Lindy Effect and User Retention on Wikidata

Goran S. Milovanović
DataKolektiv, Owner/Data Scientist, WMDE

Jan Dittrich
UX Design and Research, WMDE

Martin Gerlach
Research Scientist, WMF Research

WIKI DATA CON
2021 - a sustainable future for Wikidata

How it started

State of the Wikidata Community Survey

Wikidata Community Survey 2021

Wikidata Activities

Since which year have you been active on Wikidata?

2017

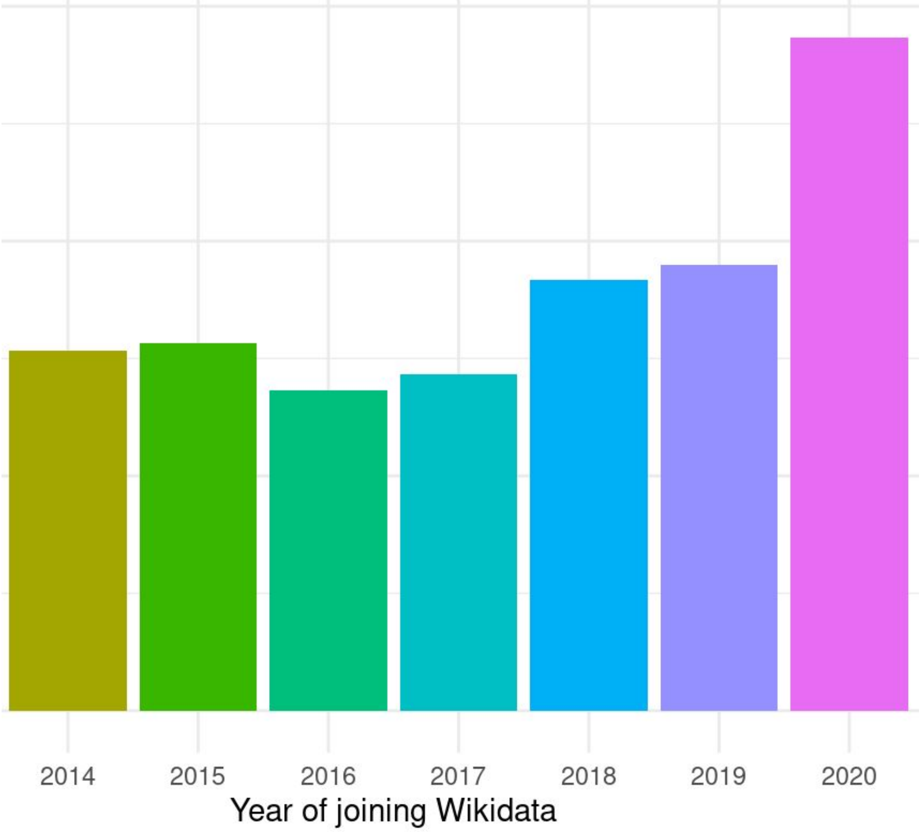
How it started

State of the Wikidata Community Survey

How it started

State of the Wikidata Community Survey

What about the account age?
For how long do editors stay with us to contribute?



Motivation

State of the Wikidata Community Survey

- **For how long to editors stay with us?**
- **How can we retain editors?**
- **How and when can we retain new editors?**
    - **Which kind of new editors leave us and when?**
- **Can we predict if and when an editor will leave the project?**

Related to ongoing research around Wikipedia Editor retention – but there is few research on retention of Wikidata editors!

Do we observe a Lindy effect?

“The Lindy effect (also known as Lindy's Law[1]) is a theorized phenomenon by which the future life expectancy of some non-perishable things, like a technology or an idea, is **proportional** to their current age.”
(Says Wikipedia)

Phabricator: User Retention Wikidata: A model for "participating since" patterns in the 2021 Wikidata Community Survey (T282563)

Do we observe a Lindy effect?

“The Lindy effect (also known as Lindy's Law[1]) is a theorized phenomenon by which the future life expectancy of some non-perishable things, like a technology or an idea, is **proportional** to their current age.” (Says Wikipedia)

This means:

retention time from now = *p* * retention time until now

Expected remaining account age

Account age at the time of observation (now)

Analysis

Variables

1. We have collected user revision histories for **399,967 users** - everyone who ever registered on Wikidata -from the [WMF Data Lake (wmf.mediawiki_history)](): from the beginning of time and until the most recent available data (Sep/Oct 2021).
2. Each **user revision history** is coded as a sequence of active and inactive (< 5 edits) months, e.g. 00111010101111110101010...).
3. Lengthy **0+$** sequences were edited to end in **00000$** e.g. 001110101011111100000~~000000000~~.
4. If a user revision history ends in five or more consecutive months of inactivity, we say that the **user has left Wikidata**.
5. Of course, users sometimes leave and then return: we count the number of user **reactivations**.
6. We count the number of active months in each user's revision history: **total user activity**.

Analysis

How likely are users to leave Wikidata?

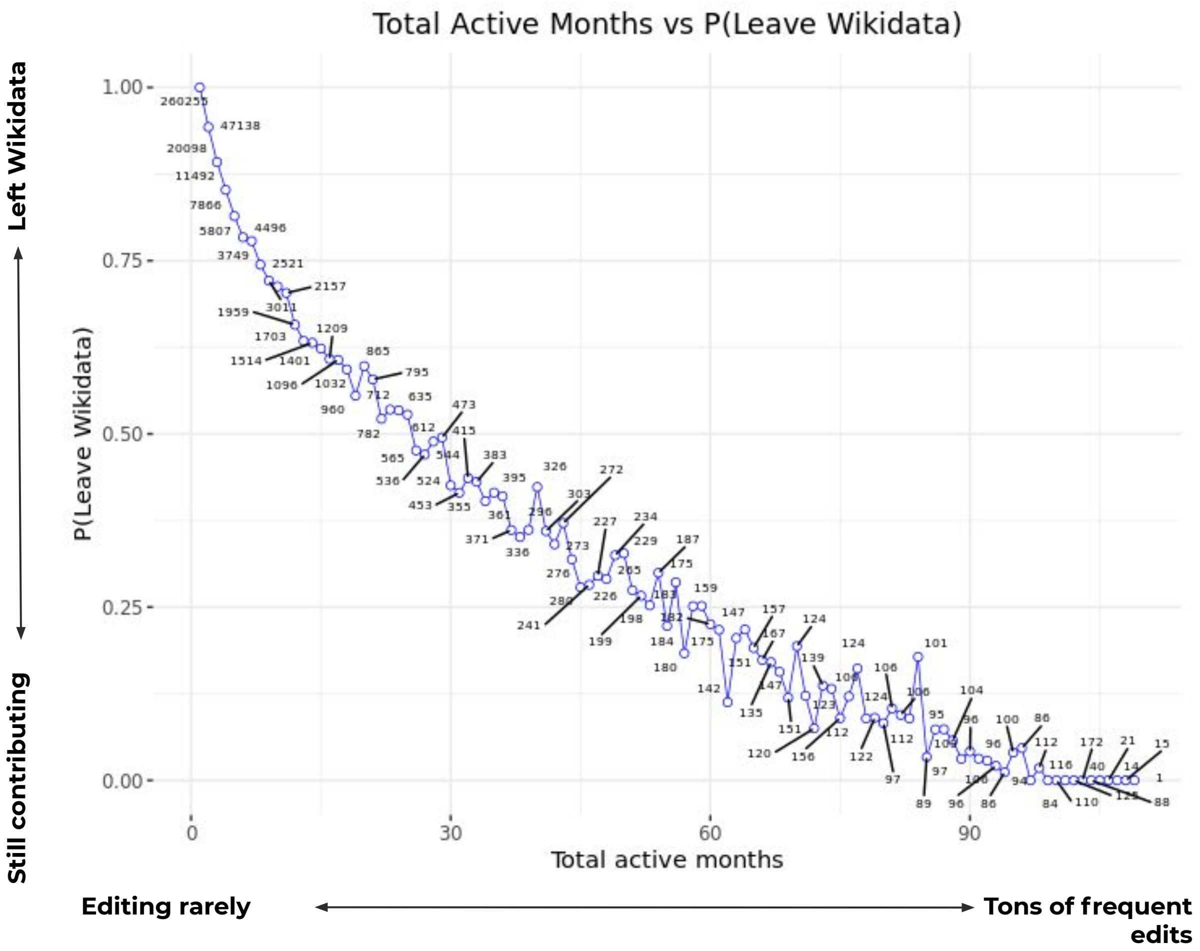
Users are very likely to stop editing: we’ve found that the probability to stop editing is **0.92**

Only **8%** 400K users who have ever registered in Wikidata were active at the time of analysis.

In the next slide we show that this probability of .92 heavily depends upon user revision history.

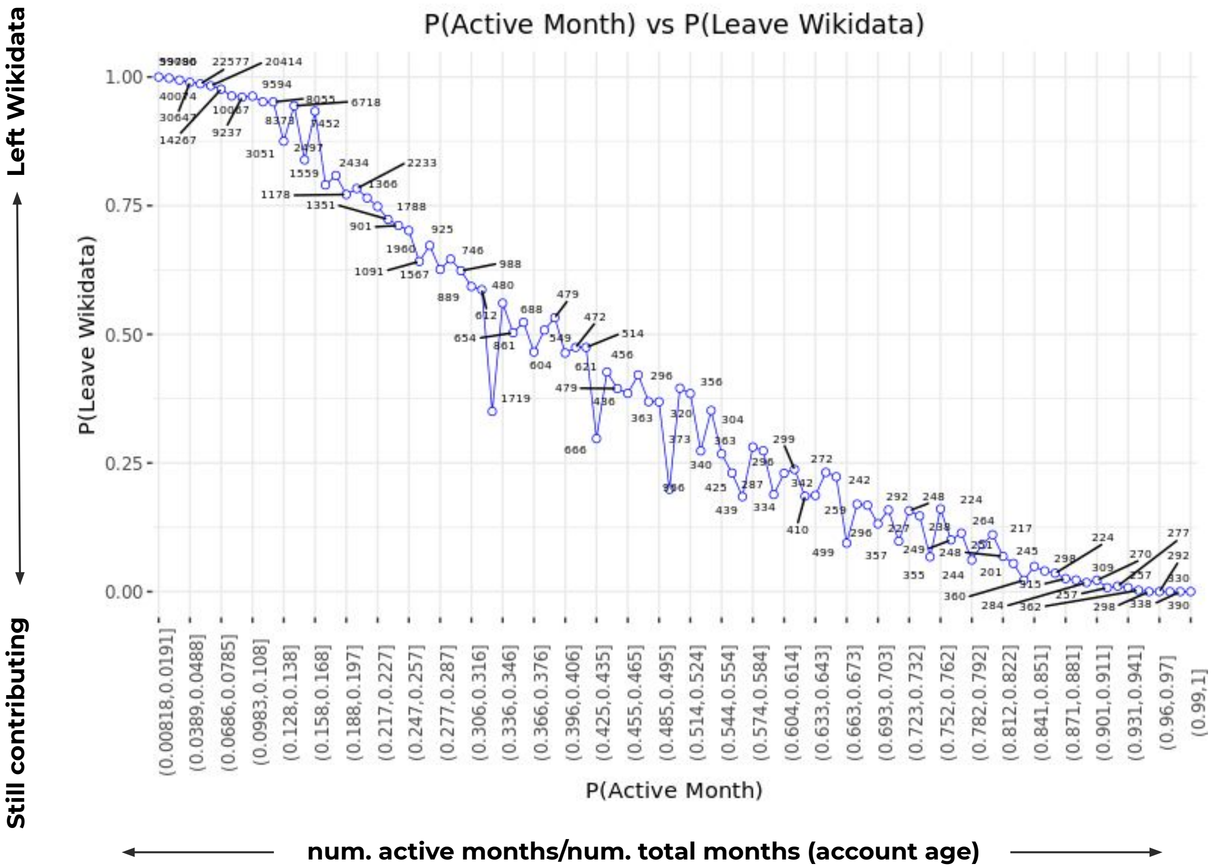
Analysis

How likely are users to leave Wikidata?



Analysis

How likely are users to leave Wikidata?



Analysis

Lindy Effect

Lindy Effect or not?

If the Lindy effect holds than the Survival function of the account age is Pareto.

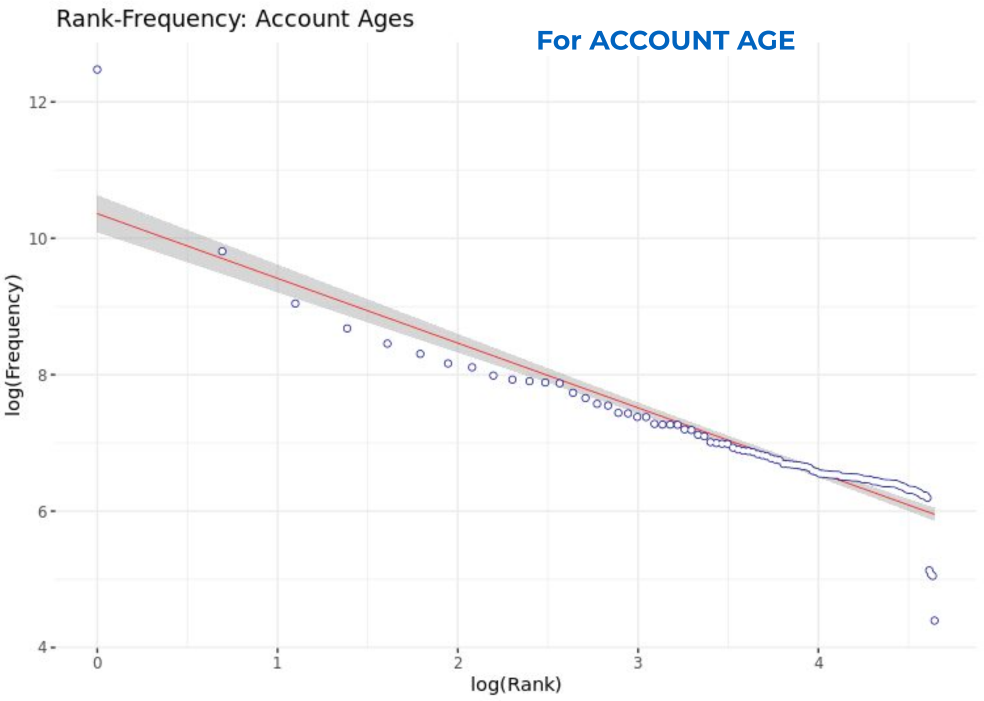
Eliazar, I. (2017). Lindy's Law. *Physica A: Statistical Mechanics and its Applications*. Vol. 486, 15 November 2017, pp. 797-805.

Hence we need to test if the Wikidata account age follows a power-law or not!

Analysis

Lindy Effect

Lindy Effect or not?



Frequency: how often do we observe an account age of N months?
Rank: we simply rank the frequencies, assigning ordinal numbers: 1, 2, 3,..
log(Rank) vs log(Frequency) plot is linear if a power law holds.

Analysis

Lindy Effect

Lindy Effect or not?

For ACCOUNT AGE

{poweRlaw} R package (based on: Clauset. A., Shalizi, C. R. & Newman, M. E. J. (2009). *Power-Law Distributions in Empirical Data. SIAM Rev.*, 51(4), pp. 661–703.)

Method A	Method B
Estimate x_{min}	Take empirical x_{min}

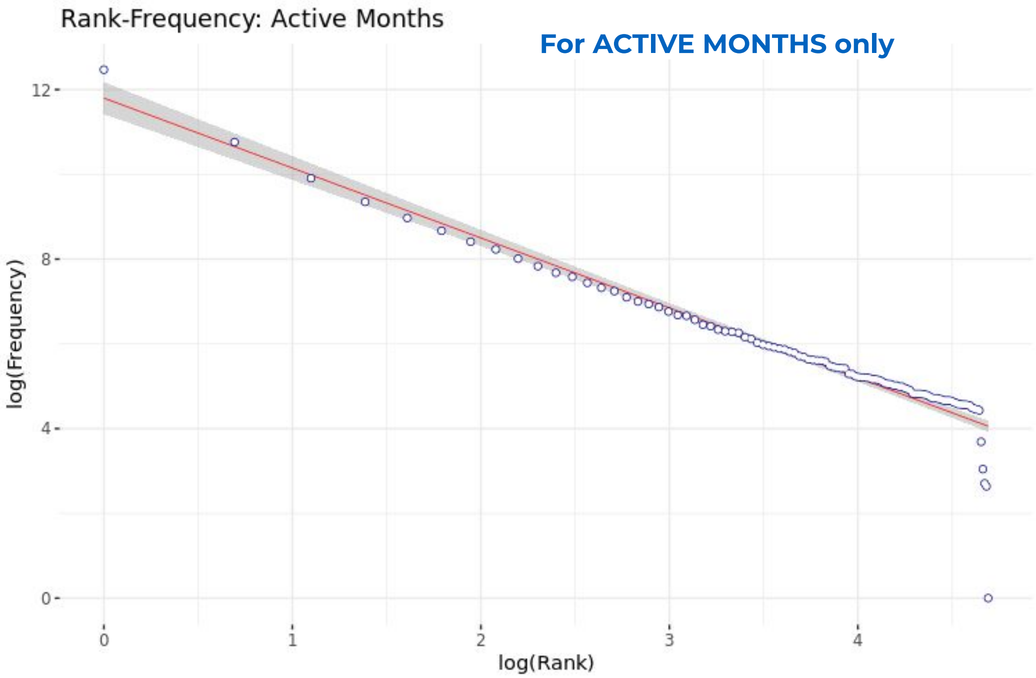
H_0: Power law
H_1: Not a power law

x_{min} = 7	**x_{min} = 6**
α = 1.78	**α = 2.81**
Bootstrap p (1000 sims) **p = 0**	**Bootstrap p (1000 sims)** **p = 0**
Result: NOT a power law	**Result: NOT a power law**

Analysis

Lindy Effect

Lindy Effect or not?



Frequency: how often do we observe an account with N active months months?
Rank: we simply rank the frequencies, assigning ordinal numbers: 1, 2, 3,..
log(Rank) vs log(Frequency) plot is linear if a power law holds.

Analysis

Lindy Effect

Lindy Effect or not?

For ACTIVE MONTHS only

Clauset. A., Shalizi, C. R. & Newman, M. E. J. (2009). *Power-Law Distributions in Empirical Data*. *SIAM Rev.*, 51(4), pp. 661–703.

Method A	Method B
Estimate x_{min}	Take empirical x_{min}

H_0: Power law
H_1: Not a power law

x_{min} = 2	**x_{min} = 1**
α = 1.82	**α = 1.98**
Bootstrap p (1000 sims) **p = 0**	**Bootstrap p (1000 sims)** **p = 0**
Result: NOT a power law	**Result: NOT a power law**

Analysis

Lindy Effect

Lindy Effect or not?

Well, no.

But this **does not** mean that

- people who are editing since a long time are **not** less likely to leave, or that
- newcomers do **not** have the highest probability to drop out.

This means **only** that a particular scaling is not present:

$$\mathrm{E}[T - t | T > t] = p \cdot t$$

Expected remaining account age ***T* - *t***...

... given that the user is already around for more than ***t*** months.

(with **T** following a power-law distribution with **a = 1 + 1/p**)

Analysis

User Retention

User Retention

Can we predict if a user will "stay" or "leave", given their user revision histories and additional features?

XGBoost for Binary Classification
FEATURES

- Number of user reactivations
- H - The Shannon Diversity Index (i.e. the entropy present in the distribution of active/inactive months, scaled)
- Account age (months)
- Number of active months (months)
- P(Active Month)
- Number of revisions - content NSs
- Number of revisions - talks NSs
- Average number of revisions per month - content NS
- Average number of revisions per month - talk NSs
- Mean length of inactivity periods (months)
- Median length of inactivity periods (months)

Analysis

User Retention

User Retention

XGBoost for Binary Classification
TRAINING

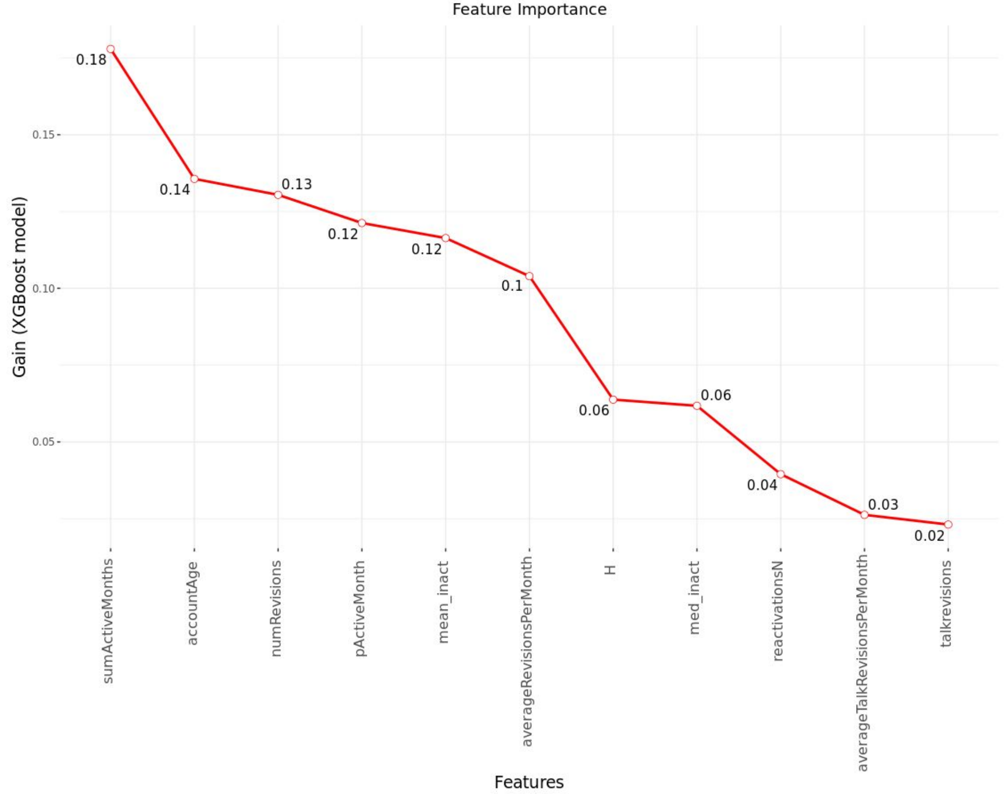
- Split into train and test dataset
- Heavy downsampling and upweight in the training dataset because of huge class imbalance
- Search through a very constrained parameter space, cross-validation across eta, max_depth and subsample only
- Shallow trees (max_depth = 5 or 10), lots of them (n_rounds = 10,000)

Analysis

User Retention

User Retention

XGBoost for Binary Classification
FEATURE IMPORTANCE



Analysis

User Retention

User Retention

Can we predict if a user will "stay" or "leave", given their user revision histories and additional features?

XGBoost for Binary Classification
RESULTS

- **TPR** = 0.91, **FNR** = 0.09
- **FPR** = 0.10, **TNR** = 0.90

- **AUC** = 0.969 - a bit better than currently the best model of Wikidata user retention that was reported in the literature (c.f. the DeepFM approach - but the authors used a different "leave" criterion then we did, so the models are not really comparable)

- **Bayesian analysis**, starting from P(User Leaves) = .92, shows that **P**(User leaves|Model says user will leave) = **.99**

Code/Data

GitHub

https://github.com/wikimedia/analytics-wmde-WD-WikidataAdHocAnalytics/tree/master/WD_UserRetention

- **01_WD_userRetention_ETL.R** - extract raw data from the wmf.mediawiki_history from Hadoop (denormalized revision history, all WMF projects); available (dumps) from WMF Analytics Datasets
- **02_WD_userRetention_Analytics.R** - data pre-processing, visualizations, power-law hypothesis testing, XGBoost model

Data (in the repo)

- **WD_UserRetention.csv** - raw dataset used to extract user revision histories; user IDs are anonymized
- **WD_UserRetention_TalkRevisions.csv** - raw dataset, revisions in the talk namespaces; user IDs are anonymized and matched with WD_UserRetention.csv
- These two datasets are used in 02_WD_userRetention_Analytics.R to produce all analytics datasets used in the study

Questions

WIKI DATA CON
2021 - a sustainable future for Wikidata