et al. Huang and Fu (2013) showed in the context of microtask crowdsourcing, but also because non-registered edits might also come from applications implementing automatic edits via the Wikidata API, and hence show a different behaviour.
- *Type of things edited*: We distinguish between item edits and non-item edits. Item edits are edits done to create, update or delete an item in the knowledge base (e.g. an entity, a class). Non-item edits are edits done in other kinds of pages such as project and user pages. We only focus on edits done on items which are part of the knowledge graph.
- *Means to edit*: There are various interfaces to edit Wikidata (e.g. the wiki, Wikidata games, etc.). We differentiate between edits done using tools and edits done without tools, because in the former case users do not decide what item to work on next, nor the type of edit to do. To classify edits into these two groups we use the *OAuth* tags database provided by Wikimedia and scan the edit comments for any other trace left by tools listed in Wikimedia directories (including Gadgets. User scripts and external tools).
- *Type of edit*: we use the list of actions registered in Wikidata's backend [8] to distinguish between the major actions (e. g. set a label, update a claim, delete a claim or add a reference).

We have published all the preprocessed data, as well as our research data online [9].

6 Quantitative Analysis of the Wikidata Edits

Before addressing the topics raised in the definition of hypotheses, we obtain descriptive statistics that allow us gain a better understanding of the data set. Out of the complete set of 350+ million edits, 1.5+ million edits are done by anonymous users and 261+ million edits are done by bots. We exclude both sets of edits from our analysis, as we are interested in studying human editing behaviour. We limit our analysis therefore to a raw data set of 87+ million edits.

As expected, the number of edits done with tools exceeds the number of edits done manually (see Table 1). The number of distinct editors editing manually (without tools) is higher than the ones using tools. An explanation to this fact might be that there are sporadic editors who make a few edits in the wiki to try out the system or to maintain specific targets, but do not get involved with tools like the Wikidata Game or QuickStatements.

Given that editing behaviours may show different type of patterns when using tools, we decided to focus our analysis on the set of edits done by registered users without tools. The

[8] Wikibase actions `https://www.mediawiki.org/wiki/Wikibase/API/de`

[9] Preprocessed Wikidata History `https://github.com/criscod/wikidata_editors_evolution_jcscw2018`

	Registered Without tools	Registered With Tools
Number of edits	35,069,629	52,345,356
Items edited	7,633,131	13,065,045
Non-Items edited	176,892	1,392
Number of distinct editors	142,643	6,060

Table 1 Different types of edits in Wikidata's history (from October 2012 until July 2016) that we consider in our analysis.

data used by default for all the results presented in all the following sections is therefore referring to this set of 35+ million edits done over 7+ million items of the knowledge base.

6.1 Volume of Edits

To understand the influence that different users have in the system, we computed the total number of edits made by each user and plotted the histogram. As it can be seen in Figure 1, there are many users who make a low number of edits and few users who make a high number of edits, as expected. The median of the edit counts is 2 edits, while the standard deviation is 7223 edits. This kind of behavior is actually similar to what we observe in other crowd-powered systems. In the context of paid microtask crowdsourcing, participation is typically dominated by few workers who complete most of the workload (Franklin et al., 2011). Similarly, in citizen science projects (e. g. GalaxyZoo) very few users perform many tasks, while the vast majority tags less than 30 images each (Lintott et al., 2008).

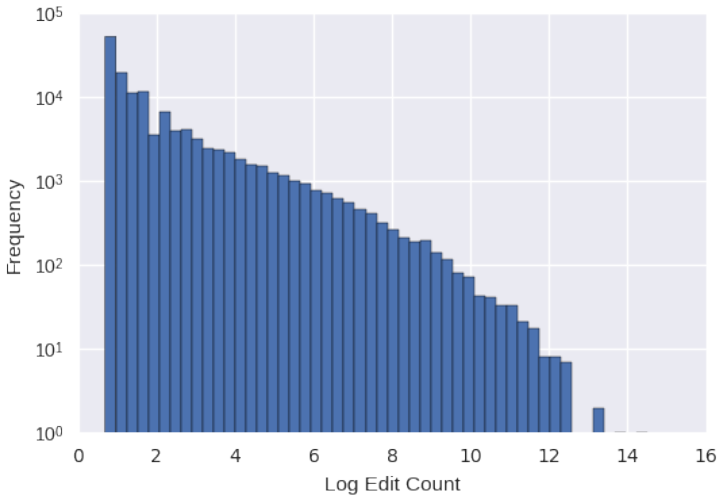

Fig. 1 Total number of edits done by each Wikidata user.

We also looked at the extent to which items are crowdsourced, by computing the number of distinct editors who have edited each item. Again there is a clear a long-tail in the distribution, as there are many items that have been edited by few editors, while there are few items that have been edited by many editors. The median is 1 editor per item, while the standard deviation is 3.1 editors per item.

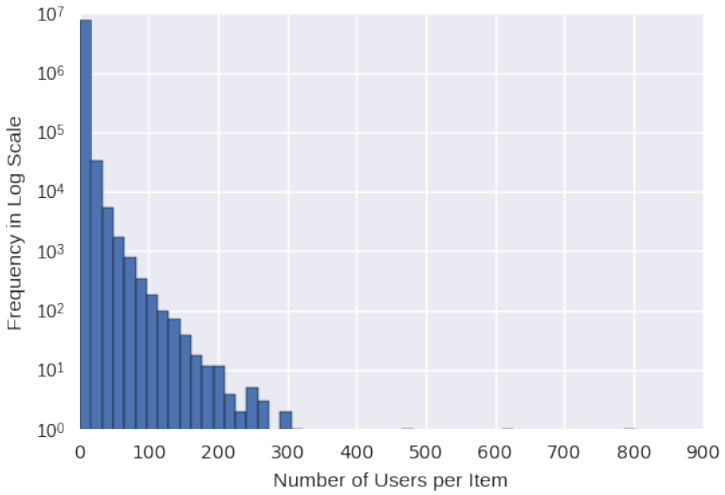

Fig. 2 Histogram of editors per item.

Finding 1.1: There is a skewed distribution of edit counts (i. e. few editors with many edits and vice versa).

Finding 1.2: There is a skewed distribution of editors per item (i. e. few items are edited by many editors and vice versa).

Figure 3 shows the boxplot with the counts of edits, by year in which editors started to edit. The median decreases with the years, which might be related to the fact that in the beginning there is a broader "blank space" to be edited.

6.2 Change in the User Base over the years

When we observe the timestamp of the first and last seen edits of editors, we can analyze for each year in our dataset the number of editors who joined, as well as the number of editors who were seen last that year, and the number of editors who joined but were seen last that year. As it can be observed from Figure 4, the number of people joining per year increased from 2012 until 2015, while from 2015 until 2016 this number decreased. Exactly the same behavior is observed for the last seen edit and the people whose first and last edit is seen in the same year.

Fig. 3 Edit counts per starting year (counts in log scale).

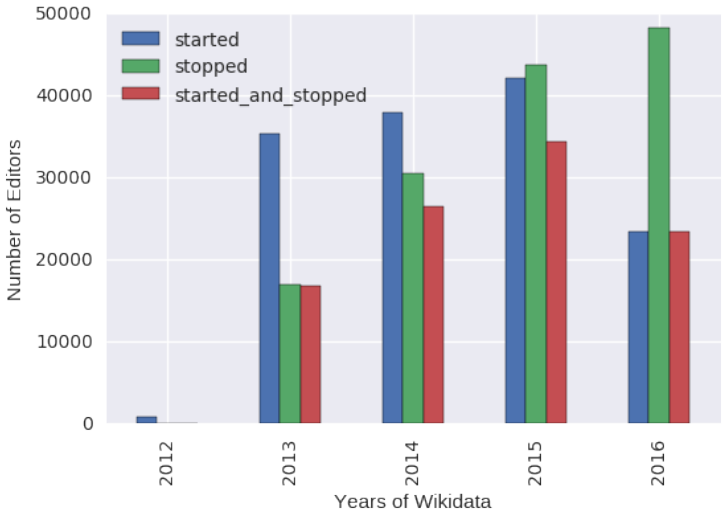

Fig. 4 Change in User Base.

6.3 Editor Lifespan

Wikidata (and Wikipedia) editors often have breaks (i. e. periods of time when they do not edit). For this reason, the Wikimedia Foundation introduced the notion of *active* and *inactive* users, to differentiate between users who are present and who users who "having a pause". A user is defined as active if in the last 30 days she did 5+ edits. However, a user that is within a period of inactivity has not necessarily abandoned the project. To measure the lifespan of editors, we need to distinguish between being temporarily inactive and gone. Therefore, we look at editors as being in one of three possible status: "active", "inactive" and "gone". The

distinction between active and inactive is given by Wikimedia, to distinguish those who are gone, we need an additional definition. Instead of defining an arbitrary threshold (e. g. of one year) of time after which we label editors as gone, we calculated it empirically. In order to do so, we analyzed the length (in months) of inactivity periods for all Wikidata editors. Surprisingly, the longest gap is of 16 months, which means that there were editors who, after 16 months without editing, came back to Wikidata and edited again. Looking at the percentiles we decide to use 9.967 months as a threshold in our data set to define editors that are gone and editors that are still in the system (either in active or inactive mode). Once we labeled the dataset according to this threshold, we encountered that from the total of 140,330 editors (who edited items) 77.698 editors appear to be gone by July 2016. That is, around $55\,\%$ of the editors have abandoned the project.

To better understand the distribution of editors lifespan, we decided to analyze only editors who are gone (because only in that case we can be sure that we are looking at completed lifespans). Figures 5 and 6 shows the histogram for the editors lifespan. We can see that there are many users with a very short lifespan, and only few are long-lasting editors. There are editors who have been editing for almost 3 years.

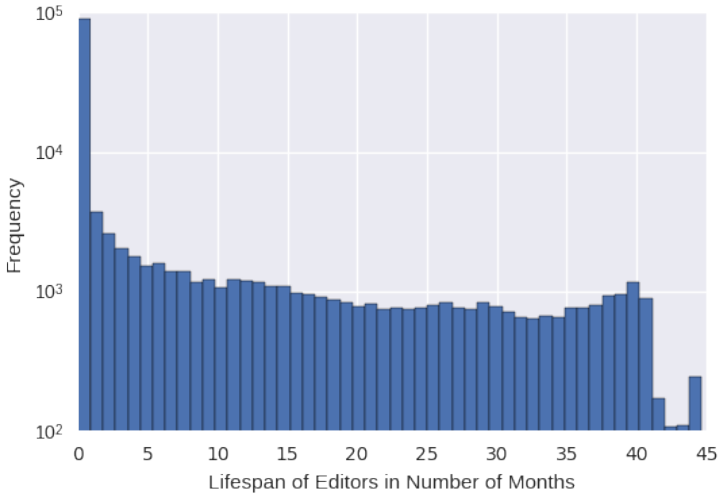

Fig. 5 Lifespan of all Wikidata editors in our data set.

When comparing this drop-out ratio to other systems, we find similar results. User participation in different types of online collaborative platforms shows similar patterns to the ones we have observed in Wikidata. Generally speaking, in online communities about 90% of the participants are inactive content consumers while about only 10% actively contributes content for long periods of time (Stewart et al., 2010). Other examples include participation in MOOCs being very skewed with numbers of participants completing the online course varying between 5% and 10% (Clow, 2013).

Figure 7 relates lifespan and number of edits done within the observed time frame between first and last edit. Obviously, with bigger lifespan editors may have bigger edit counts. However, interestingly the behavior is not linear, meaning that there are still people who are either slower or less committed, who have longer lifespan but the same number of edits.

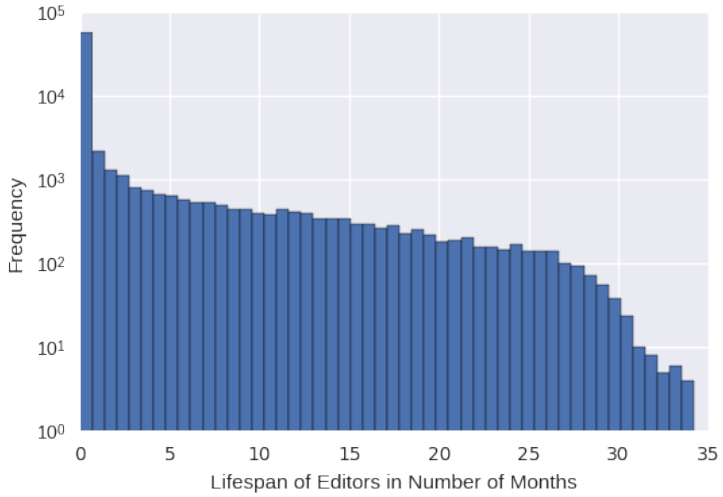

Fig. 6 Lifespan of Wikidata editors (only users that have abandoned the platform are shown).

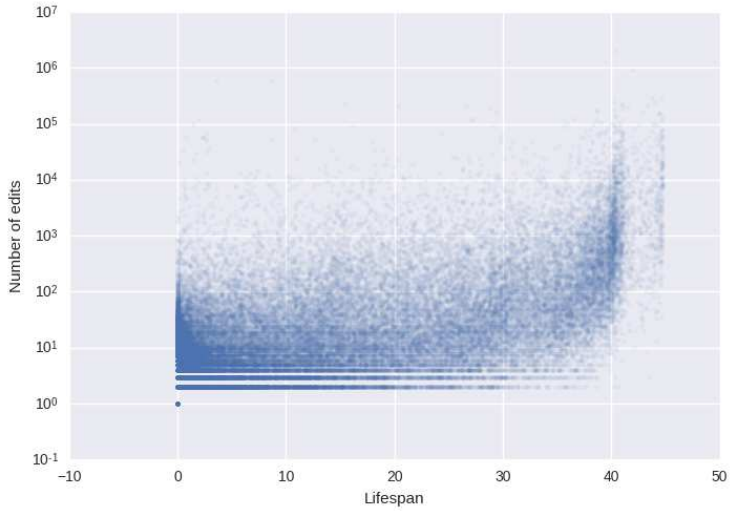

Fig. 7 Number of edits vs lifespan.

Finding 2.1: There is a slightly skewed distribution of lifespan (i. e. few editors were in the project for many months and vice versa)..

Finding 2.2: There is not a linear relation between the lifespan and the volume of edits done by editors.

7 Longitudinal Analysis of Wikidata Edit History

In this section we report about the exploratory analysis that we have done to understand the evolution of the editors in various dimensions. The goal of this analysis is to understand the trends that appear over time, and to find different groups of editors that should be addressed differently when designing engagement strategies.

7.1 Methodology

In this section, we describe the granularity of the evolution analysis, the cross-sectional indicators that we obtain for each editor over their lifetime, and the two goals that we set to distinguish between power and weak users.

7.1.1 Granularity of the Evolution Analysis

We analyze the evolution of editing behaviour from two different perspectives: first, we look at how editors behave as they get more experienced after doing batches of edits. To group edits, we define editing sessions, a common technique in Web-based systems. Second, we observe the way editors edit over the natural months in which Wikidata has been online. Note that the latter allows us to account for the exact occurrence of inactivity periods.

Analysing the time between edits A session is a slot of time in which a user takes a sequence of (often related) actions, that are temporally isolated from other sequence of actions. In our case, sessions are sequences of edits. In the literature we find works that analyze sessions in the context of Wikipedia (Geiger and Halfaker, 2013), but sessions have not yet been explored in Wikidata. It is interesting to do so, because the tasks that we can observe in Wikidata and in Wikipedia have a different nature and granularity. We followed the methodology defined by (Geiger and Halfaker, 2013), which has been widely accepted in the Wikipedia community. For each editor, we compute all the temporal differences between pairs of edits and plot a histogram (see Figure 8). Here we analyze all the set of registered human editors (who do not use tools), because sessions do not depend on the fact of having abandoned Wikidata or not. It is worthwhile mentioning that to generate the sessions, we consider both edits on items and edits on non-item pages (e. g. item discussion pages, user talk pages), as users tend to combine both types of edits in their sessions.

Defining the length of an editing session We analyzed the distribution of the edit differences (with logarithmic bucketing), and fitted it (with error $\chi^2 < 0.001$) as a sum of three distributions: the first peak, is a log-normal centered at around 1 second. The second distribution is an exGaussian distribution centered at around 10 seconds. The third distribution, centered at about 1 day is another log-normal distribution. We interpret the distributions analogously to what Geiger et al. did, looking at the long and short differences. From right to left, we interpret that the third distribution (with largest differences) represents the inter-session differences. The second distribution looks like the intra-session differences, whereas the first distribution with very small differences looks like intra-session edits. These small differences between intra-session edits seem very peculiar to Wikidata (and different from Wikipedia), because here users may edit multiple items simultaneously (as we will shortly illustrate). We looked at the set of edits in this distributions and we identified that many of these differences referred to deletions

and merge items edits. Moreover, Wikidata's GUI allows users to update multiple claims simultaneously, while giving the instruction to save all the updates after one click. Also, users may have multiple tabs open and edit interchangeably across them. The calculated threshold for defining using the aforementioned three-peaks fitting sessions when considering exclusively edits done without tools is 4.37 hours. Hence, we (programmatically) define a new session when two consecutive edits are separated by a time difference of at least 4.37 hours.

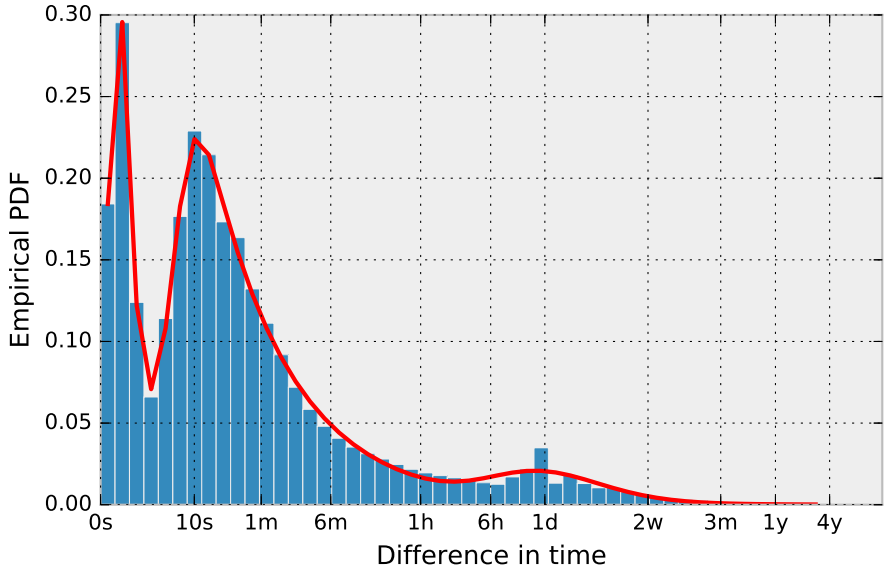

Fig. 8 Distribution of edit differences. The red, continuous line represent the best fit with two log-normal and one expGaussian distribution.

Finding 3.1: In Wikidata we find shorter times between edits than in Wikipedia.

Finding 3.2: We empirically define new sessions after 4.37 hours of inter-edit time, around 4 times longer than in Wikipedia.

From this model we compute the threshold that discriminates between inter-session and intra-session edit differences as about 4hours and 30 minutes. We can notice that, apart from the small second peak, the structure is similar to the Wikipedia session one, as shown in (Geiger and Halfaker, 2013). We labeled our data accordingly and grouped edits into sessions.

7.2 Editing Behaviour Indicators

We take into account the indicators used in the Wikimedia community [10] and we extend them with others defined by us. In particular, we look at three dimensions: (I) Editor contribution, (II) Editor participation, and (III) Diversity of the edits accomplished by editors.

Measuring Editor Contribution In order to measure editor contribution at each point in time we measure the following indicators:

- **i1**: number of edits. It looks at any kind of edit (from creation, to update and deletion).
- **i2**: average number of edits per item. It identifies any kind of edit done to items (either items of the schema- or the instance-level).
- **i3**: number of items edited.

Measuring Editor Participation We measure participation in terms of the time dedicated to the task of editing. The indicator that we use is:

- **i4**: number of seconds between the first and last edit in the timeframe being analysed. If there is only one edit, we decide to define i4 as 0 seconds. Note that this indicator is only relevant for a session-based analysis.

Measuring the Diversity of Tasks Accomplished by Editors To measure the diversity of of tasks that editors do, we distinguish between the different types of actions registered by Wikibase [11]:

- **i5**: diversity of types of edits. We use a standard measure for diversity, the the Shannon-Entropy (Shannon, 2001), to calculate the diversity of types of edits as

$$H(T) = -\sum_{t \in T} prob(T = t) \times \log prob(T = t).$$

  The probability is defined as the frequency of type t. When we compute i5, we normalize the entropy based on the number of edits done in the session/month analyzed.

Since we are interested in observing and comparing the **evolution of these indicators over time**, we first compute each of these indicators at several points in time and second, we fit a linear model using the RANSAC algorithm (Fischler and Bolles, 1981), a robust linear model estimator able to deal effectively with outliers. For each editor we obtain the values for their slope, intercept and R2. These values allow us to understand the general trend over time of an indicator for an editor. For example, when we fit a linear model on the different observations for i1, we can identify if a user is increasing, decreasing or maintaining the number of edits over time by looking at the slope of the fitted linear model. The intercept provides information about the scale and R2 indicates the error between the fitted model and the actual shape of the i1 time-based indicators.

7.2.1 Criteria for Identifying Power and Weak Users

The Wikidata community acknowledges the value of both (a) editors who contribute with a high number of edits, and (b) editors who contribute for a long time. As we see in Figure 7, these two dimensions are not necessarily always related. The optimal case is to have editors who contribute with many edits and stay a long time in the system. However, each of these

[10] Research Metrics `https://meta.wikimedia.org/wiki/Research:Metrics`

[11] Wikibase actions `https://www.mediawiki.org/wiki/Wikibase/API`

Evolution Granularity	Editing Behaviour Indicators	Goal of Analysis
Session-based	i1, i2, i3, i4, i5	Volume of Edits
Month-based	i1, i2, i3, i5	Volume of Edits
Session-based	i1, i2, i3, i4, i5	Lifespan
Month-based	i1, i2, i3, i5	Lifespan

Table 2 For different options to study the evolution of editing behaviour. i1 is the number of edits. i2 is the average number of edits per item. i3 is the number of items edited. i4 is the number of seconds between the first and last edit in the session – only valid for session-based analysis. i5 is the diversity of types of edits.

dimensions separately is useful. The former implies that the knowledge base may grow or improve (depending on the concrete edits), whereas the latter means that there are editors who could eventually be available in a call for participation.

For this reason, we consider that users can be power users or weak users in two different ways: (i) in terms of the volume of their contribution and (ii) their lifespan. We set this two-fold goal as the focus of our empirical analysis, and examine the evolution of editing behaviour following the 4 different configurations listed in Table 2:

7.2.2 Empirical Findings

The main purpose of the exploratory phase of our analysis is to identify differences in the evolution of behaviour between groups of editors, mainly editors with high volume of edits vs. low volume of edits, and editors with high lifespan editors vs. low lifespan editors.

Different Behaviours: Figures 9 to 15 show several scatterplots, where each depicts the slope (x-axis) and intercept (y-axis) for a particular indicator for each editor, in one particular evolution granularity. Figure 9 for example shows the plot for indicator i1, with a session-based evolution. We decide to plot these two dimensions along the x and y axis, ignoring R2 because it does not provide a useful signal to discriminate different behaviours between these groups of editors, at least visually (in Section 8, we will see that even such indicator will be useful for prediction).

We observe that there are the editors with constant slope (i. e. they perform according to past sessions/months), while others have positive or negative slope, with different values for intercept (negative, zero and positive). This finding confirms that there are different evolutions of editing behaviours present among the community, where some people increase the number of edits over time, other people show a decay in the time they invest, and other editors show a constant performance.

The colour of each dot indicates whether the editor represented by the dot has a high or low volume of edits or long or short lifespan (depending on the particular instance). As it is clearly differentiable in Figures 9 and 10, power editors (having long lifespan), seem to have a constant behaviour in terms of the output they produce over the months (vertical green cluster in the figures). In contrast, weak editors tend to change their behaviour, having either an increasing or decreasing evolution of their contribution (diagonal white cluster in the figures). In the case of session-based evolution, most of the editors with constant contribution are editors with long lifespan or high volume of edits, but the differentiation between the two groups is not as perfect as with the monthly evolution.

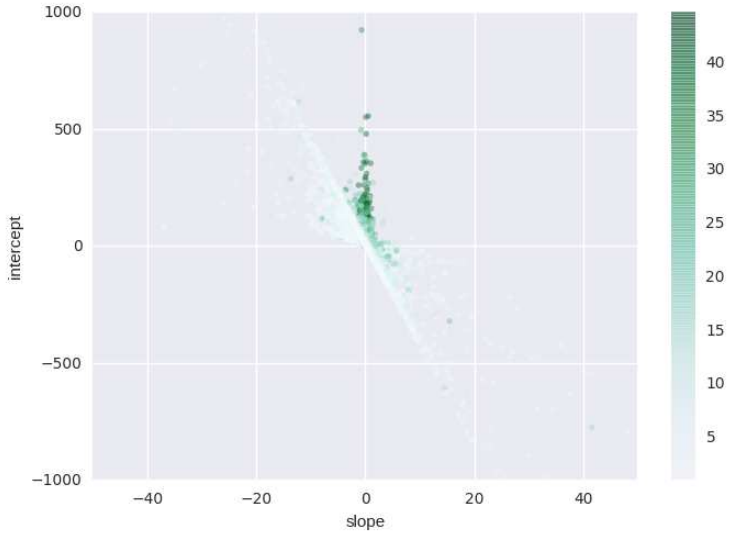

Fig. 9 Scatter plot indicating for each editor the slope and intercept obtained by the RANSAC algorithm for indicator i1 (number of edits), in a month-based analysis of evolution, depicting the editors' lifespan.

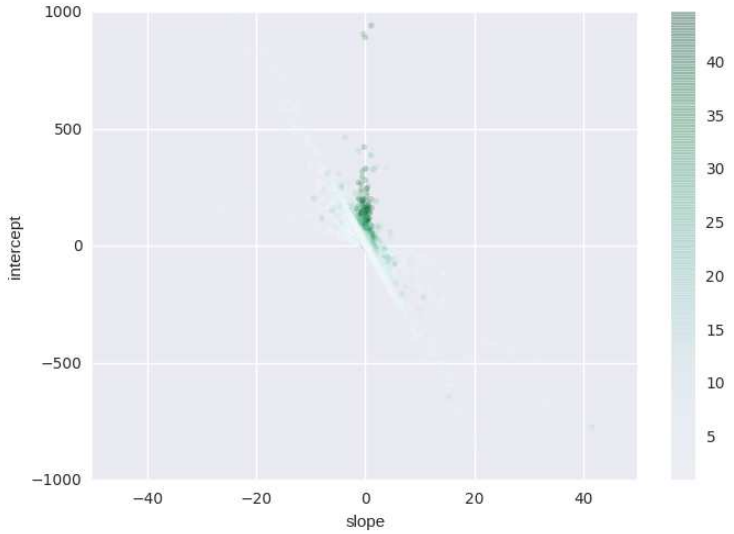

Fig. 10 Scatter plot indicating for each editor the slope and intercept obtained by the RANSAC algorithm for indicator i3 (number of items), in a month-based analysis of evolution, depicting the editors' edit count.

Finding 4: Editors with long lifespan have a constant contribution over months, while editors with short lifespan do not.

As for the way the participation evolves over sessions, we can see in 12 that editors with a higher total volume of edits have a constant participation, while editors with lower volume

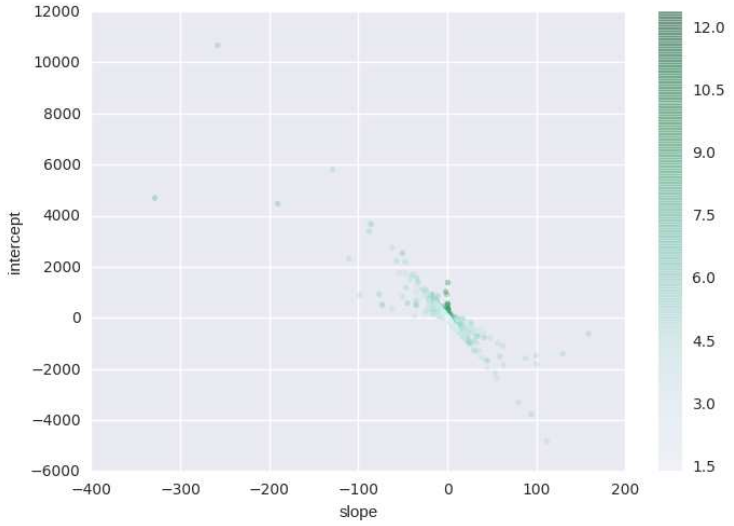

Fig. 11 Scatter plot indicating for each editor the slope and intercept obtained by the RANSAC algorithm for indicator i1 (number of edits), in a month-based analysis of evolution, depicting the editors' edit count.

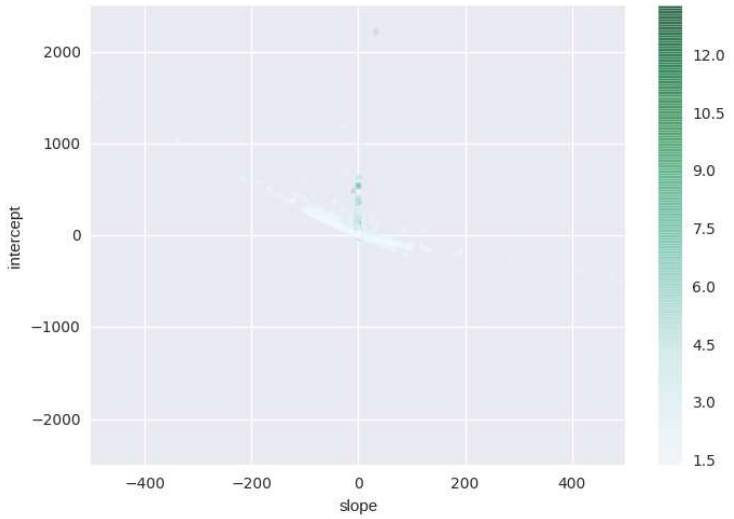

Fig. 12 Scatter plot indicating for each editor the slope and intercept obtained by the RANSAC algorithm for indicator i4 (seconds per session), in a session-based analysis of evolution, depicting the editors' edit count.

of edits have a rather increasing or decreasing evolution. When it comes to comparing the participation over sessions of editors with long and short lifespan 13, the separation between the two groups is less clear (because some power users also have a non-constant participation evolution), but it is still visible.

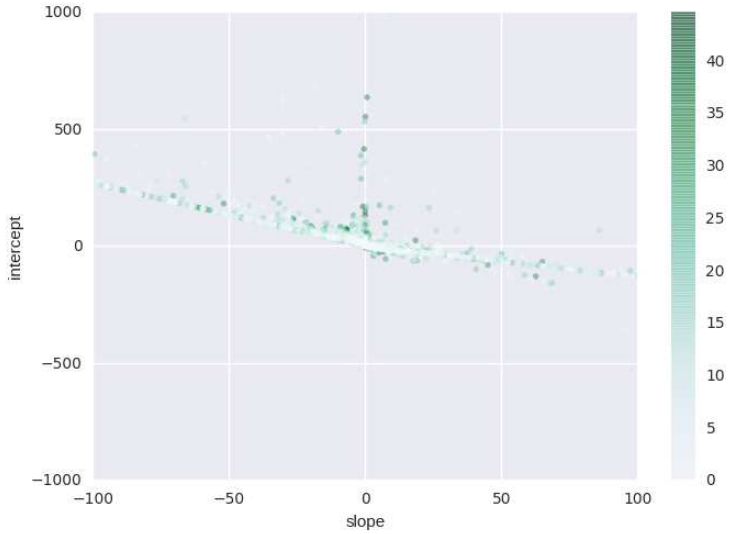

Fig. 13 Scatter plot indicating for each editor the slope and intercept obtained by the RANSAC algorithm for indicator i4 (seconds per session), in a session-based analysis of evolution, depicting the editors' lifespan.

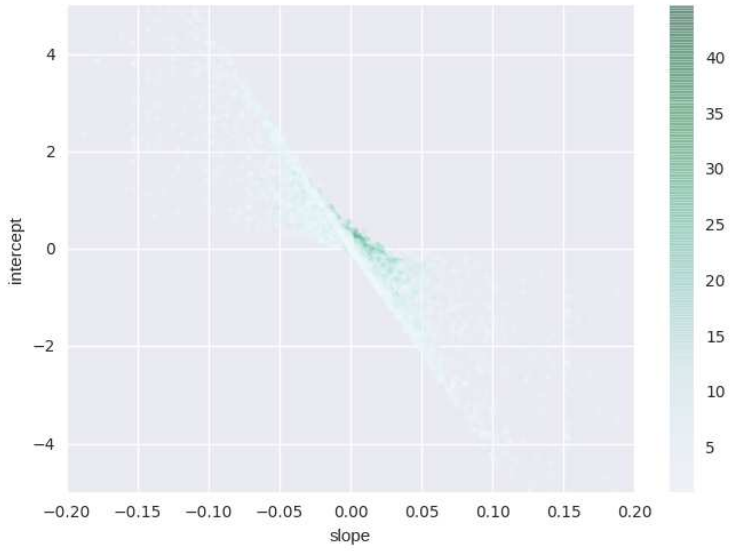

Fig. 14 Scatter plot indicating for each editor the slope and intercept obtained by the RANSAC algorithm for indicator i5 (diversity of type of edits), in a month-based analysis of evolution, depicting the editors' lifespan.

Finding 5: Editors with a high volume of edits have a constant participation over sessions, while editors with low level of volume of edits do not.

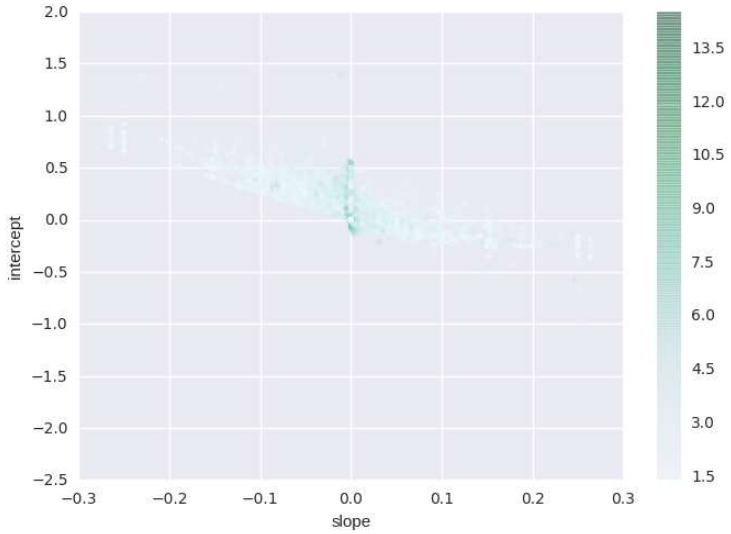

Fig. 15 Scatter plot indicating for each editor the slope and intercept obtained by the RANSAC algorithm for indicator i5 (diversity of type of edits), in a month-based analysis of evolution, depicting the editors' editcount.

Regarding the diversity of type of edits, editors with longer lifespan tend to increase the diversity over the months, while editors with smaller lifespan can either decrease or increase the diversity (cf. Figure 14). As for editors with higher total volume of edits, they seem to keep the diversity of the types of edits, while editors with lower volume of edits increase it or decrease it (cf. Figure 15).

Finding 6: Editors with a long lifespan tend to increase the diversity of the type of their edits, while editors with short lifespan can either increase or decrease it over the months.

Note that the RANSAC algorithm removes from the data set those editors who have an insufficient number of observations to draw any conclusion on the evolution of the behavior.

Given this result, we can compare the distribution of lifespan and edit count of different slope intervals. Figures 16 and 17 show a valuable example, comparing the histogram for the lifespan of editors with a slope with an absolute value smaller or bigger than 0.2, for indicator i1 (number of edits per session). There is a clear difference in the distribution of both histograms: the editors with absolute slope value smaller than 0.2 (and therefore closer to zero), have bigger lifespans and the frequency of bigger lifespan is also higher than for the editors with absolute value bigger than 0.2.

7.3 Model Building

In order to confirm our intuition about the relationship between lifespan/edit count and the indicators, we use a Random Forest regressor with the whole dataset as training data and visualize the predicted values on the indicator space.

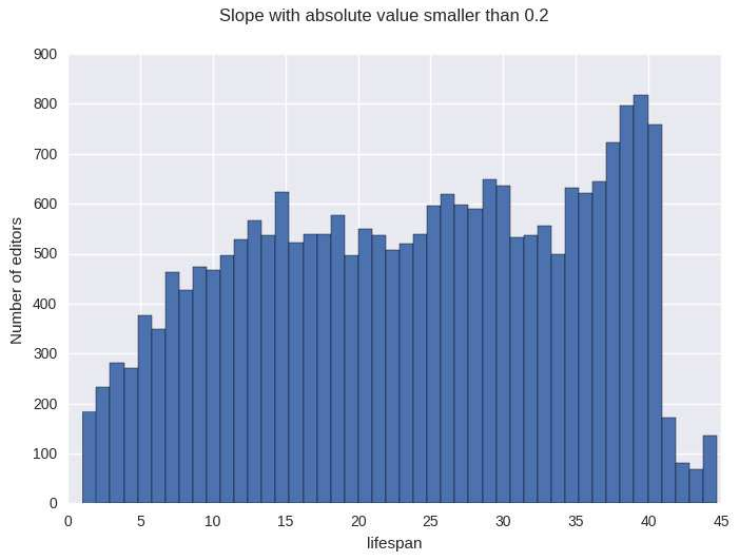

Fig. 16 Histogram showing the lifespan for editors with slope of an absolute value smaller than 0.2.

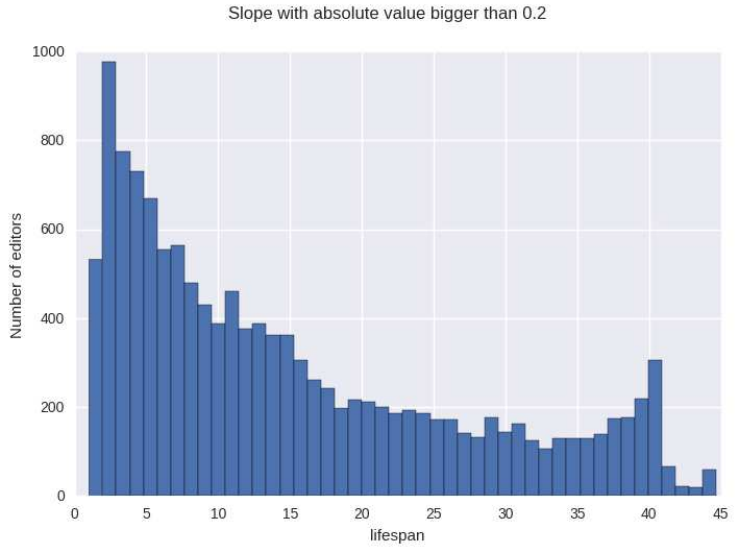

Fig. 17 Histogram showing the lifespan for editors with slope of an absolute value bigger than 0.2.

In Figure 18 and 19, we visualize a projection on the slope and intercept of the predicted lifespan for i1 and i2 respectively of the month based indicators. The predicted values corroborate the intuition about the absolute value of the slope as being a good indicator of high lifespan: for i1, a slope close to zero (constant behaviour) corresponds to high lifespan; for i2, we see that the second and fourth quadrant are associated with low lifespan, following the pattern visualized in Figures 9- 15. The corresponding projection for the session based indicators (Figures 20 and 21) are less clear, presumably because the interaction between all

different indicators are more involved, and a two-dimensional projection is not able to simply visualize them. We verify the prediction capabilities of such model in the next section.

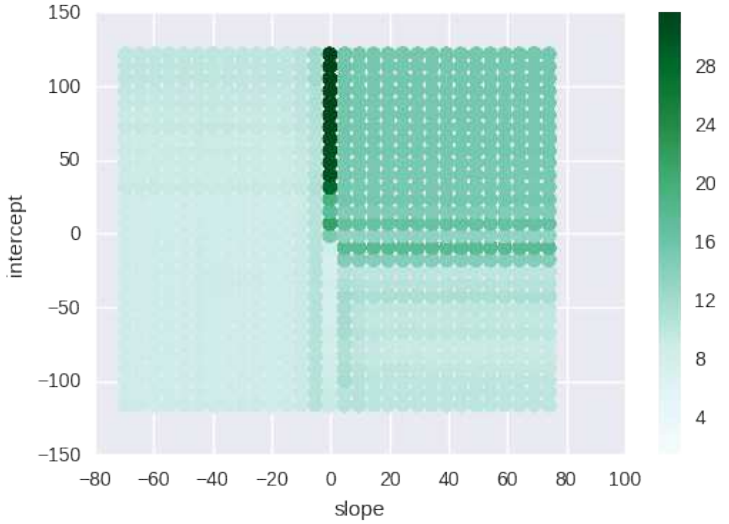

Fig. 18 Random Forest lifespan prediction on a projection of slope and intercept of i1 for the month-based indicator.

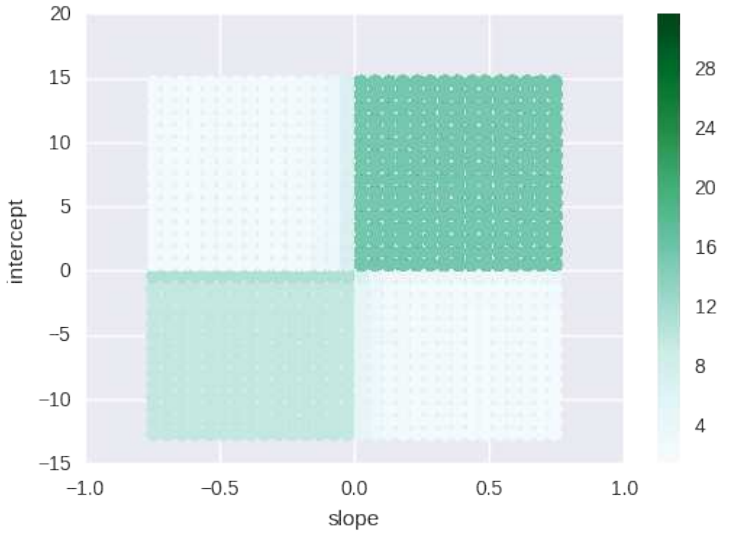

Fig. 19 Random Forest lifespan prediction on a projection of slope and intercept of i2 for the month-based indicator.

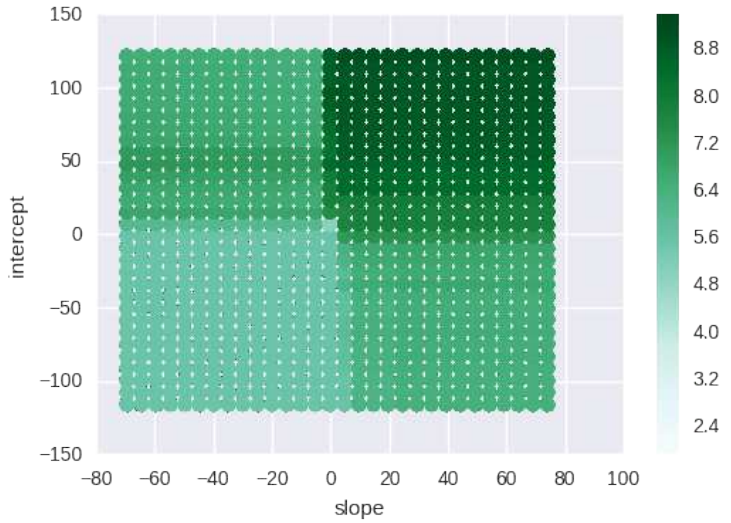

Fig. 20 Random Forest count edits (log scale) prediction on a projection of slope and intercept of i1 for the session-based indicator.

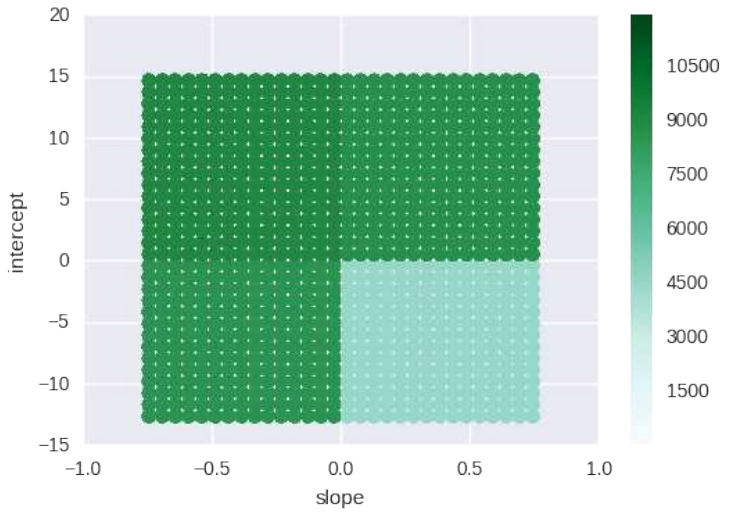

Fig. 21 Random Forest count edits prediction on a projection of slope and intercept of i2 for the session-based indicator.

7.4 Hypotheses Revision

As we can see along the Figures of Section 7.2.2, the difference between editors with long / short lifespan and editors with high / low volume of edits, in terms of the way they evolve is in some cases obvious and in others not. We can confirm Hypothesis 1, because we see that editors with high lifespan show a constant contribution, while other do not. The same applies to high volume of edits, although with less strength (cf. Finding 4). We can also

confirm Hypothesis 2 in the case of lifespan, because we see that people with longer lifespan maintain a constant participation, while editors with shorter lifespan do not. However, when it comes to volume of edits, the measurements do not help differentiating the two groups of editors as clearly. Hypothesis 3 is rejected, because we see that both in long and short lifespan, and in high and low volume, editors tend to increase the diversity of the type of actions they accomplish. Hence, evolution in contribution and evolution in participation are good indicators to differentiate standard and power editors in (especially in terms of lifespan), while diversity alone it is not sufficient.

8 Predicting Volume of Edits and Lifespan of Editors

Motivated by the findings presented in the previous section, we define our prediction tasks as follows: Given an editor, and a set of observations based on the aforementioned productivity, participation and diversity indicators, we predict

- whether the editor will be contributing with a high or level volume of edits and
- whether the editor will contribute for a long or short lifespan.

To solve these two prediction problems, we use supervised learning methods. More specifically, we use binary classifiers that take as input the information about the slope, intercept and R2 obtained by applying the RANSAC algorithm on the multiple indicators-based measurements, across sessions and across months. To define the thresholds for high / low volume of edits and long / short lifespan, and create the 4 classes (i. e. power and standard editors as for lifespan, and power and standard editors as for volume of edits), we observe the distribution of both volume of edits and lifespan, and decide that 15 months and 100 edits are suitable numbers to empirically distinguish between the different classes of editors that we define in terms of volume of edits and lifespan respectively.

We select two different classifiers: Random Forest[12] and a Logistic Classifier to compare their performance in terms of precision, recall, f1-score, and support. We evaluate the classifiers in different settings: (a) for a session-based evolution, we run the prediction evaluation with training data of size 100, 200, and 300 sessions. (b) For a month-based evolution, we pick training data of size 3, 5, and 10 months. The data that we use at this point is the complete set of edits done by humans without tools over items. We considered filtering out editors whose status is not 'gone', as in such case we cannot really state their lifespan with certainty. However, we noticed that after applying the RANSAC algorithm, there were no editors whose state is active and her lifespan is shorter than 15 months. The only editors who are not gone, have a lifespan longer than 15 months and therefore can also be labeled as having a long lifespan (no matter what the exact final value for the lifespan will be).

Figures 22 and 23 show the average F1-scores obtained for each class (power and standard) by the two classifiers, after running 10-fold subsample validation, for each training data size. In all cases, the Random Forest classifier outperforms the Logistic Classifier. If we compare the two plots of Figure 22 to the two plots of Figure 23, we observe that the former show stable F1-scores, even when augmenting the number of months or sessions in the training data. In the latter plots, there is a trend to increase the F1 with a bigger training data size.

According to the results, we obtain a higher F1 when we predict lifespan than we predict volume of edits.

[12] The Random Forest parameters chosen are: 100 estimators and bootstrap technique with subsample class balancing.

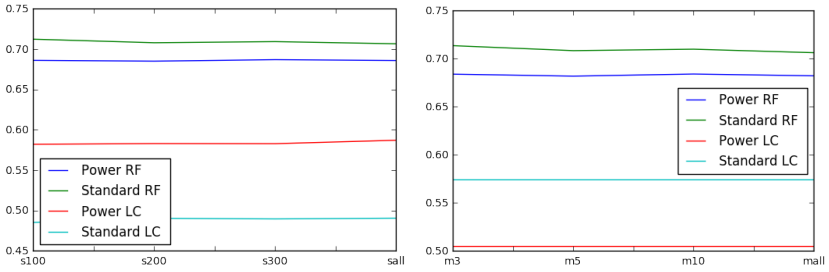

Fig. 22 Plots comparing the F1-score for each class (power and standard) obtained by two classifiers (a Logistic Classifier and the Random Forest classifier) when predicting the volume of edits that editors will make. The first plot shows the F1-score evaluation using 100, 200, 300 sessions of edit history per editor as training data, while the second plot shows the evaluation using 3, 5, 10 months of edits per editor as training data.

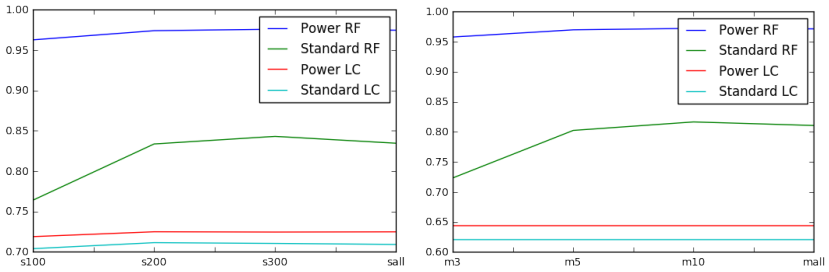

Fig. 23 Plot comparing the F1-score for each class (power and standard) obtained by two classifiers (a Logistic Classifier and the Random Forest classifier) when predicting the lifespan that editors will have, using 100, 200, 300 sessions (first plot) and 3, 5, 10 months (second plot) of edit history per editor as training data.

9 Discussion

In this section we discuss the findings highlighted in sections 6 and 7, and present some of the implications of our research results for the Wikidata community.

9.1 Summary of Findings

In summary, with this work we found that:

- There is a skewed distribution of edit counts.
- There is a skewed distribution of editors per item.
- There is a slightly skewed distribution of lifespan.
- There is not a linear relationship between the lifespan and the volume of edits done by editors.
- In Wikidata we find shorter times between edits than in Wikipedia.
- We empirically define new sessions after 4.37 hours of inter-edit time, around 4 times longer than in Wikipedia.
- Editors with long lifespan have a constant contribution over months, while editors with short lifespan do not.

- Editors with a high volume of edits have a constant participation over sessions, while editors with low level of volume of edits do not.
- Editors with a long lifespan tend to increase the diversity of type of their edits, while editors with short lifespan can either increase or decrease it over the months.

9.2 Interpretation of Findings

Participation inequality (Yasseri et al., 2012) is present in a vast amount of Web systems, where often only the 1% of users contribute heavily, 9% of users contribute sporadically and the remaining 90% are so-called lurkers, who consume information (reading and observing) but do not actively contribute. Hence, the skewed distribution of edit counts is an expected behaviour. Still, it is a relevant descriptive statistic that can be used to understand the order of magnitude of the set of editors whom should be addressed, and define parameters of the editor retention / reactivation strategy (e. g. an upper bound for the number of expected edits to be improved). The (also) skewed distribution of editors per items indicates that not many item descriptions are really crowdsourced, which in the case of Wikidata might be especially detrimental, considering that the knowledge base is meant to be the aggregate of what primary Web sources state. Features like qualifiers, ranks and references in statements allow Wikidata to portray existing plurality, and therefore, the more people involved in curating information about an item, the higher the chance to capture this plurality and the lower the risk of having certain kind of bias during the data collection.

The distribution of the lifespan shows that there a couple of thousands of people who have been contributing for almost the 4 years that we studied, which is priceless. Again, having this descriptive statistic can help configure the retention / reactivation strategy. Currently, Wikimedia encourages editors after they reached editing milestones (e. g. editors get congratulated via the wiki after they achieve their 100th edit). So, similarly, we could think of acknowledging people for being in the project for a particular amount of time. The fact that the relationship between lifespan and volume of edits is not linear indicates that there are some people who make a contribution of for example a thousand of edits in a short time – even less than one month. Probably events such as hackathons and editathons with a specific focus (e. g. to enter the description of women Swiss scientists in Wikidata) stimulate editing activity of participants, but in some cases it might mostly during the event. Moreover, data providers might also edit in bulk for a short time to ingest one single data set into Wikidata.

Having shorter times between edits than in Wikipedia is coherent with the nature of Wikidata, because it allows people to edit structured data – that being adding references, updating a date or a quantity, or creating a link between two items. In Wikipedia, people can also correct a typo, add a citation or a link, but the main task is to write a text (either a complete page or a paragraph), and the edit is registered when the text is saved, and not after each word has been edited. In Wikidata, if statements are edited within the same item, and the editor is proficient it is very much feasible that two edits are accomplished within one or few seconds. The high number of consecutive edits done in or under 5 seconds can be due to the fact that some actions, like merging items and deleting items (only granted to special editors) can be executed one after the other in a very short time. Furthermore, labels in various languages can be edited all at once. It can also be that some editors edit multiple items in parallel (e. g. in several tabs open at the same time) to perform similar kind of actions, like for example geolocate, or add descriptions. Another explanation for the high number of very short times between edits is that we are able to identify edits by tools that

leave a register trace (e. g. # petscan), but it is not possible to filter out API calls executed from tools that other developers might have implemented and did not tag or even advertise.

The results referring to the differences between power and standard editors suggest that power editors have habits in terms of their contributions – edits done each month – and the participation – the time spent in sessions. The existence of habits reflects not only the conviction to contribute, but it also shows that these editors successfully manage to find work they can do. From conversations we had with editors, we know that some power editors add information about the film they have watched, or edit information about the person they heard about in the news. Others look everyday through the list of recent and unpatrolled changes. There are different events that trigger editing action, and power editors follow them regularly. One could argue that if contribution and participation are both constant, then there is no clear signal of a learning effect that leads to editors becoming faster in accomplishing their tasks. We assume that the explanation is that power editors have a learning effect mostly in the beginning and later their efficiency becomes stable. Since they typically have many edits, it is possible that this initial learning effect is invisible to the linear model fitted. The increasing trend for the diversity of the types of edits present in power editors is aligned with the fact that intrinsically motivated people tend to "seek out novelty and challenges, to extend and exercise one's capacities, to explore, and to learn."(Ryan and Deci, 2000).

Between the two prediction problems that we set up (i. e. predicting the lifespan range and predicting the range of volume of edits), predicting the lifespan has a higher priority, because it gives us the key information about when we should address standard editors that will become inactive. Therefore, being able to predict lifespan more accurately than the volume of edits is a positive result.

9.3 Implications

Based on our findings, we identify three major areas where the Wikidata community could focus:

- Increasing the crowdsourcing ratio per item: in the same way that there are indicators to draw attention to the relative (in)completeness of items (see the ReCoin tool `https://www.wikidata.org/wiki/User:Ls1g/Recoin`), there could be indicators measuring the plurality of item descriptions and the number of editors involved in the description (as a crowdsourcing ratio). The goal of showing such information would be to request the help of further editors and improve the plurality of the item description – however that is defined.
- Acknowledging long lifespan and high activity periods: the retention literature conveys that it is better to interact with users before they leave because the probability of making them go back to the system once they drop-out are lower. Hence, we encourage the Wikidata community to add lifespan and high activity recognition to the acknowledgments that are implemented in the wiki. The system could congratulate editors for being in the system for a long time or for having a high activity peak. The message could report historic statistics or even show flashbacks about a significative edit done by the editor a while ago. As a reward, the community could grant these editors a special privilege in their wikidata-versary.
- Encouraging a behavioral change that makes standard editors become power editors: with Wikidata's increasing size and complexity, it is becoming more and more challenging for editors to master the variety of tools to edit, query and visualize Wikidata's data,

and find things they can contribute to. In our work, we found out that a very high rate of editors have a very short lifespan, and their evolution show that their contribution and participation is not constant. That is, there are many editors with the potential to become more active. Usually, early dropouts are motivated by any of the multiple possible phenomena such as people lacking the conviction for free knowledge, people not finding the way to contribute and Wikidata not being able to develop a sense of "addiction"[13]. Organizing training and dissemination events to educate people about the value of free knowledge is the solution to (i) and that is something that Wikimedia already does. For (ii) and (iii), Wikidata needs to provide a methodology and tools that allow these editors find valuable work to do and develop a reinforcing editing habit. Previous works have shown that suggesting things to edit in Wikipedia can be useful both for the users and the system (Cosley et al., 2007; Wulczyn et al., 2016), and providing article feedback can also help readers transition into editors (Halfaker et al., 2013). Given that the dynamics defined in the Wikipedia community is similar to the Wikidata community, we believe that there is space to develop a solution that can guide Wikidata editors and help them become more active and transition close to power editors (if they are willing to do so). We propose to design a system that helps standard editors find their editing mission. Missions could be defined individually or collectively (i. e. shared with other editors). To encourage a change in their behaviour, it would be important to consider behavioral change theories (Yasseri et al., 2012; Michie et al., 2011) suggest that change is more effective when the person frames intentions and goals, has the chance to self-regulate him/herself, and can freely select from available choices. So, as a design principle, the system would let them define what they would like to achieve, and decide what they finally would like to work on let, rather than impose or assign work to do. The system, still, would need to reduce the number of editing possibilities to a manageable and attractive set of options. And such an algorithm would need to exploit the main difference between Wikidata and Wikidata: the structure in the data and the fine granularity of traceable actions. Another key feature of such a solution could be a tight collaboration between power and standard users. Standard editors would define their intentions (e. g. editing for the city of Zurich) and identify themselves with roles (e. g. someone would like to become a quality ninja, but she does not know how yet). Power editors would set calls for actions and define data needs. The system could enable a 1:1 contact between power and standard editors, so as to share and disseminate best practices, recommendations on habits and know-how. Wikimedia currently enables mentorship [14]. Our system would aim at systematizing some actions in this process. The predictive models defined in this paper would be useful to estimate the amount interaction (and help) that editors would need, as well as to compute the time when the system interacts with the editor, whose behaviour we would like to improve (in terms of engagement). [15]

[13] Causes to drop out in Wikipedia by the community `https://www.wikizero.com/en/Wikipedia:WikiProject_Editor_Retention#`

[14] Wikipedia Mentorship `https://en.wikipedia.org/wiki/Wikipedia:Mentorship`

[15] This idea, together with the major findings of this research were presented in a talk at WikidataCon `https://goo.gl/vKH1kj`. The Wikidata community appreciated the findings and welcomed this proposal to improve editor attrition.

9.4 Limitations

Our work has two major limitations: *First*, we do not provide a qualitative analysis of the edits. While we distinguish among different types of edits (see indicator i5 7.2), we focused on contribution, participation and diversity of the edits without observing features of the items (e. g. topics and categories of the items), or the quality of the actual edits (e. g. whether the new statement created by an editor is semantically accurate). Clustering items by topics would be useful for computing further diversity indicators, whereas understanding the quality of the edits would help us categorize editors in different ways. Labeling the quality of edits can be beneficial for filtering out editors whom we may not want to retain – people who intentionally provide incorrect, hence, disruptive edits – and consequently can help us design more accurate measures against attrition. Yet, predicting the lifespan is a useful information by itself, because it gives a hint about the moment when we need to intervene to encourage behavioural change. Likewise, the prediction of the volume of edits gives an indication about the magnitude of the editor's work. The fact of knowing if they are primarily good or bad edits will only change the interpretation and the way we will proceed (e. g. in the case of malicious editors, the shorter the lifespan and the lower the volume of edits, the better; and in the case of helpful editors it is exactly the opposite)

Second: we do not reveal the reasons that lead to editor attrition. Understanding the issues that standard editors face is crucial for designing a solution to the attrition problem. That is why, we plan to run a survey in the Wikidata community, to address this question.

While these are natural extensions of our work, they are complex topics that are worth standalone articles. Automatically identifying the quality of edits, for example, is a highly challenging task (Sarabadani et al., 2017). In fact, the definition of data quality in Wikidata is still under debate among the research and volunteer communities [16]. Similarly, partitioning the knowledge base according to different topical domains could be done in many various ways and would require studying the application of specific techniques such as topic modelling in knowledge graphs. These two tasks are out of the scope of this study, which by design aims at understanding the differences in participation and dedication between power and standard editors, and predicting the group to which editors will belong in the future.

10 Conclusions and Future Work

In this paper we have performed a longitudinal cross-sectional analysis of Wikidata's editor community behaviors aiming at understanding how different types of editing behaviors lead or not to high volume and long-term engagement with the community.

Our results show that:

- The number of new editors joining the Wikidata community has been increasing over time, but it decreased recently.
- The distribution of contributions is very skewed with few editors contributing most of the edits and many editors performing just a few edits.
- To correctly define an edit session, it is necessary to consider the intra-session edits distribution (as well as intra-session differences and inter-session differences), obtaining a threshold of 4.3 hours, around 4 times longer than in Wikipedia.
- There is no linear relation between editors' lifespan and the volume of edits they provide.

[16] `https://www.wikidata.org/wiki/Wikidata:Requests_for_comment/Data_quality_framework_for_Wikidata`

- Power editors tend to show a constant contribution over the months and a constant participation over the sessions, while standard editors show instead an increasing or decreasing tendency.
- Power editors tend to increase the diversity of the types of edits, but this dimension alone is not clearly separating the two sets of editors, because some standard editors also increase their diversity over time.
- Despite the unbalanced nature of the data (i. e. few editors with many edits or long lifespan), it is possible to automatically predict the future the volume of edits and lifespan duration of an editor based on the available edit history of Wikidata editors. We are able to obtain better prediction results than a naive classifier in each of the 4 tested configurations. We are able to predict lifespan better than volume of edits (with an average F1 score above 0.9) in both session- and month-based evolution.

Our results are relevant to the Wikidata community, because they shed some light on the way power and standard editors in Wikidata evolve differently over time. Having these insights is useful –especially now that there is still limited knowledge about the way the community progresses– to design methods that encourage standard editors to contribute more, and hopefully also longer, as they experience progress. Additionally, our results and observations may be of use to other crowdsourced knowledge curation and maintenance projects with similar characteristics to better engage their communities of contributors.

While this work is leveraging a very large dataset of activity logs, in the future we plan to complement our work with a qualitative study performed by surveying and interviewing representative Wikidata editors sampled from the different categories we looked at in this work. As future work, we can also include the *quality of edits* as a feature which can be measured with the revert rate as a proxy. Furthermore, as a follow-up work, we plan to work on the implementation of the solution for attrition and retention management proposed in Section 9, in collaboration with Wikimedia.

11 Acknowledgments

We would like to thank Michele Catasta for his feedback at an early stage of this research, and the rest of the participants of our Dagstuhl Research Meeting “Crowdsourcing Research - Transcending Disciplinary Boundaries”. We also would like to thank Michael Luggen for his help to set up one of the machines used for the experiments of this project. This project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 732328, as well as from the COST Action IC1302 - Keystone.

During the manuscript reviewing process, several authors changed their affiliation. Part of the work presented in this paper was carried out while Cristina Sarasua was affiliated with the University of Koblenz-Landau (Germany) and visited the University of Sheffield (United Kingdom), Gianluca Demartini was affiliated with the University of Sheffield (United Kingdom) and Djellel Difallah was affiliated with the University of Fribourg (Switzerland).

References

Alvarez, Michael R. (2016). *Computational Social Science: Discovery and Prediction*, Analytical Methods for Social Research. Cambridge: Cambridge University Press.

Ang, Lawrence; and Francis Buttle (2006). Customer Retention Management Processes: A Quantitative Study. *European Journal of Marketing*, vol. 40, no. 1/2, pp. 83–99.

Clow, Doug (2013). MOOCs and the Funnel of Participation. *LAK '13. Third Conference on Learning Analytics and Knowledge*. New York: ACM, pp. 185–189.

Cosley, Dan; Dan Frankowski; Loren Terveen; and John Riedl (2007). SuggestBot: Using Intelligent Task Routing to Help People Find Work in Wikipedia. *IUI'07. Proceedings of the 12th International Conference on Intelligent User Interfaces*, IUI '07. New York: ACM, pp. 32–41.

Cox, David R. (1992). Regression models and life-tables. *Breakthroughs in statistics*. Springer, pp. 527–541.

Cuong, To Tu; and Claudia Müller-Birn (2016). SocInfo'16. Applicability of Sequence Analysis Methods in Analyzing Peer-Production Systems: A Case Study in Wikidata. *Social Informatics*. Berlin: Springer, pp. 142–156.

Danescu-Niculescu-Mizil, Cristian; Robert West; Dan Jurafsky; Jure Leskovec; and Christopher Potts (2013). No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities. *WWW 2013. 22nd International World Wide Web Conference, Rio de Janeiro, Brazil, May 13-17, 2013*. New York: ACM, pp. 307–318.

Difallah, Djellel; Michele Catasta; Gianluca Demartini; and Philippe Cudré-Mauroux (2014). Scaling-Up the Crowd: Micro-Task Pricing Schemes for Worker Retention and Latency Improvement. *HCOMP'14, Second AAAI Conference on Human Computation and Crowdsourcing,*. AAAI, pp. 50–58.

Dittus, Martin; Giovanni Quattrone; and Licia Capra (2016). Analysing Volunteer Engagement in Humanitarian Mapping: Building Contributor Communities at Large Scale. *CSCW '16. Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work Social Computing*. New York: ACM, pp. 108–118.

Druck, Gregory; Gerome Miklau; and Andrew Mccallum (2008). Learning to Predict the Quality of Contributions to Wikipedia. *WikiAI'08. Proceedings of the Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*. Palo Alto: AAAI Press, pp. 7–12.

Duhigg, Charles (2012). *The Power of Habit: Why We Do What We Do in Life and Business*, Vol. 34. Random House.

Fischler, Martin A; and Robert C Bolles (1981). Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, vol. 24, no. 6, pp. 381–395.

Franklin, Michael J.; Donald Kossmann; Tim Kraska; Sukriti Ramesh; and Reynold Xin (2011). CrowdDB: Answering Queries with Crowdsourcing. *SIGMOD 2011. Proceedings of the ACM SIGMOD International Conference on Management of Data, Athens, Greece, June 12-16, 2011*. New York: ACM, pp. 61–72.

Gandica, Yérali; Joäo Carvalho; and Fernando Sampaio dos Aidos (2015). Wikipedia editing dynamics. *Physical Review E*, vol. 91, no. 1, pp. 012824.

Geiger, Stuart R.; and Aaron Halfaker (2013). Using Edit Sessions to Measure Participation in Wikipedia. *CSCW 2013. Computer Supported Cooperative Work, San Antonio, TX, USA, February 23-27, 2013*. New York: ACM, pp. 861–870.

Gordini, Niccolo; and Valerio Veglio (2017). Customers Churn Prediction And Marketing Retention Strategies. An Application of Support Vector Machines Based On the Auc Parameter-Selection Technique In B2B E-Commerce Industry. *Industrial Marketing Management*, vol. 62 pp. 100–107.

Halfaker, Aaron; Aniket Kittur; and John Riedl (2011). Don't Bite the Newbies: How Reverts Affect the Quantity and Quality of Wikipedia Work. *Proceedings of the 7th Inter-*

national Symposium on Wikis and Open Collaboration, 2011, Mountain View, CA, USA, October 3-5, 2011. New York: ACM, pp. 163–172.

Halfaker, Aaron; Oliver Keyes; and Dario Taraborelli (2013). Making Peripheral Participation Legitimate: Reader Engagement Experiments in Wikipedia. *CSCW 2013. Computer Supported Cooperative Work, San Antonio, TX, USA, February 23-27, 2013*. New York: ACM, pp. 849–860.

Huang, Shih-Wen; and Wai-Tat Fu (2013). Don't Hide in the Crowd!: Increasing Social Transparency Between Peer Workers Improves Crowdsourcing Outcomes. *CHI '13. ACM SIGCHI Conference on Human Factors in Computing Systems, Paris, France, April 27 - May 2, 2013*. New York: ACM, pp. 621–630.

Iba, Takashi; Keiichi Nemoto; Bernd Peters; and Peter A. Gloor (2010). Analyzing the Creative Editing Behavior of Wikipedia Editors Through Dynamic Social Network Analysis. *Procedia - Social and Behavioral Sciences*, vol. 2, no. 4, pp. 6441 – 6456.

Lintott, Chris J; Kevin Schawinski; Anže Slosar; Kate Land; Steven Bamford; Daniel Thomas; Jordan M. Raddick; Robert C Nichol; Alex Szalay; Dan Andreescu; et al. (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, vol. 389, no. 3, pp. 1179–1189.

Müller-Birn, Claudia; Benjamin Karran; Janette Lehmann; and Markus Luczak-Rösch (2015). Peer-production System or Collaborative Ontology Engineering Effort: What is Wikidata? *OpenSym'15. Proceedings of the 11th International Symposium on Open Collaboration*. New York: ACM, pp. 20:1–20:10.

Michie, Susan; Maartje M van Stralen; and Robert West (2011). The Behaviour Change Wheel: A New Method for Characterising and Designing Behaviour Change Interventions. *Implementation Science*, vol. 6, no. 1, pp. 42.

Nov, Oded (2007). What Motivates Wikipedians? *Communications of the ACM*, vol. 50, no. 11, pp. 60–64.

Panciera, Katherine; Aaron Halfaker; and Loren Terveen (2009). Wikipedians Are Born, Not Made: A Study of Power Editors on Wikipedia. *Proceedings of the ACM 2009 International Conference on Supporting Group Work*. New York: ACM, pp. 51–60.

Piscopo, Alessandro; Christopher Phethean; and Elena Simperl (2016). Wikidatians are born: paths to full participation in a collaborative structured knowledge base. *HICSS 2017. 50th Hawaii International Conference on System Sciences, Hilton Waikoloa Village, Hawaii, USA, January 4-7, 2017*. AIS Electronic Library (AISeL), pp. 4354–4363.

Ponciano, Lesandro; and Francisco Brasileiro (2014). Finding Volunteers' Engagement Profiles in Human Computation for Citizen Science Projects. *Human Computation*, vol. 1, no. 2,.

Rosenberg, Larry J; and John A. Czepiel (1984). A Marketing Approach for Customer Retention. *Journal of Consumer Marketing*, vol. 1, no. 2, pp. 45–51.

Ryan, Richard M; and Edward L Deci (2000). Self-determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-being. *American Psychologist*, vol. 55, no. 1, pp. 68.

Sarabadani, Amir; Aaron Halfaker; and Dario Taraborelli (2017). Building automated vandalism detection tools for Wikidata. *WWW 2017. Proceedings of the 26th International Conference on World Wide Web Companion*. pp. 1647–1654.

Schmachtenberg, Max; Christian Bizer; and Heiko Paulheim (2014). Adoption of the Linked Data Best Practices in Different Topical Domains. *ISWC 2014, The Semantic Web - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*. Berlin: Springer, pp. 245–260.

Shannon, Claude Elwood (2001). A Mathematical Theory of Communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55.

Singer, Philipp; Denis Helic; Andreas Hotho; and Markus Strohmaier (2015). Hyptrails: A Bayesian Approach for Comparing Hypotheses about Human Trails on the Web. *WWW 2015. Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, May 18-22, 2015*. New York: ACM, pp. 1003–1013.

Stewart, Osamuyimen; David Lubensky; and Juan M. Huerta (2010). Crowdsourcing Participation Inequality: A SCOUT Model for the Enterprise Domain. *HCOMP'10. Proceedings of the ACM SIGKDD Workshop on Human Computation*. New York: ACM, pp. 30–33.

Strohmaier, Markus; and Claudia Wagner (2014). Computational Social Science for the World Wide Web. *IEEE Intelligent Systems*, vol. 29, no. 5, pp. 84–88.

Verhoef, Peter C. (2003). Understanding the Effect of Customer Relationship Management Efforts on Customer Retention and Customer Share Development. *Journal of Marketing*, vol. 67, no. 4, pp. 30–45.

Vrandečić, Denny; and Markus Krötzsch (2014). Wikidata: a Free Collaborative Knowledge Base. *Communications of the ACM*, vol. 57, no. 10, pp. 78–85.

Walk, Simon; Denis Helic; Florian Geigl; and Markus Strohmaier (2016). Activity Dynamics in Collaboration Networks. *ACM Transactions on the Web (TWEB)*, vol. 10, no. 2, pp. 11.

Walk, Simon; Philipp Singer; Lisette Espín Noboa; Tania Tudorache; Mark A. Musen; and Markus Strohmaier (2015). Understanding How Users Edit Ontologies: Comparing Hypotheses About Four Real-World Projects. *ISWC 2015. Proceedings of the 14th International Conference on The Semantic Web - ISWC 2015 - Volume 9366*. Springer-Verlag New York, Inc., pp. 551–568.

West, Robert; Ingmar Weber; and Carlos Castillo (2012). A Data-driven Sketch of Wikipedia Editors. *WWW 2012. Proceedings of the 21st World Wide Web Conference, Lyon, France, April 16-20, 2012 (Companion Volume)*. New York: ACM, pp. 631–632.

Wulczyn, Ellery; Robert West; Leila Zia; and Jure Leskovec (2016). Growing Wikipedia Across Languages via Recommendation. *WWW 2016. Proceedings of the 25th International Conference on World Wide Web, Montreal, Canada, April 11 - 15, 2016*. New York: ACM, pp. 975–985.

Yasseri, Taha; Robert Sumi; and János Kertész (2012). Circadian Patterns of Wikipedia Editorial Activity: A Demographic Analysis. *PLoS ONE*, vol. 7, no. 1, pp. 1–8.

Zaveri, Amrapali; Anisa Rula; Andrea Maurino; Ricardo Pietrobon; Jens Lehmann; and Sören Auer (2016). Quality assessment for linked open data: A survey. *Semantic Web Journal*, vol. 7, no. 1, pp. 63–93.