D3.1 Knowledge extraction models

Project ref. no.	957185
Project title	Möbius: The power of prosumers in publishing
Project duration	1st March 2021 – 30th of March 2023 (36 months)
Website	www.mobius-project.eu
Related WP/Task	WP3 / T3.1
Dissemination level	Public
Document due date	28/02/2022 (M12)
Actual delivery date	
Deliverable leader	EUT
Document status	**Draft** / Released / Validated / Submitted

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 957185

Revision History

Version	Date	Author	Document history/approvals
0.1	01/02/2022	Mihnea Tufis, Cristian Consonni, Julià Vicens, David Laniado (Eurecat); Constantin Popa (University of Trento)	Draft version
0.2	23/02/2022	Mihnea Tufis, Cristian Consonni, Julià Vicens, David Laniado (Eurecat); Constantin Popa (University of Trento)	Complete version for internal review
0.3	25/02/2022	Alexandru Stan (IN2)	Review
0.4	01/03/2022	Gino Querini, Enrico Turrin (FEP)	Review
0.5	02/03/2022	Thomas Van Dam (IMEC)	Review
0.6	04/03/2022	Mihnea Tufis, Cristian Consonni, Julià Vicens, David Laniado (Eurecat); Constantin Popa (University of Trento)	Final version

Executive Summary

In this document, we present the metrics, methods, and results of the analysis of data from existing prosumer communities performed in T3.1, that set the bases for the development of the Prosumer Intelligence Toolkit.

Research questions were defined after conversation with project partners and stakeholders, and interviews and focus groups with publishers, organized in the framework of WP2. For addressing these research questions, we collected and analyzed digital traces of user interactions among thousands of users from three popular prosumer platforms in which users co-create content: a fanfiction community where users create works based on existing fictional universes, and review one another's work (AO3); a fandom wiki, where users create content collaborating on editing wiki pages to document any element related to a fictional universe (Fandom); and a social reading platform, where users create original stories and read and comment on one another's work (Wattpad).

The central part of the document includes:

- An analysis of popularity dynamics in AO3, and a model to predict works that will become very popular in the near future, based on previous history. We have further re-adapted the model to apply it on tags, and predict topics that will become popular. Both with work and tag popularity prediction, we obtained a satisfactory accuracy with a simple and interpretable model. This responds to an emerging need of publishers and stakeholders to identify trending content and topics, and may help to understand trends in the interests of readers and writers, and to identify valuable content to be considered for publishing.
- An analysis of social interactions between AO3 users, modelled as graphs, studying structural properties of the social networks resulting from different kinds of interactions, and the centrality of the users in the community. We characterized each user by the combination of their centrality as a producer (feedback received an author) and as an active consumer (feedback given as a reader) in the network, finding consistent clusters across the major communities, that suggest the existence of different emerging profiles, among which the ones we dubbed *superproducers*, *superconsumers*, and *superprosumers,* as the users who have the highest levels of centrality in one of the two dimensions, or in both. We believe this "map" of prosumer roles may be helpful to understand the composition of a community and identify relevant users for specific aims.
- An analysis of collective dynamics in Fandom wikis, studying how activity on different tasks and spaces evolves over time and in different phases of community growth, for different communities, with an investigation on peaks of activity and their nature in the life of a community. The most edited pages result to be related to the main characters of each fictional universe. The amount of activity devoted to parallel spaces beyond

editing the main content of the wiki (e.g. coordination, communication, technical aspects) varies substantially across communities; as a general tendency, the proportion of the effort spent for personal communication and interactions increases during the periods of higher activity in a community.

- An analysis of the dynamics, and the language and emotion of users' feedback on a sample of very popular books from two diverse Wattpad categories: teen literature and classics. We found a tendency to have more activity on the first and last chapters of a book, and we have shown how language in the feedback around each book, and across books, can be characterized and compared through different tools for language and emotional analysis.

Finally, the document presents lines for future research and some preliminary ideas for the development of interactive dashboards in the Prosumer Intelligence Toolkit, to be developed in task T3.2.

This document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained herein. All logos, trademarks, imagines, and brand names used herein are the property of their respective owners. Images used are for illustration purposes only.

Table of contents

List of Figures

List of tables

Terminology and Acronyms

AO3	*Archives of Our Own*
EXP	*Exponential distribution*
LGN	*Lognormal distribution*
PDF	*Probability density function*
PL	*Power Law distribution*
TPL	*Truncated Power Law distribution*
HCI	*Human-Computer Interaction*

1. Introduction

One of the aims of the Möbius project is to guide the publishing sector in dealing with the emerging prosumer paradigm. To this end, task T3.1 is devoted to developing scalable digital methods for examining and extracting actionable knowledge relevant for the publishing industry from open prosumer communities of writers, readers, and fans.

The term “prosumer”, coined in 1980 by American futurist Alvin Toffler, became increasingly popular to describe a new parading in which users are not just passive consumers, but are "individuals who consume and produce value, either for self-consumption or consumption by others, and can receive implicit or explicit incentives from organizations involved in the exchange" (Lang et al, 2020). In this new paradigm, traditional customer intelligence methods, that consider users as passive consumers who will buy or not buy the product and try to analyse their behaviour and preferences in order to maximize sales, are unable to capture and boost the value created by the users as producers.

Such creative potential for the publishing sector can be exemplified with cases like “Fifty Shades of Grey” by author E.L. James, that got to sell over 100 million copies worldwide, and was initially a fanfiction creation based on the successful Twilight series, or the Netflix original film «The Kissing Booth», based on a 2011 story published in the social reading platform Wattpad by a 15-years old user, that was read by 19 million people on Wattpad before it was turned into a series of books.

In this framework, the project aims to explore ways to untap the potential still largely hidden in this kind of prosumer communities, as well as to identify which analysis methods, metrics, techniques, and tools can be useful in this setting.

While a big effort has already been made in previous research to develop methods and algorithms for mining social media and online platforms, we believe the particularities of the context of fanfiction and prosumer communities make it unique, so that it needs to be addressed by a specific modelling endeavour.

In task T3.1 we aim to develop methods and metrics to extract actionable knowledge from prosumer communities in the publishing sector, that will be the basis for the development of a Prosumer Intelligence Toolkit in T3.2, aimed to effectively streamlining cooperation with prosumer communities the publishing workflows.

We decided to focus on existing successful communities, where thousands of users are already co-creating content, and at the same time generating rich records of interactions that can help study their individual and collective behaviour on these platforms. Accessing and analysing data from such platforms, we can extract knowledge and develop methods that can be applied in other settings were prosumers are involved. In particular, we can generate relevant information for our stakeholders by modelling community dynamics, studying and measuring different aspects of their interactions, identifying relevant content and users, and performing predictions.

1.1 Research questions

In order to define the scope and objectives of this work and align them with the needs of the end-users of the Prosumer Intelligence Toolkit that will be developed in T3.2, we had several meetings with Möbius partners and stakeholders, including publishers who participated in interviews and focus groups organized within the work in WP2.

Based on this input, we collected different needs and formulated directions and research questions for our work, organized in 4 groups:

- Popularity dynamics:
  - Which are the popularity dynamics of content in these platforms?
  - How can we predict which items will become more popular?
  - How can we predict which topics will become more popular?

- Social dynamics:
  - How can we identify the most relevant prosumers?
  - How can we identify controversial content?
  - How can we characterize different kinds of prosumers?

- Community dynamics:
  - How do successful communities grow, and organize their work in different kinds of activity over time?
  - Are there peaks of activity, and how can we characterize them?

- Emotions and language:
  - How can we characterize language features and emotional content of the community feedback on some work?
  - How can we compare the emotional content of the feedback received by different works?

Some needs from the stakeholders could not be addressed in our work, for different reasons:

- Some requests focused on demographic information, which we do not have access to on any of the platforms available for the analysis. Therefore, although these are relevant aspects to be studied, we had to discard them in the scope of this analysis. However, we foresee how these aspects could be integrated within the current analysis, should demographic data on the users of any of these platforms become available at some point.

- Some requests focused on aspects such as the click and reading patterns of users, information that is only available to the companies or organizations who run the platforms. On the other hand, this kind of analysis of users'/customers' behaviour as mere consumers of information is already widely performed in digital platforms, while here we aim to focus on the creative potential of prosumers, taking an innovative perspective that builds on accessible digital traces to study co-creation patterns in these communities.
- Some requests focused on understanding the whole activity of a user across a platform, looking at whether they are interested in a specific topic/genre/fictional universe, or more, and finding patterns and preferences. This was especially related to understanding the "consuming" interests and preferences of users in a traditional way, like it is widely done with recommender systems. The main issue with this approach is that this would require to have access to the whole record of activity of a user on the platform, at least for a consistent set of users. This is in contrast with the way in which we retrieved the data, i.e., focusing on selected communities/fictional universes and delimiting in this way the data collection. So, for each user we can only observe their activity within the scope of the communities for which we have data. This limitation is in most cases imposed by the way data are organized and retrieved; only in the case of AO3 it would have been possible to follow a user-centric approach to data scraping, but this would have presented two kinds of issues: on one hand, it would not have allowed to get the full picture on a given community, which was instead our priority, in line with the approach already defined in the project's grant agreement; on the other hand, it would have implications for user privacy as the focus would be on individual users, making it more problematic to make results publicly available. For example, showing the activity of individual users in dashboards would raise privacy issues, while focus on content and communities makes it easier to treat and show individual user' data only at anonymized or aggregated level.
- Some requests focused on activity on social media, as an emerging. While this would undoubtedly add a layer of comprehension of popularity dynamics, that often are not limited to a single platform, again it is outside the scope of the project, especially due to the difficulty of retrieving data from relevant social media platforms mentioned by the partners and stakeholder, like TikTok or Instagram, that have very strict policies and barriers to access their data.

Although relevant to stakeholders (as indicated for example by publishers), we had to exclude these types of input, that unfortunately are out of scope for our research, mainly for the lack of data that would allow us to carry out the corresponding analysis, and we delimited the scope of this work on the research questions listed above. We hope to be able to integrate some of these directions in our future work and the topics (at least some) are highlighted in general as deserving further analysis.

1.2 Platform selection

In the project's grant agreement, we had foreseen to focus our analysis on three platforms, one fandom wiki platform, Fandom.com, and two fanfiction communities, Archives of Our Own (AO3), and Fanfiction.net. However, we could not retrieve data from the third platform, due to the terms of service that do not allow for automatic scraping; we wrote several messages to ask for permission to retrieve data in the frame of this study, but we did not receive any answer, so we had to discard this option. Instead, we decided to focus on the popular social reading platform Wattpad, for which we managed to obtain a dataset from a recent study (Pianzola et al, 2021).

This way, we believe we are able to provide a broader view on prosumers in relation to the publishing sector, covering a more diverse set of platforms:

- **Archives of Our Own (AO3)**[1] we focus on a large and very active fanfiction platform, where users can publish works and review each other's works; it is similar in many aspects to fanfiction.net, so we assume the results could be in large part extrapolated to the context of fanfiction.net.
- **Fandom.com**[2] allows us to look at the process of prosumers creating a kind of encyclopaedia around a fictional universe, documenting characters, plots, episodes, places, etc.; this resembles creating a kind of Wikipedia for each fictional universe. Indeed, the wiki engine and interface are the same as in Wikipedia (MediaWiki). However, the context is quite different, and so may be the social dynamics; while there are plenty of studies that analyse many different aspects of collaboration in Wikipedia, much less effort has been spent to study the communities that edit Fandom wikis.
- **Wattpad**[3] is a social reading platform, where users can write original stories, read other users' stories, and comment or discuss on specific paragraphs.

As it can be seen, these three scenarios are quite different in several aspects.

First, the context and aim of these communities is different: while on AO3 and Wattpad users develop new stories, on Fandom users collect knowledge about existing fictions; while AO3

[1] https://archiveofourown.org/

[2] https://www.fandom.com/

[3] https://www.wattpad.com/

and Fandom are communities of fans, and rely on existing fictional universes, on Wattpad authors mostly create new original stories.

Interface, and interaction possibilities also differ across these platforms: Fandom is based on wiki technology, where users edit common artifacts, wiki pages, and all actions are recorded in the edit history; AO3 is based on an open source infrastructure, managed by the community itself, and presents an old forum-like interface allowing different kinds of interactions between users; Wattpad is run by a company, and has a more modern interface that allows users to comment on specific paragraphs of the text.

Finally, also the data that can be retrieved from these platforms differ: in Fandom we have access to the whole edit history for each page from the selected communities; in AO3 we have bookmarks and comments left by each user, with the thread structure of the discussion resulting from the comments and replies around each work; in Wattpad we have the text of the comments and the paragraph to which it is referred, but not their indentation structure.

1.3 Roadmap

Therefore, given all the specificities of these three different platforms, both from a social, technical, and data perspective, we decided to run different specific analyses on each of them, taking advantage of their particularities, starting from the research questions formulated above, and focusing on the aspects that we considered more relevant for the communities from each platform.

In particular, we took advantage of the availability of fine-grained social interaction data for a large amount of works over a long period of time from AO3, to focus for this platform especially on the first two sets of research questions, centred on popularity dynamics and social dynamics.

On Fandom, as discussed above, we do not have original works created by users, and the wiki paradigm does not allow a piece of content to be associated with an individual author, but to the whole community (or at least to all the users who edited that wiki page), therefore we decided to take advantage of this platform to investigate mainly questions from the third set, related to community dynamics over time.

Finally, on Wattpad we do not have the same richness of information as for AO3 on social interactions, while we have access to the text of a large number of comments for selected popular books; we decided to focus on the fourth set of research questions, and characterizing textual features, emotions and sentiments in the feedback on different books.

1.4 Document structure

This document is structured in three main sections, each devoted to the analysis performed on one of the three selected platforms.

In Section 2 we will present the analysis of data from AO3, for which on one hand we will deepen into the investigation of popularity dynamics and present a model to predict the popularity of content and topics, and on the other hand we will model social interactions through social network analysis and characterize different profiles of prosumers based on their behaviour as "producers" and as "active consumers" of content.

Then, in Section 3, we present the analysis for Fandom, where we study community dynamics over time and focus on looking at different kinds of activity over time (i.e., activity on different kinds of pages), and at the temporal evolution of activity identifying peaks of edits, and specific pages concerned by these spikes of activity.

In Section 4 we present the analysis of data from the Wattpad social reading platform, focusing on the distribution of comments along books and chapters, and on the comparison of emotions in the comments on different books.

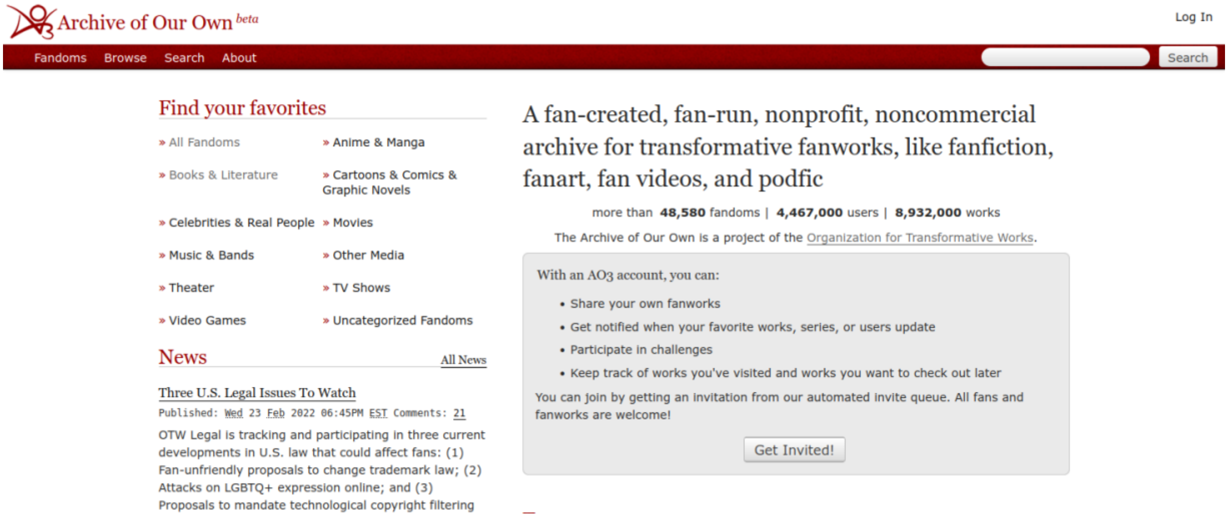
Finally, in Section 5 we present conclusions and discuss the next steps and lines for future work.

2. Analysis of a fanfiction community: Archives of Our Own (AO3)

2.1 Platform description

The Archive of Our Own (AO3) self-defines as "a non-commercial and non-profit central hosting site for transformative fan works such as fanfiction and, in the future, other transformative works such as fanart, fan videos, and podfic. The AO3 is built on open-source archiving software designed and built by and for fans".[4]

The web site is one of the main references worldwide for fanfiction work, and as reported in its home page it currently hosts almost 9 million works created by the users, organized in over 40 thousand fandoms, with over 4 million registered users.



Find your favorites

» All Fandoms
» Anime & Manga
» Books & Literature
» Cartoons & Comics & Graphic Novels
» Celebrities & Real People
» Movies
» Music & Bands
» Other Media
» Theater
» TV Shows
» Video Games
» Uncategorized Fandoms

News All News

Three U.S. Legal Issues To Watch

Published: Wed 23 Feb 2022 06:45PM EST Comments: 21

OTW Legal is tracking and participating in three current developments in U.S. law that could affect fans: (1) Fan-unfriendly proposals to change trademark law; (2) Attacks on LGBTQ+ expression online; and (3) Proposals to mandate technological copyright filtering

A fan-created, fan-run, nonprofit, noncommercial archive for transformative fanworks, like fanfiction, fanart, fan videos, and podfic

more than **48,580** fandoms | **4,467,000** users | **8,932,000** works

The Archive of Our Own is a project of the Organization for Transformative Works.

With an AO3 account, you can:

- Share your own fanworks
- Get notified when your favorite works, series, or users update
- Participate in challenges
- Keep track of works you've visited and works you want to check out later

You can join by getting an invitation from our automated invite queue. All fans and fanworks are welcome!

Figure 1 - Screenshot from AO3's home page.

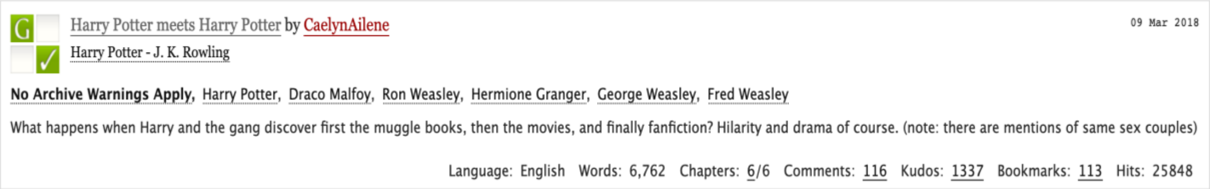
[4] https://archiveofourown.org/faq/about-the-archive?language_id=en#whatisao3. Retrieved, 22/02/2022.

The community has been studied as a successful example of gift culture (Riley, 2015), self-organization and self-governance (Fiesler, 2018), and feminist HCI, where the platform was designed and coded primarily by women to meet the needs of the online fandom community, and design decisions were informed by existing values and norms around issues such as accessibility, inclusivity, and identity (Fiesler et al, 2016). The rich and open tagging system, giving place to a folksonomy of keywords and categories emerging from users' behaviour, has also been object of several studies (Dalton, 2012; Price & Robinson, 2021).

2.1.1 User actions

The foundation for any metrics and models for content popularity are the actions undertaken by the users of the platforms.

The three types of user actions in AO3 are *comments, kudos,* and *bookmarks*.



Harry Potter meets Harry Potter by CaelynAilene 09 Mar 2018

Harry Potter - J. K. Rowling

No Archive Warnings Apply, Harry Potter, Draco Malfoy, Ron Weasley, Hermione Granger, George Weasley, Fred Weasley

What happens when Harry and the gang discover first the muggle books, then the movies, and finally fanfiction? Hilarity and drama of course. (note: there are mentions of same sex couples)

Language: English Words: 6,762 Chapters: 6/6 Comments: 116 Kudos: 1337 Bookmarks: 113 Hits: 25848

Figure 2 - User actions and interactions with AO3 content

Comments

Comments are a method of leaving feedback on a work. Each work, or each chapter of a work, has a textbox at the end where users can input their remarks. One can also respond to comments other people have left, which will form a comment thread. Comments are located at the bottom of a work's page, beneath the list of kudos. Note that a work can have multiple pages of comments; each page will show 20 comment threads. Comment threads are like conversations: one person leaves a comment, and all subsequent replies to that comment are part of one comment thread.

By default, one can comment anonymously on any work in the Archive; however, the work creator can also disable this function if they choose so. When a user deletes their account, any comments they had left with that account will be attributed to the username "Account Deleted." When a user changes username, any comments and/or kudos left under the old username will update to the new username.

Kudos

The word "kudos" is from ancient Greek, meaning "glory" or "renown". One modern definition is "praise given for achievement". As an Archive feature, kudos are a quick and easy way to let a creator know that one likes their work. A user can only give kudos to a work once, even if they have multiple pseudonyms under their account. When leaving kudos while accessing a particular chapter of a work, these are linked to the work as a whole. Kudos cannot be deleted at this time, neither by the user who gave it, nor by the one who received it. All kudos originating from a deleted account are attributed to a guest, instead of the original username.

Bookmarks

An AO3 bookmark is a way for a user to mark a work that they wish to remember, retrieve more easily, or make a note about. Bookmarks created on the Archive can also serve as recommendations to other users. Any work or series posted on the Archive can be bookmarked. One can also add personal notes or tags to the work. The notes can be anything, such as brief reviews or self-reminders. One can also mark a bookmark as a Rec (Recommendation), in which case it will be included in search results when a user searches for Recs. When viewing bookmarks, any that have been marked as a Rec will display with a heart icon on the blurb. Unlike kudos, bookmarks can be edited and/or removed.

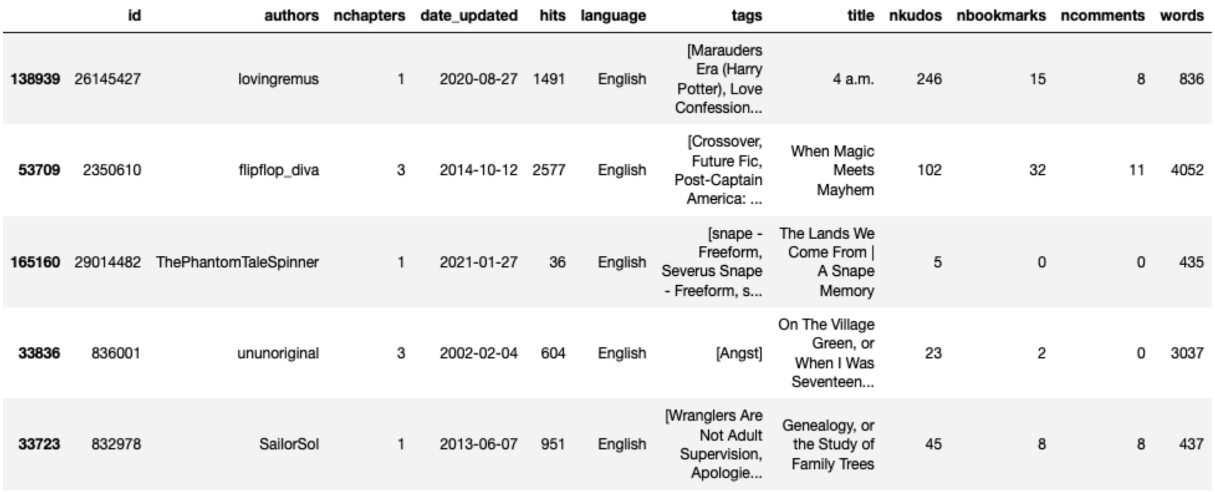
2.2 Basic dataset statistics

Works

Figure 3 illustrates a sample of work records from the data set scraped form AO3 for the purpose of this analysis. It contains the following fields for each work:

- *id* – the unique ID of the record, as it is stored in our database
- *authors* – the author of a chapter within a work
- *nchapters* – number of chapters by an author within a work
- *date_updated* – the date the last chapter was updated
- *hits* – the number of clicks a chapter has received
- *language*
- *tags* – a list of tags attached to a chapter
- *title* – the title of the chapter
- *nkudos* – number of kudos received by a chapter
- *nbookmarks* – number of times a chapter was bookmarked
- ncomments – number of comments a chapter has received

- *words* – the number of words in the chapter



	id	authors	nchapters	date_updated	hits	language	tags	title	nkudos	nbookmarks	ncomments	words
138939	26145427	lovingremus	1	2020-08-27	1491	English	[Marauders Era (Harry Potter), Love Confession...	4 a.m.	246	15	8	836
53709	2350610	flipflop_diva	3	2014-10-12	2577	English	[Crossover, Future Fic, Post-Captain America: ...	When Magic Meets Mayhem	102	32	11	4052
165160	29014482	ThePhantomTaleSpinner	1	2021-01-27	36	English	[snape - Freeform, Severus Snape - Freeform, s...	The Lands We Come From | A Snape Memory	5	0	0	435
33836	836001	ununoriginal	3	2002-02-04	604	English	[Angst]	On The Village Green, or When I Was Seventeen...	23	2	0	3037
33723	832978	SailorSol	1	2013-06-07	951	English	[Wranglers Are Not Adult Supervision, Apologie...	Genealogy, or the Study of Family Trees	45	8	8	437

Figure 3 - Sample from the AO3 works dataset

	Marvel	Harry Potter	Sherlock Holmes	Lord of the Rings	Percy Jackson	Twilight	Warriors
Records	444804	280310	124337	24234	19240	10974	2763
Authors	83915	66609	24396	6871	7543	4913	1421
Titles	356240	228320	105771	22449	18338	10359	2728
Languages	45	49	34	31	19	21	10

Table 1 - Basic characteristics of AO3 fandoms

Table 1 presents the basic statistics on the communities in the AO3 dataset, showing the size of each community in terms of records, authors, titles, and languages, while Table 2 summarises the main properties (total, mean standard deviation and median) for each of the main features (chapters, kudos, bookmarks, comments, words) of all the works in all the fandoms.

		Marvel	Harry Potter	Sherlock Holmes	Lord of the Rings	Percy Jackson	Twilight	Warriors
Chapters	Total	1.29M	1.01M	0.36M	0.09M	0.07M	0.06M	0.01M
	Mean	2.92	3.62	2.92	3.92	3.65	6.31	5.07
	Std	7.35	8.72	7.38	10.74	8.17	11.34	9.35
	Median	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Kudos	Total	101.82M	49.34M	17.35M	1.85M	3.14M	0.98M	0.07M
	Mean	228.92	176.03	139.58	76.42	163.27	89.77	28.54
	Std	587.18	591.08	412.32	245.48	316.22	225.77	69.12
	Median	78.0	45.0	50.0	22.0	74.0	27.0	10.0
Bookmarks	Total	14.62M	8.73M	2.49M	0.26M	0.37M	0.19M	9696
	Mean	32.87	31.15	20.04	11.13	19.37	17.77	3.51
	Std	121.9	147.68	100.0	60.07	54.04	57.12	10.77
	Median	7.0	4.0	3.0	2.0	6.0	4.0	1.0
Comments	Total	11.96M	6.61M	2.86M	0.35M	0.36M	0.15M	0.02M
	Mean	26.9	23.61	23.06	14.63	19.2	14.18	9.58
	Std	132.73	139.81	110.28	90.58	95.92	55.34	48.37
	Median	6.0	3.0	5.0	2.0	5.0	2.0	1.0
Words	Total	3299M	2961M	766M	243M	169M	204M	26M
	Mean	7417.84	10563.71	6165.81	10032.92	8806.12	18619.41	9615.16
	Std	21926.47	33020.84	19330.09	33010.74	26198.51	41964.01	22235.52
	Median	2158.0	2225.5	1707.0	1860.5	2242.5	4247.0	2115.0

Table 2 - Works characteristics by AO3 fandoms

The main conclusions are:

1. The fandoms generating the largest amount of content are clearly Marvel and Harry Potter. Sherlock Holmes also stands out as #3, far behind the first two, but also way ahead of the rest. Warriors is clearly the smallest fandom from those studied.
2. Authors produce on average 2-3 chapters per fandom. Interestingly, it's the lowest contributing fandoms (Twilight, Warriors) in which authors average more contributions (6-7). However, with the distribution being skewed, it is safer to characterise this production based on the median (equal to 1, irrespective of the fandom).
3. Marvel remains the most appreciated fandom in terms of kudos-per-chapter (mean of 229). Interestingly, Percy Jackson, one of the fandoms with fewer chapters, seems to have attracted just as much appreciation (μ=163) as the one ranked at #2, Harry Potter (μ=176). This comparison is even more obviously in favour of Percy Jackson, when we look at its median (74), very close to Marvel's (78) and above that of Harry Potter (45).
4. A similar conclusion can be drawn about the bookmarks: despite a higher average of the top-2 fandoms, a look at the median shows that Marvel and Percy Jackson standing out (7 and 6 per chapter, respectively), while Harry Potter seems to be more similar to Twilight (4 each).
5. Concerning the comments received, it is again the chapters from Marvel and Harry Potter that clearly stand out, together with those from Sherlock Holmes (averaging around 23 comments per chapter). Once again, due to the nature of the distribution, the median paints a slightly different picture, with Marvel, Sherlock Holmes and Percy Jackson attracting more comments per chapter (5-6 per chapter).
6. Finally, we look at the length of the chapters produced in each fandom. Most of them average between 6,000-10,000 words and it is one of the "smaller" fandoms, Twilight, that clearly exhibits the longer contributions. This is even clearer when we analyse the medians, as we distinguish 3 groups of fandoms: i) Twilight, ii) Marvel, Harry Potter, Percy Jackson, Warriors, and iii) Sherlock Holmes and Lord of the Rings.

Last, in Table 3 we look at the production activity from an author's perspective in each of the fandoms, by summarising the productions (chapters or words) of each author in each fandom.

		Marvel	Harry Potter	Sherlock Holmes	Lord of the Rings	Percy Jackson	Twilight	Warriors
Chapters	Mean	15.45	15.25	14.9	13.84	9.31	14.09	9.86
	Std	42.34	47.44	53.81	40.61	27.96	44.9	24.75
	Median	4.0	4.0	3.0	3.0	3.0	4.0	3.0
Words	Mean	39.31K	44.45K	31.42K	35.38K	22.46K	41.58K	18.69K
	Std	116K	153K	114K	113K	92K	158K	54K
	Median	7500	7821	4780	5094	4643	8217	3727

Table 3 - Characteristics of authors' contributions in AO3 fandoms

1. On average, authors seem to produce a comparable number of chapters, irrespective of the fandom. The exceptions are Percy Jackson and Warriors, in which authors produce considerably fewer works. However, this effect disappears when looking at the median chapters per author, which sits at around 3-4 chapters independently of the fandom.
2. Authors in Harry Potter and Twilight appear to write longer contributions, while authors in Percy Jackson and Warriors seem to write shorter ones.

Tags

Unfolding the list of tags attached to each chapter under each work, we obtain an impressive dataset of 332.66 million records, amounting to 865,518 unique tags over the seven AO3 fandoms analysed. As we will discuss later, this huge number of tags imposes some computational restrictions on the rest of the analyses.

	Marvel	Harry Potter	Sherlock Holmes	Lord of the Rings	Percy Jackson	Twilight	Warriors
Tags	519.8K	254.7K	107.9K	26.1K	42.7K	15.7K	6.3K
Chapter-Avg.	0.4	0.25	0.3	0.29	0.61	0.26	0.64
Title-Avg.	1.46	1.11	1.02	1.16	2.33	1.52	2.33

Table 4 - Unique number of tags used in each AO3 fandom

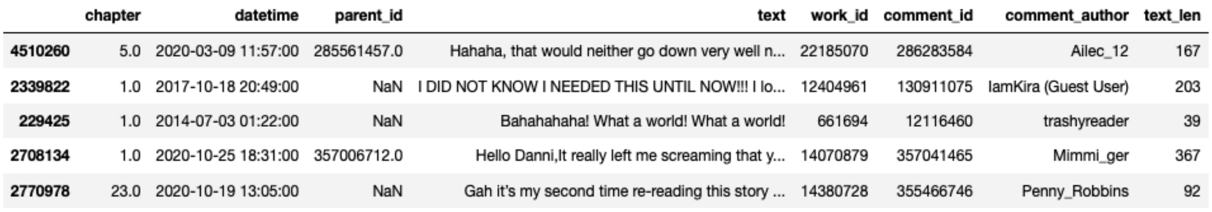
Nevertheless, when looking at the average tags per chapter and tags per title, these numbers are not so high. In fact, on average, chapters barely get tagged, while titles receive 1-2 tags.

Concerning the frequency of use of unique tags, Table 5 gives a list of the top 10 tags across all AO3 fandoms and the number of times they have been used.

Tag	Frequency
Angst	3648101
Fluff	3131127
Hurt/Comfort	2657218
Slow Burn	2216790
Alternate Universe - Canon Divergence	2192835
Romance	1639111
Alternate Universe	1517331
Anal Sex	1247356
Humor	1243801
Angst with a Happy Ending	1215256

Table 5 – Top 10 most frequent tags across AO3

Comments



	chapter	datetime	parent_id	text	work_id	comment_id	comment_author	text_len
4510260	5.0	2020-03-09 11:57:00	285561457.0	Hahaha, that would neither go down very well n...	22185070	286283584	Ailec_12	167
2339822	1.0	2017-10-18 20:49:00	NaN	I DID NOT KNOW I NEEDED THIS UNTIL NOW!!! I lo...	12404961	130911075	IamKira (Guest User)	203
229425	1.0	2014-07-03 01:22:00	NaN	Bahahahaha! What a world! What a world!	661694	12116460	trashyreader	39
2708134	1.0	2020-10-25 18:31:00	357006712.0	Hello Danni,It really left me screaming that y...	14070879	357041465	Mimmi_ger	367
2770978	23.0	2020-10-19 13:05:00	NaN	Gah it's my second time re-reading this story ...	14380728	355466746	Penny_Robbins	92

Figure 4 - Sample from the AO3 Comments data set

Figure 4 illustrates a sample of comment records from the data set scraped form AO3 for the purpose of this analysis. It contains the following fields for each comment:

- *chapter* – the number of the chapter within a work, that a user has commented on
- datetime – comment timestamp (ISO format)
- *parent_id* – identifier of the comment that a user is replying to. Null if this is a root comment
- *text* – the content of the comment
- *work_id* – the id of the work that received the comment
- *comment_id* – unique id of the comment
- *comment_author* – user that left the comment
- *text_len* – the length of the comment in characters

2.3 Community dynamics

Activity distributions

To understand the theoretical distributions of the user actions, we plot the empirical data for four of the main fandoms (Marvel, Harry Potter, Sherlock Holmes, Lord of the Rings) on a log-log scale (see Figure 5). While a power law distribution is typically expected in content production systems such as AO3, we would like to validate our hypothesis by comparing with other similar distributions: exponential (EXP), lognormal (LGN), power law with exponential cut off (truncated power law) (TPL).

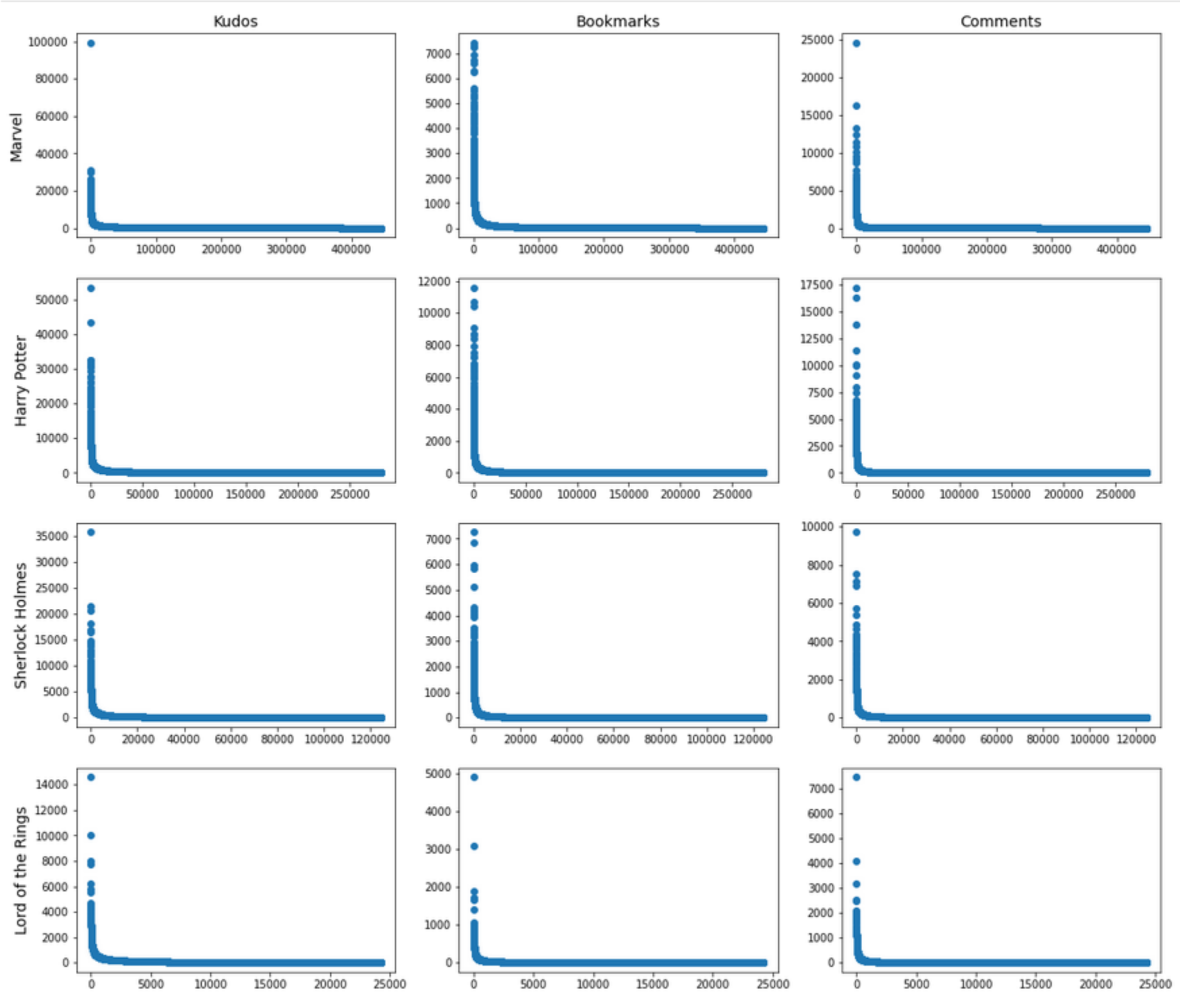


Figure 5 - Distribution of AO3 users' interactions

Table 6 summarises this analysis, by comparing the goodness of fit between the most reasonable power law fitting the data on one hand, and the 3 alternative distributions on the other hand. This is done for each category of user actions and for each of the 4 main fandoms. The columns describe the results of this comparison by means of two statistics:

- The ratio R describing which of the 2 distributions fits the data better: a positive R supports the first distribution; a negative R supports the second.
- The p-value. We consider a p-value smaller than 0.05 to be a significant result in support of one of the distributions.

		PL vs. EXP		PL vs. LGN		PL vs. TPL	
		R	p	R	p	R	p
Harry Potter	Comments	20067.25	0.00	-252.97	0.00	-321.64	0.00
	Kudos	62.43	0.00	-2.83	0.13	-3.88	0.01
	Bookmarks	143.11	0.00	-5.58	0.04	-7.32	0.00
Lord of the Rings	Comments	2007.81	0.00	-6.76	0.02	-13.76	0.00
	Kudos	330.15	0.00	-7.23	0.02	-8.57	0.00
	Bookmarks	833.36	0.00	-10.52	0.01	-13.01	0.00
Marvel	Comments	23544.19	0.00	-399.63	0.00	-421.03	0.00
	Kudos	31.86	0.16	-2.34	0.33	-1.62	0.07
	Bookmarks	368.01	0.00	-61.71	0.00	-70.12	0.00
Sherlock Holmes	Comments	6508.43	0.00	-96.89	0.00	-110.72	0.00
	Kudos	2815.81	0.00	-107.58	0.00	-119.04	0.00
	Bookmarks	3482.68	0.00	-111.58	0.00	-122.43	0.00
Percy Jackson	Comments	1005.03	0.00	-2.91	0.14	-5.82	0.00
	Kudos	17.48	0.04	-1.43	0.29	-1.80	0.06
	Bookmarks	97.22	0.00	-11.03	0.00	-12.24	0.00
Twilight	Comments	1063.72	0.00	-40.86	0.00	-45.20	0.00
	Kudos	10.51	0.13	-1.84	0.21	-2.52	0.02
	Bookmarks	87.55	0.00	-7.57	0.02	28.74	0.00

Warriors	Comments	450.74	0.00	-0.68	0.41	-2.35	0.03
	Kudos	50.07	0.00	-1.37	0.32	-1.56	0.08
	Bookmarks	115.01	0.00	-1.35	0.28	-2.04	0.04

Table 6 - Comparison between the goodness of fit of selected distributions of AO3 user interactions

We can conclude that the data fits the Lognormal and the Truncated Power Law distributions better than the Power Law. We visualise these insights in Figure 6, in which we plot the empirical distribution of the data over the three theoretical distributions mentioned before.

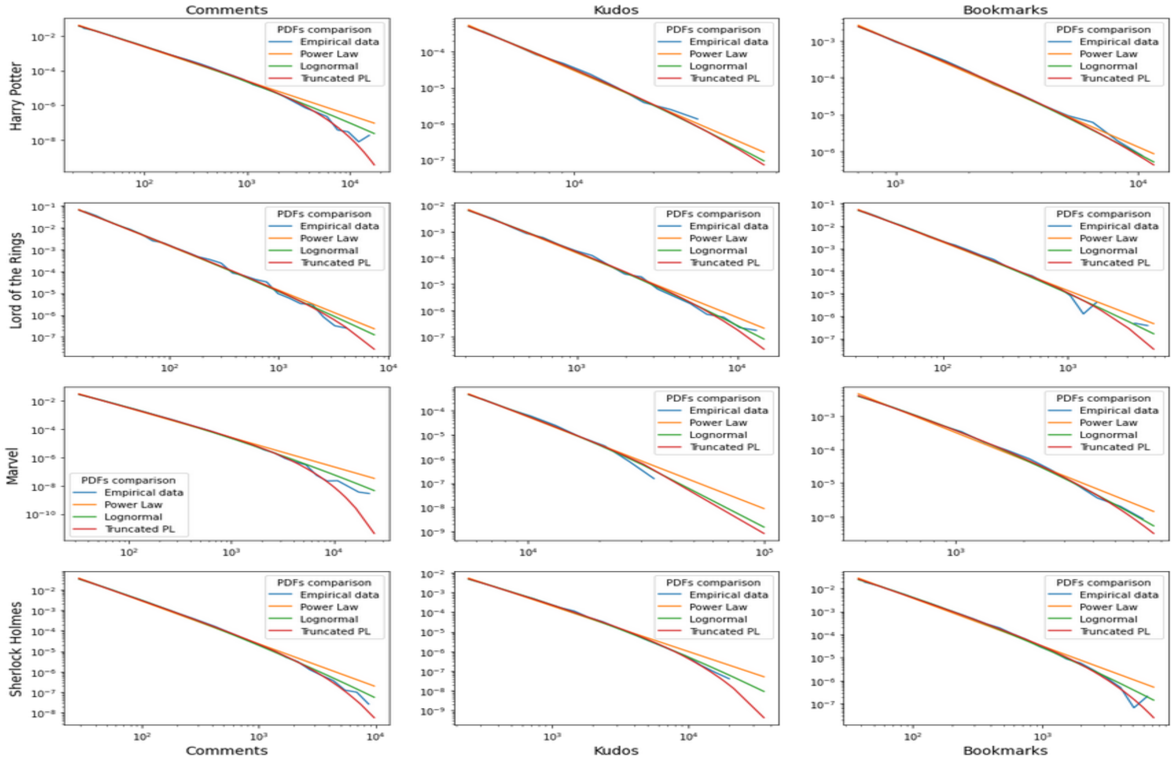


Figure 6 - PDFs fits using selected distributions: power law (orange), a lognormal (green) and a truncated power law (red)

Correlations

The distributions of the user actions seem to be very similar, although applicable at different scales. To understand their magnitude, Table 7 synthesises the correlations between each pair of user actions, for each of the analysed fandoms.

Fandom	Comments vs. Bookmarks	Comments vs. Kudos	Bookmarks vs. Kudos
Marvel	0.54	0.49	0.89
Harry Potter	0.64	0.60	0.93
Sherlock Holmes	0.51	0.50	0.92
Lord of the Rings	0.64	0.58	0.90
Percy Jackson	0.50	0.52	0.88
Twilight	0.67	0.66	0.94
Warriors	0.68	0.65	0.93

Table 7 - Correlations between AO3 users' interactions

Production activity

In this section, we try to quantify the production activity of AO3 users: number of works, number of words that they wrote, number of comments and the total number of contributions (as the sum of works and comments). Figure 7 shows the distribution of the production activity in works, words, and chapters. As was the case with the consumption behaviour, there is only a very small number of users (in this case authors), that produce a lot of content, while the overwhelming majority of users produce only a little content.

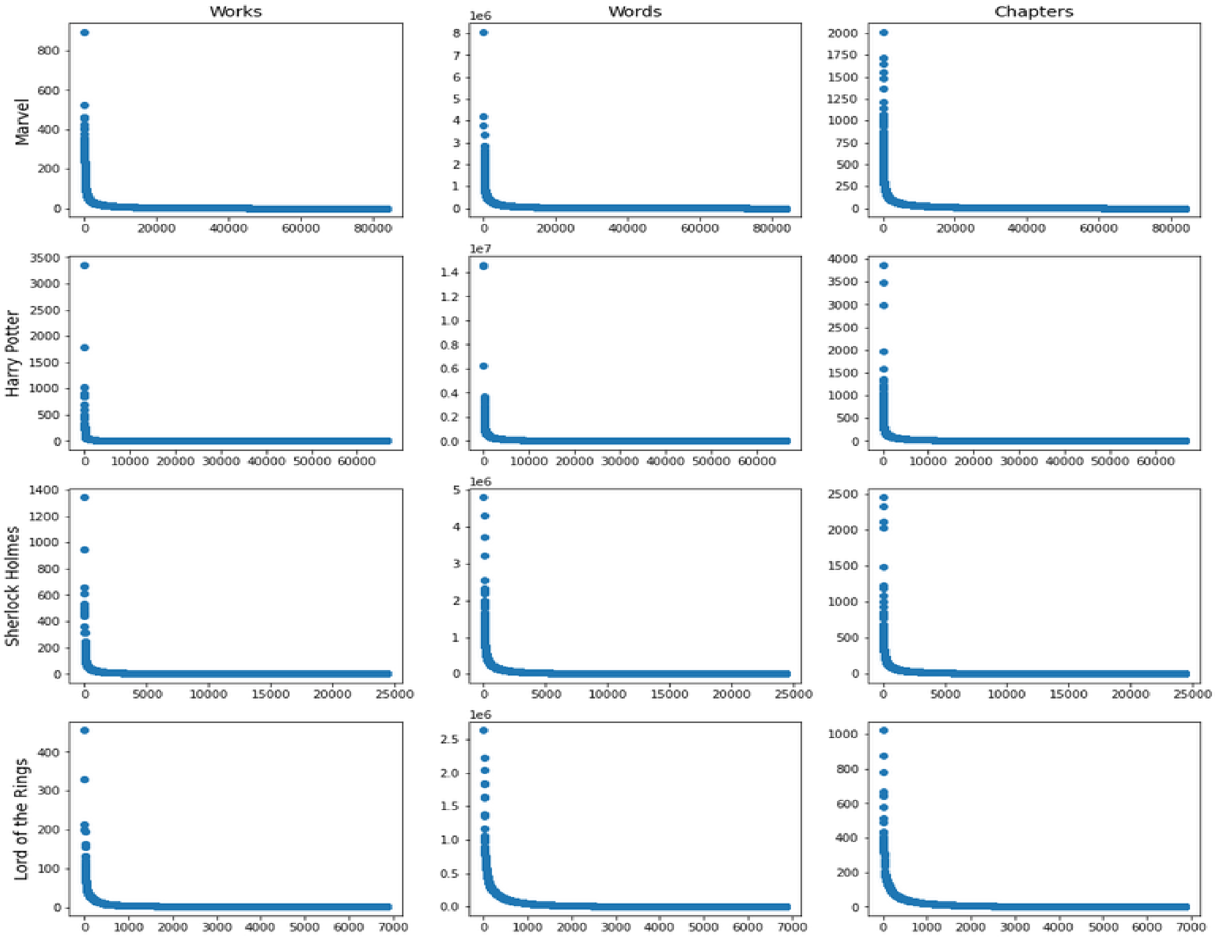


Figure 7 - Distributions of the content produced by AO3 authors in the largest fandoms

2.4 Popularity dynamics

Popularity metrics

Popularity in content production systems can refer to several different things: number of “likes”, bookmarks, number of comments, scores (“stars”) in a reputation system etc., to name the most obvious. Therefore, before proceeding, we need to define how we measure popularity in our analyses of AO3. This definition needs to be aligned both with our requirements in terms of a prediction model and what the platform has to offer in terms of scraped data.

Work popularity

To visualise the dynamics of popularity over time (Figure 8), we consider the first 90% of the comments and plot their cumulative sums over the number of days from initial publication. We sampled among the top 20% chapters in the four largest fandoms (Marvel, Harry Potter, Sherlock Holmes, Lord of the Rings) and plotted their cumulative comments over time. We further created 4 separate classes of popularity corresponding to the 95th, 90th, 85th and 80th percentiles, resulting in:

- Red, top 5%
- Green, top 5-10%
- Blue, top 10-15%
- Orange, top 15-20%

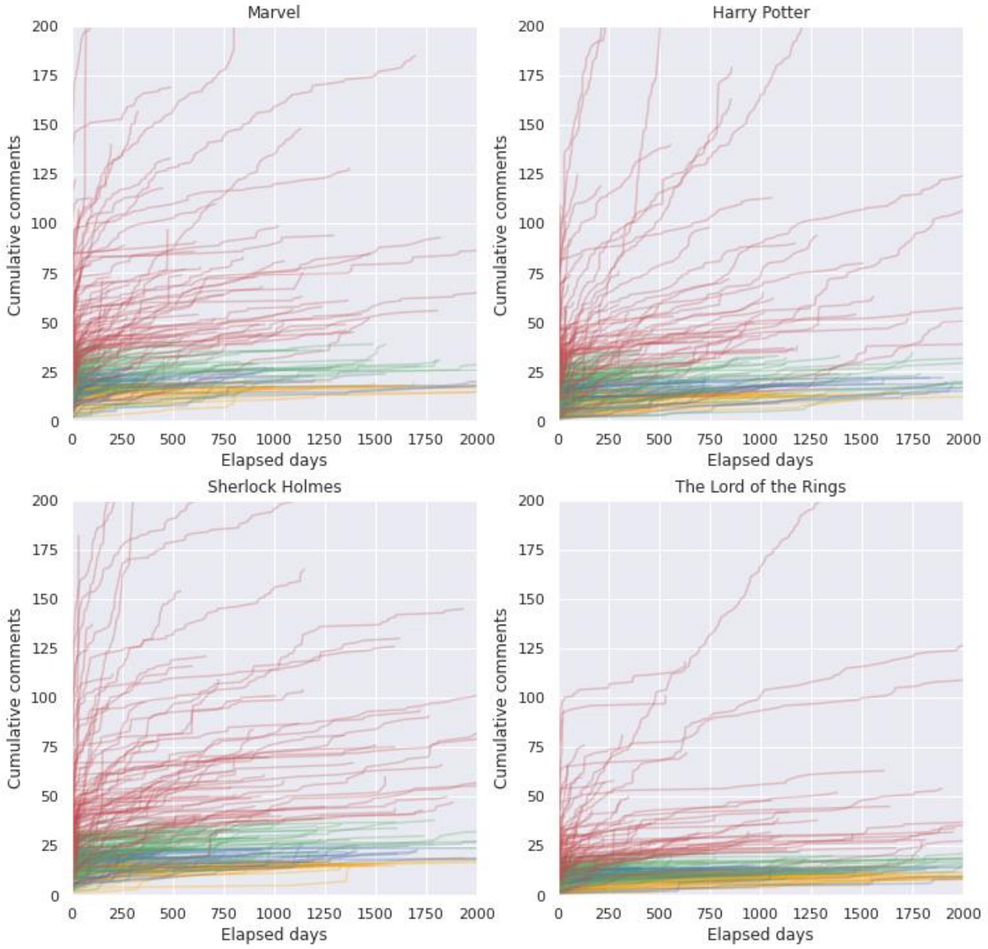


Figure 8 - Chapters popularity growth for the largest AO3 fandoms. Colour code: top 5% chapters in red, top 5-10% in green, top 10-15% in blue, top 15-20% in orange.

Very few chapters are successful from the onset, receiving most of their comments in the first 100 days. These are the chapters for which there is no clear overlap in the charts in Figure 8. Unfortunately, this is not the case for most of the chapters, for which the initial overlap between colours gives a visual intuition of the difficulty (if not impossibility) of predicting the total number of comments far into the future, based on the initial performance of a chapter.

In Harry Potter in particular, the overlap between the orange, blue and green lines persists even after years, while in Sherlock Holmes we see examples of chapters that suddenly become popular 1-2 years after staying unnoticed.

Typically, the overlap is very noticeable between adjacent groups of colours (but not only!) before the 500-days mark, but very often even beyond. As a conclusion, the early popularity of a chapter does not correlate well with the popularity it can attain many years after.

Previously we discussed the correlation between the total number of comments and bookmarks. Now, we want to study this interaction in the context of the lifetime of a work, considering monthly data. More specifically, we are interested to check how correlated are the vectors that hold the cumulative number of bookmarks and comments for each month after the publication of a work. This experiment doesn't include the kudos since there is no date and time information associated with them.

Fandom	Monthly	Cumulative
Percy Jackson	0.35	0.80
Twilight	0.26	0.82
Warriors	0.21	0.77

Table 8 - Correlations between total comments and bookmarks in AO3

While we know that the cumulative correlations are very strong, the same cannot be said for the monthly data. This indicates that even though the two are correlated cumulatively, this effect doesn't necessarily happen in the same time frame.

Figure 9 illustrates the way in which works accumulate feedback over time, up to eventual saturation. It plots the percentage of comments from total at different milestones (1 week, 1 year, 2 years, 4 years, 8 years, 16 years).

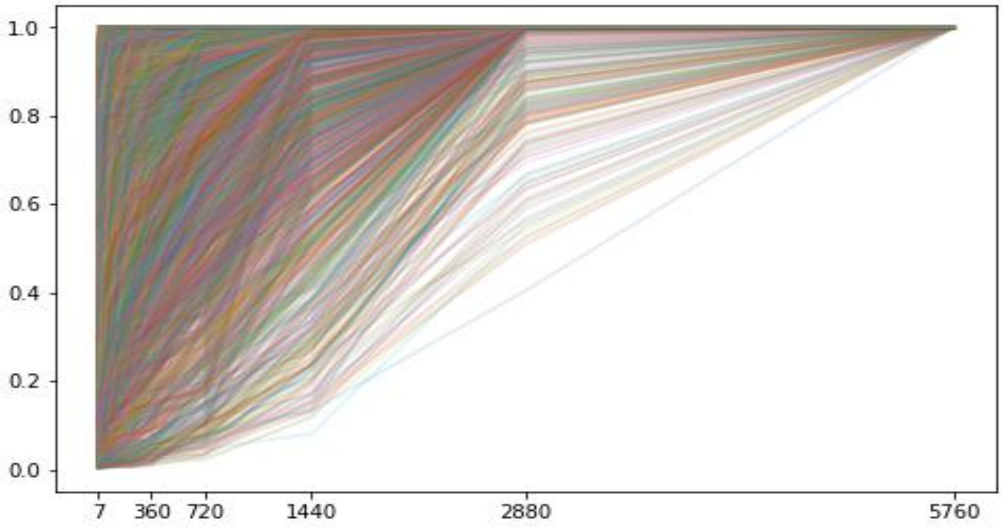


Figure 9 - Saturation patterns for works popularity in AO3

Tag popularity

In Section 2.2 (subsection Tags) we discussed about the large number of tags that are generated over the entire AO3 content. This quickly causes computational issues when trying to replicate the analysis of lifetime dynamics previously performed for works.

The quantile plot for the AO3 tags (see Figure 10 – left and Table 4) illustrates how a very small number of tags have very large frequencies and vice-versa. Applying a knee locator indicates that the inflection point in the plot in Figure 10 – left corresponds to the quantile at 0.971. Thus, we perform the rest of the analysis with tags corresponding to the top 2.9% frequencies; these are exactly 25096 unique tags, with a minimum frequency of 1081 appearances.

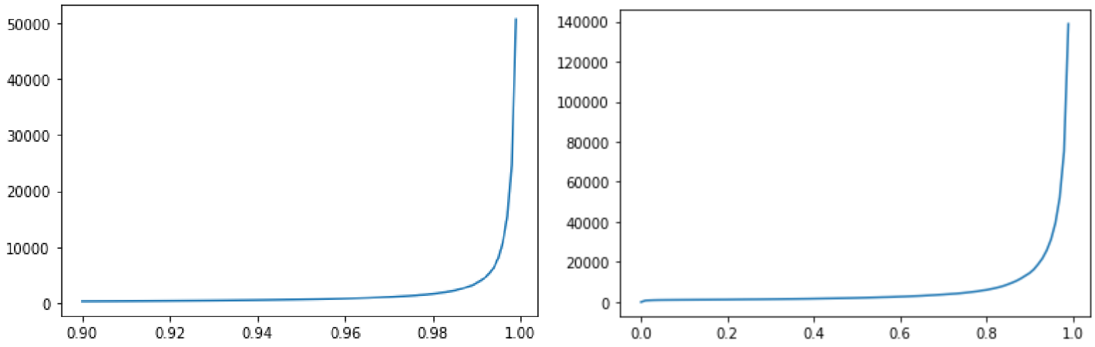


Figure 10 - Quantile plots of all AO3 tags (left) and of top (2.9%) AO3 tags

We plot the lifetime of these tags in a similar way to that of works, by creating 3 colour-coded classes indicating the level of relative popularity attained by each tag. These levels are established by analysing the quantile plot corresponding to the top 2.9% (see Figure 11):

- red – over the 95th percentile (established by knee locator)
- green – between the 80th (established visually) and 95th percentiles
- blue – between the 60th (established visually) and 80th percentiles

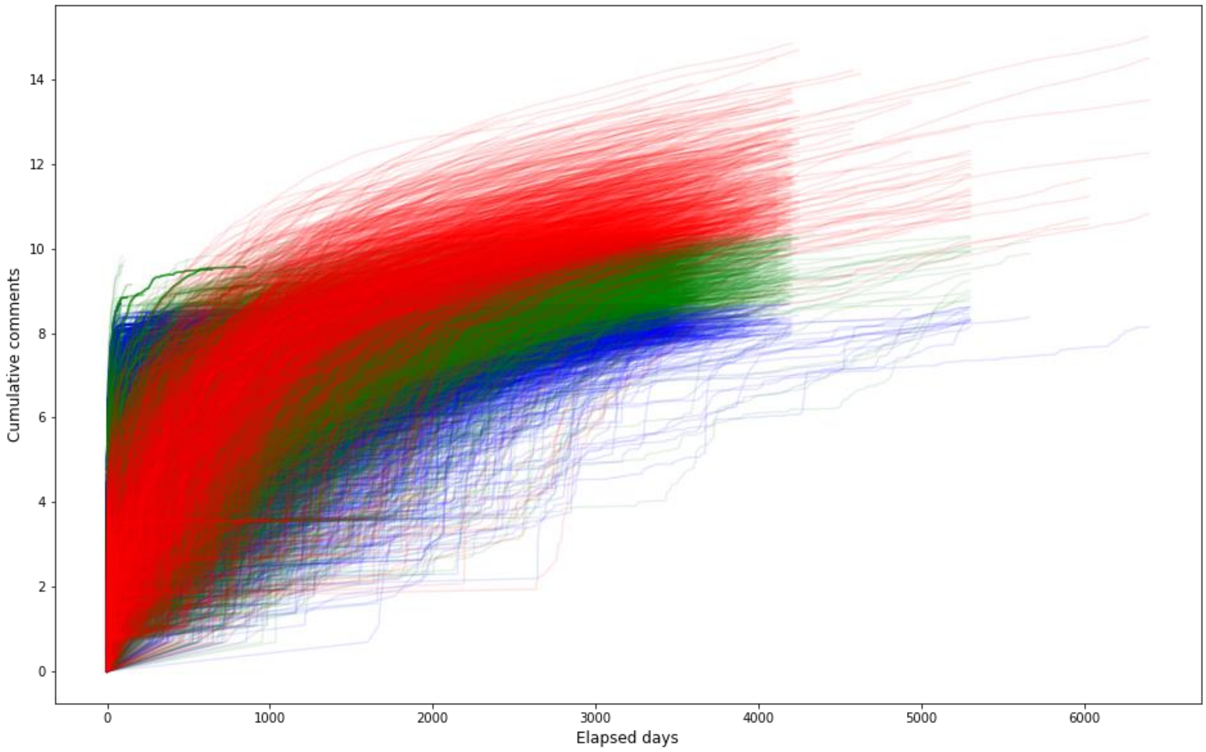


Figure 11 - Lifetime dynamics of top (2.9%) AO3 tags

As in the case of works, we see that differentiations between these classes are very difficult in the first 800 days of a tag's life. This effect may persist well after 1200 days (more than 3 years!), indicating that it can be even more difficult to predict the success of a tag based solely on its initial performance.

Some "early shooters" indeed draw attention (the region to the left of the red stream); however, they end up staying in the lower (blue) popularity zone.

An interesting effect, likely due to the association of several tags to the same title (or chapter), is that most of these tags seem to evolve together, visually creating the stream effect observable in Figure 11. There appear to be very few tags that evolve on their own, completely independently of others.

2.5 Popularity prediction model for fanfiction communities

2.5.1 Predicting popularity of works

We define the popularity of a work through the feedback it receives. Typically, the feedback should be as simple as the **total number of comments** received by a work (or chapter). However, such a metric can easily become biased towards works for which only a small number of users engage in a long thread (e.g., flames). Thus, we choose to define feedback as the **number of distinct users** that comment on a given work (or chapter).

We can now formulate the prediction problem that we want to solve:

Considering the feedback received during the past P days, will a work reach top N, within the next F future days?

The problem has three parameters:

1. P – the number of past days over which feedback is considered. The longer the model can “look” into the past, the more reliable the prediction. However, it is desirable to take decisions as fast as possible and thus reducing P, while still achieving reliable predictions.
2. F – the prediction horizon or how many days into the future do we wait for a work to become popular.
3. N – when ranking the works based on the received feedback, what is the cut-off line for classifying them as popular?

The model uses a logistic regression to predict whether a work will become popular F days after a time reference (t*), based on the total feedback acquired by a work until t* and the feedback variation between the last P days and t*.

Because of the evolution of the users’ engagement with the platform over time, the model also depends on the moment at which we start training it. Figure 12 shows the total number of comments over time. We notice that the activity in the two largest fandoms has grown very rapidly starting with 2018. For the model to make accurate predictions for the subsequent years, it should use data that is as recent as possible. However, choosing a date which is too recent might result in too small a training sample, which would be insufficient for training a valuable model. Thus, a trade-off between the two needs to be made when choosing the start date.

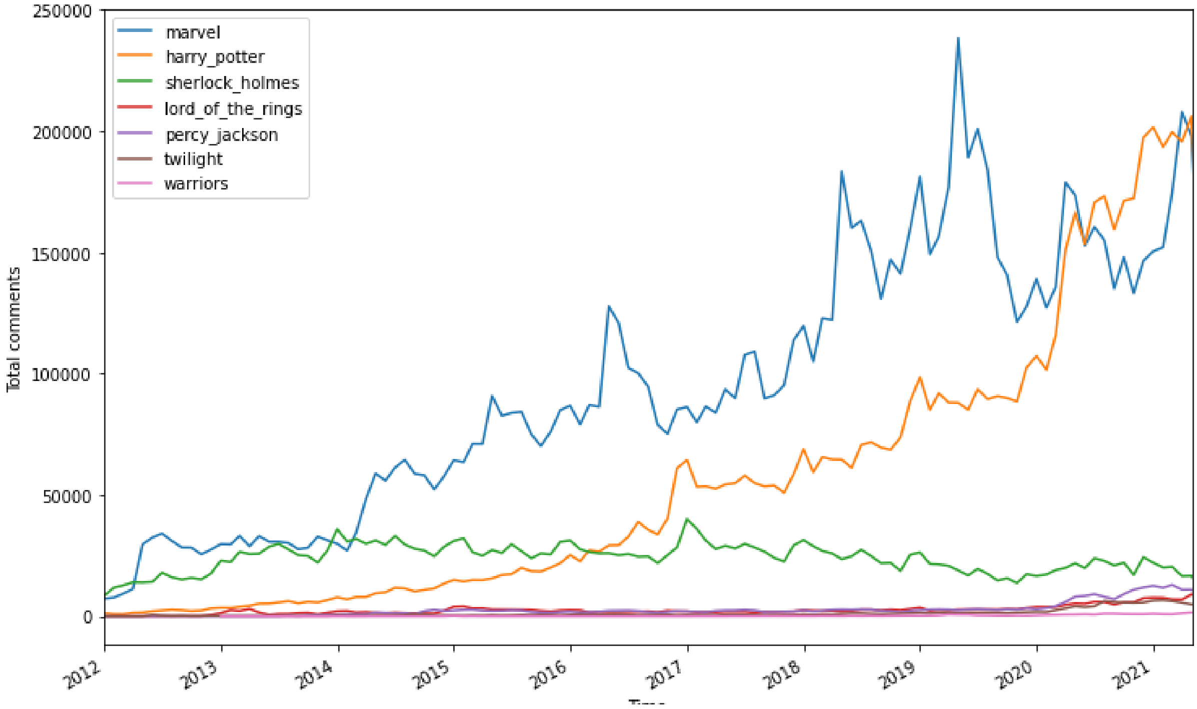


Figure 12 - Evolution of feedback in AO3

We also recall that to predict whether a work will make it in the top N most popular in the next F days following a reference time t* (y), we train a logistic regression that considers the total feedback up to t* and the feedback variation between the past P days and t*. The equation of this regression is the following and includes interaction factors between the two predictors.

$$\ln\left(\frac{y}{1-y}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \beta_5 x_2^2$$

$$y = probability\ of\ being\ top - N, \qquad x_1 = feedback(t^*), \qquad x_2 = \Delta feedback(P, t^*)$$

Finally, we scale the input features and perform an 80-20 train-test split, before training the logistic regression model to obtain the following coefficients.

Coefficient	β_0	β_1	β_2	β_3	β_4	β_5
Value	-9.59	0.15	1.52	-0.01	0.37	-0.01

Table 9 - Work popularity prediction model - Coefficients

Considering the scaling prior to training, the values of the coefficients are useful for understanding the importance of each of the predictors. Thus, we notice that the coefficient corresponding to the feedback variation predictor has the largest importance in the model ($\beta_2 = 1.52$), 5 times larger than second most important, corresponding to interplay between the two predictors ($\beta_4 = 0.37$). At the opposite end, it appears that the quadratic form of the 2 features have the smallest impact on the model ($\beta_3 = \beta_5 = -0.01$).

We trained different models, for various values of the triple past days – future days – top works (P, F, N), and we decided to observe the top 1% (N=0.01) for F=360 days into the future.

P (days)	Precision	Recall	F1-score
30	0.75	0.87	0.80
60	0.79	0.90	0.84

Table 10 - Work popularity prediction results

The results of the validation on the test set show a satisfactory performance of what is a parsimonious model, very similar to results obtained by slightly more elaborated models reported in the literature (Zeng et al., 2013). Increasing the observation horizon from P=30 days to P=60 days, seems to bring a consistent improvement in term of both precision and recall.

2.5.2 Predicting popularity of tags

Another important topic of concern is that of tags popularity. If previously we were studying the popularity of a chapter or an entire work, here we are interested in using the tags associated to these and we investigate the evolution of their popularity. While this may not push a specific piece of content on publishers' radars, it can be a very interesting tool for understanding what **kind of content** is being produced or is generating interest in prosumer communities.

The model is an adaptation of the one used to predict work popularity presented in the previous subsection, and just as that one, it uses the same user feedback to quantify the popularity. The main difference is that the feedback is now being aggregated over each tag, with a specific tag capable of being attached to more than one work.

Because of the size of the data and the resulting computational limitations, the model is trained only with data corresponding to the most frequent tags, as described in Sections 2.2 and 2.4. The model thus predicts whether a tag will make it into the top N tags, F days into the future, after a reference time t*. The predictors are the feedback at the reference time, and the variation in feedback between the previous P days and the reference time. The equation of the model is the same as the one presented in the previous section for the case of work popularity prediction.

Finally, we scale the input features and perform an 80-20 train-test split, before training the logistic regression model to obtain the following coefficients.

Coefficient	β_0	β_1	β_2	β_3	β_4	β_5
Value	-0.41	7.61	11.69	-4.12	11.20	-1.89

Table 11 - Tags popularity prediction model - Coefficients

Considering the scaling prior to training, the values of the coefficients are useful for understanding the importance of each of the predictors. Thus, we notice that the coefficient corresponding to the feedback variation predictor ($\beta_2 = 11.69$) is about 1.5 times more important for the prediction than the coefficient of the predictor for past popularity ($\beta_1 = 7.61$). Interestingly, it seems that the interplay of the 2 has a primary role in the prediction ($\beta_4 = 11.20$). As in the case of works, it appears that the quadratic form of the 2 features have the smallest impact on the model.

We trained different models, for various values of the triple past days – future days – top works (P, F, N), and we decided to observe the top 1% (N=0.01) for F=360 days into the future.

P (days)	**Precision**	**Recall**	**F1-score**
30	0.85	0.91	0.87
60	0.83	0.92	0.87

Table 12 - Tags popularity prediction results

Just like before, we obtain a satisfactory performance on the test set, with a simple and interpretable model.

2.6 Social interactions analysis

2.6.1 Social network analysis models

To study interactions dynamics between users in the community, we rely on social network analysis, and we construct graphs to model different kinds

- **Feedback network (User-author network)**: each node is a user, and we establish a link from user B to user A whenever user B gives feedback to a work published by user A. This kind of network can be constructed for different feedback mechanisms: bookmarks or comments. In the second case (feedback network based on comments) we do not look at the reply thread, but just consider the fact that user B is participating in a thread about a work published by user A. The rationale is that no matter to whom user B is replying in their comment, this is a way to express feedback on user A's work. Therefore, we have a network where authors tend to receive incoming connections from all the users who express feedback on their works, and the most popular authors will have a more central position in the network.
- **Reply network (User-user network)**: each node is a user, and we establish a connection from user B to user A when user B replies to a comment written by user A. So, here we do not account for who is the author of the work about which the users are discussing, but just on the dynamics of the conversation, looking at who interacts with whom.

Both kinds of networks are directed, and may have weights according to the number of interactions between two users, in a given direction; e.g., if user B gives feedback in the form of comments on three works by user A, then we will have a weight of 3 in the edge from B to A in the feedback network; analogously, if user B replies to three comments by user A in some discussion thread, then the weight of the edge from B to A in the reply network will be 3.

However, for some metrics we need to deal with an unweighted graph, e.g., where we don't take into account the number of interactions between two users, but only whether there was interaction or not; and some metrics are defined for undirected graphs, where the direction of the edge does not matter. So, we built three different graphs for each kind of network (reply network, feedback network):

- **Directed Weighted Graphs**: directed edges, weighted by the number of interactions in a given direction.
- **Directed Unweighted Graphs**: these graphs are obtained from the previous ones (Directed Weighted Graphs) adding a filter on the weight of the edges: we discard all the edges that have weight lower than the threshold; in this case, we use a threshold of 3, to discard occasional interactions that only happened one or two times. The edges

that have not been discarded, representing more consolidated connections, are considered as unweighted.

- **Undirected Unweighted Graphs**: these graphs are obtained as well from the first ones (Directed Weighted Graphs) by collapsing the directed edges into undirected. The weight of the new edge is given by the sum of the edges that have been collapsed, in either direction. Finally, the edges are filtered based on their weight as we did for the Directed Unweighted Graphs, again with threshold 3.

2.6.2 Network structural metrics

Once defined these different graph types, we can compute different social network metrics that help us to understand the social dynamics in each community:

- **Density**: in the undirected unweighted graph, this measure indicates the ratio of the number of edges and the number of possible edges (Wasserman and Faust, 1994).
- **Giant component**: in the undirected unweighted graph, we can identify different connected components, where each connected component is defined as a group of nodes that are connected to each other through some path in the network. Usually in social network most nodes tend to be connected to each other, being part of a so-called giant component, whose size may typically overcome the 90% of the nodes. A high percentage of nodes belonging to the giant component of the network is an indicator of the cohesion of the community, while a low percentage is an indicator of fragmentation, as it indicates the existence of groups of users disconnected from the rest, e.g. from the giant component.
- **Clustering coefficient**: in the undirected unweighted graph, this measure provides information about how many closed triads (two users connected with a common node who are also connected with each other) can be found within the network; for this reason, it is also referred to as transitivity. It is defined as the percentage of closed triples in the network. This index defines the degree to which the social network is dominated by cliques (groups of users who are highly connected among each other but have significantly less connections to other users). At the extremes, a completely connected graph has *clustering coefficient = 1*, whereas a hierarchical tree has *clustering coefficient = 0*, as no loops are possible.
- **Reciprocity**: in the directed graphs, this metric represents the proportion of edges that are reciprocal, e.g., bidirectional. In other words, when a connection from a node A to a node B exists, how often does also a connection from B to A exist? Reciprocity indicates the value of this frequency.

2.6.3 Network structural metrics results

We built the different kinds of networks, introduced in this section, for the 7 communities in our dataset, and we computed the structural metrics described above.

	Nodes	Edges	Clustering coefficient	Reciprocity	Density	% nodes in giant CC
Marvel	118193	496196	0.028	0.036	0.00004	97.4
Harry Potter	96558	315425	0.016	0.026	0.00003	96.4
Sherlock Holmes	29283	112284	0.046	0.047	0.00013	95.8
Lord of the Rings	8347	12821	0.018	0.034	0.00018	90.6
Percy Jackson	9588	14717	0.009	0.012	0.00016	90.3
Twilight	5448	7499	0.010	0.013	0.00025	87.0
Warriors	784	1155	0.029	0.021	0.00188	77.4

Table 13 - Structural network metrics for the feedback networks based on comments (user-author comment network) for each fanfiction community.

Results for the feedback network based on comments, shown in the table above, show that these networks tend to have a big giant component for the larger networks, and more fragmented communities for smaller networks.

The Sherlock Holmes community seems to have the most cohesive community structure, with a clustering coefficient and a reciprocity of almost 0.05, while Percy Jackson and Twilight have the lowest values for both metrics.

The reciprocity ranges from 1.3% to 4.7% of reciprocated connections; these are low values that could be expected given the asymmetry of this kind of network, where most users have only outgoing connections and do not receive any incoming connection (I.e., they do only give feedback to other users' works, and have not published any work). Still, it is interesting to observe that despite this structural asymmetry, a certain number of connections is anyway reciprocal, with users providing feedback to each other (at least 3 times in each direction).

	Nodes	Edges	Clustering coefficient	Reciprocity	Density	% nodes in giant CC
Marvel	115783	745765	0.031	0.698	0.00006	98.2
Harry Potter	93971	443460	0.019	0.608	0.00005	97.0
Sherlock Holmes	28676	177539	0.050	0.760	0.00022	96.7
Lord of the Rings	8221	19989	0.021	0.742	0.00030	92.2
Percy Jackson	9221	21177	0.011	0.649	0.00025	92.7
Twilight	5345	10167	0.011	0.559	0.00036	87.9
Warriors	751	1635	0.036	0.626	0.00290	79.8

Table 14 - Structural network metrics for the reply networks (user-user discussion network) for each fanfiction community.

In the reply network we see that users create denser and more cohesive networks, engaging in conversations with one another, and especially replying to each other, which leads to a reciprocity of over 50%.

Also, in this kind of network the Sherlock Holmes community presents the strongest community structure, with the highest values of both clustering coefficient and of reciprocity.

2.6.4 Network centrality metrics

We look at centrality metrics that are computed for each node to measure how central it is in the network, according to different criteria:

- **Out-degree**: in the directed graphs, it indicates the number of outgoing connections towards other nodes. So, the centrality of a given node (user) represents with how many other users they have actively interacted (in the feedback networks, by giving feedback to some of their works; in the reply network, by replying to some of their comments).
- **In-degree**: in the directed graphs, it indicates the number of incoming connections. So, the centrality of a given node (user) represents how many users have interacted with them (in the feedback networks, by giving feedback to some of their works; in the reply network, by replying to some of their comments).
- **Pagerank**: Pagerank is like in-degree, but connections from relevant nodes are given a higher weight. Intuitively, the pagerank of a node represents the probability that, following a random path in the network, one will reach that node. It is computed in an iterative process, as the pagerank of a node depends on the pageranks of the nodes that link to it, however there are fast algorithms to compute it. Pagerank is widely adopted as a measure of relevance, usually only computed in directed networks.

We compute the distribution of centrality metrics for three major communities: Harry Potter, Sherlock Holmes and The Lord of the Rings.

To see how these metrics are distributed across users, and compare them with each other, we draw pairwise scatter plots in which we compare the distributions with each other, in logarithmic scale.

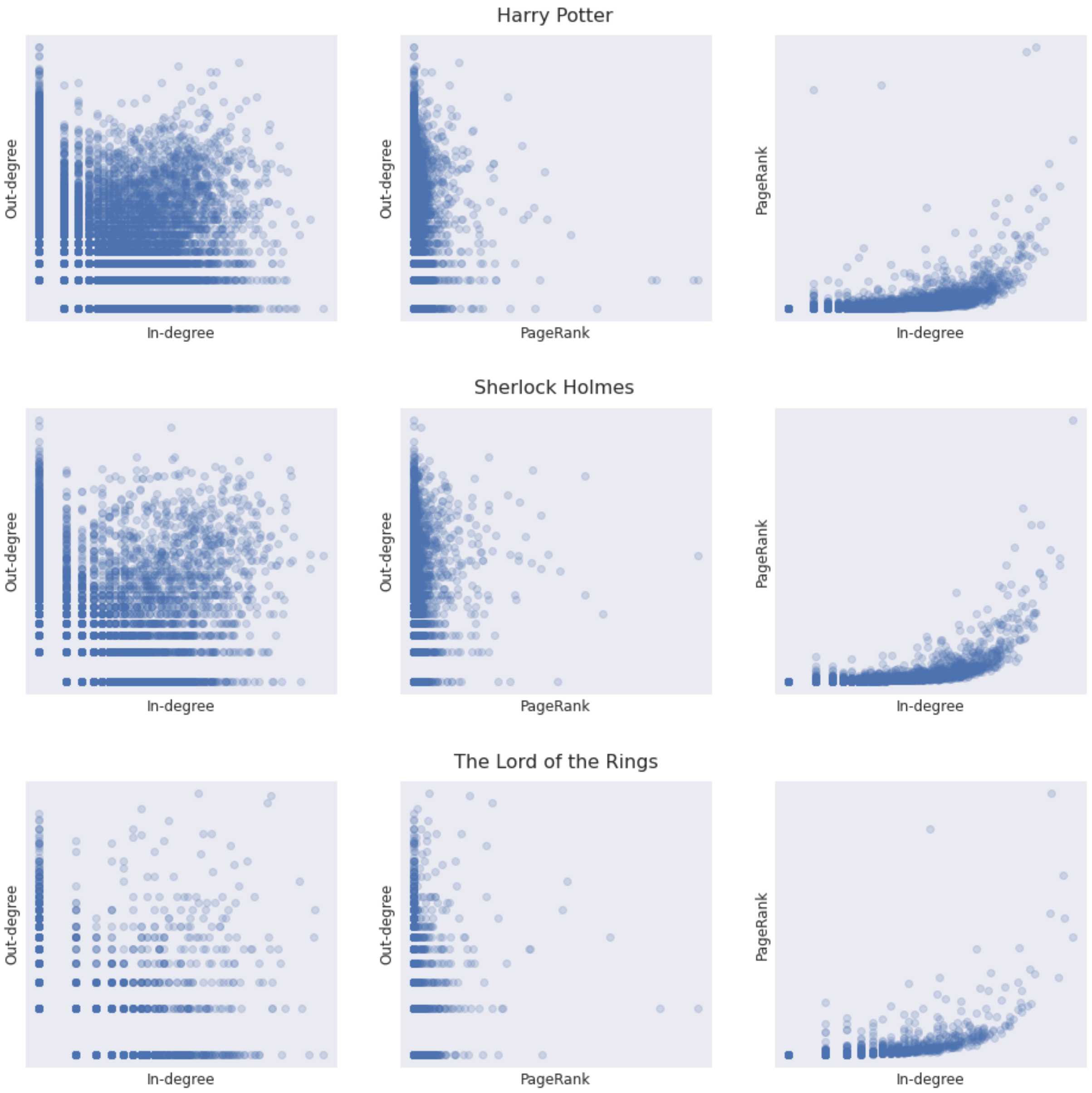


Figure 13 - Pairwise centrality metrics scatter plots for three communities: out-degree vs in-degree (left), out-degree vis pagerank (centre), in-degree vs pagerank (right).

The plots on the left column show the relation between out-degree and in-degree, i.e., between giving and receiving feedback. Intuitively, we see a concentration of dots (users) along the two axes, representing users that have a zero score for one of the two metrics: along the vertical axis, users that do only give feedback on the work of other users, but do not publish any work or receive feedback (in-degree zero); along the x axis, users that publish works and receive feedback on them, but do not give feedback to other users' work (out-degree zero). In particular, we observe that a consistent number of users in the former group may reach high levels of out-degree, having no in-degree.

We then look at the relationship between out-degree and pagerank in the central column; pagerank has a very skewed distribution, and even in logarithmic scale we can distinguish a few outliers with pagerank values much higher than the rest of the users (and generally low out-degree). In Harry Potter and The Lord of the Rings, these seem to represent a bunch of users that are very popular as authors, and do not spend time on other users' work. With respect to in-degree, which gives the same importance to feedback received from any other user, pagerank gives prominence to authors that receive feedback from other popular authors, and this seems to advantage a few users somehow specialized as authors. In Sherlock Holmes we see a difference in that the users with highest pagerank have a higher out-degree, meaning that these very popular authors also spend time providing feedback on other users' work.

Finally, on the right column of the figure we observe the relationship between pagerank and in-degree. Of course, these two metrics are quite correlated, as pagerank can be considered as a variation of in-degree, that accounts for the relevance of the nodes of incoming connections.

Considered these plots, we choose to focus on the ones in the first column (out-degree vs in-degree) for the next analysis, as they have a more even distribution, and they are based on two more immediately understandable and interpretable metrics.

2.7 Prosumer role characterization

2.7.1 Model for prosumer segmentation

Following from the analysis in the previous section, to characterize different kinds of prosumers profiles in each community we look at their position in the feedback networks in terms of in-degree and out-degree. Out-degree indicates how active is a user in giving feedback on the work of other users, while with in-degree we measure how relevant a user is as an author who receives feedback from many others. We follow the idea that each of these two metrics alone offers a limited information, while their combination may represent a good proxy for describing the role of a user in the community.

2.7.2 User clustering

We place all the users from a community on a two-dimensional plan, where on the X axis we have in-degree, and on the Y axis out-degree. In this plan we perform clustering to identify groups of users who have similar characteristics.

We use the k-means unsupervised clustering algorithm. Given a number *k* of clusters, the algorithm looks for a partition of the users that minimizes within-cluster variances.

The first step is then choosing a value for the k parameter. To this aim, we plot three metrics that are commonly used for choosing the best *k*, namely the Inertia (lower is better), Calinski-Harabasz score (higher is better), and the Davies-Boudin score (lower is better).

In Figure 14 we see the results for the feedback networks based on comments, for the three major communities in our dataset. The Inertia suggests that the improvement in the quality of the clusters grows strongly until *k=4* or *k=5,* and then presents only a moderate growth. CH seems to agree, although the improvement seems to be considered larger even when k is greater than 5. If we are willing to look at values of k greater than 5 and judge also by what the 3rd score displays, it seems that *k=9* might be the best compromise, hence we are going to use this value.

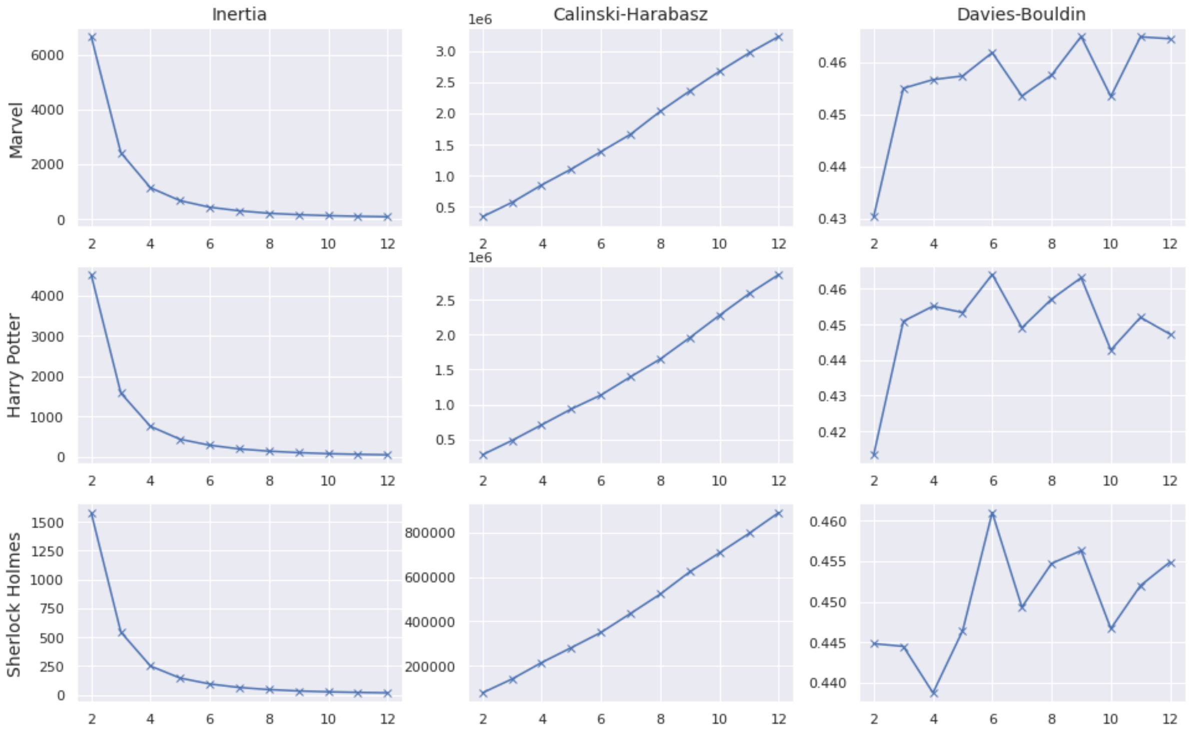


Figure 14 - Inertia, Calinski-Harabasz score and Davies-Boudin score for different values of k (on the x axis) in the feedback network based on comments.

In the feedback network based on bookmarks, we obtain the values Inertia, Calinski-Harabasz score and Davies-Boudin score shown in Figure 15, and we choose a value of k = 8.

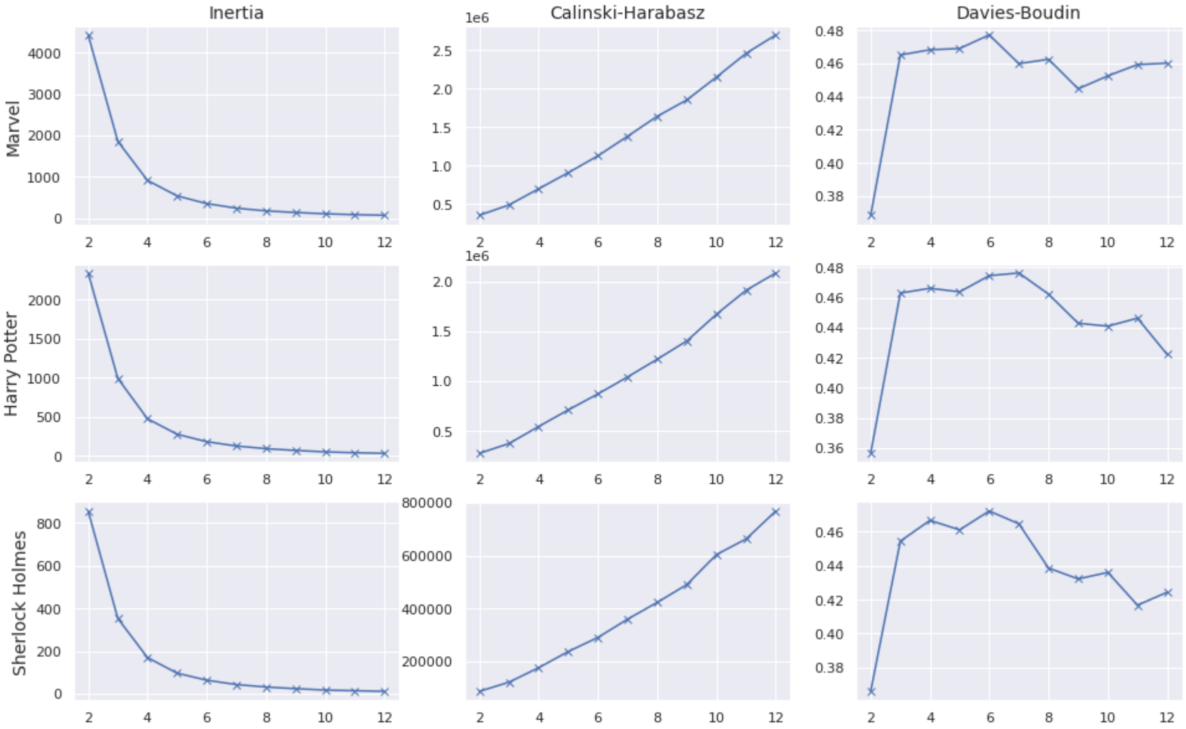


Figure 15 - Inertia, Calinski-Harabasz score and Davies-Boudin score for different values of k (on the x axis) in the feedback network based on bookmarks.

Running the k-means algorithm with *k = 9* for the feedback networks based on comments, we find the clusters shown Figure 15 for the three communities. Interestingly, results are consistent, and tend to draw the same patterns, with similar boundaries between clusters. This indicates a certain robustness of the findings.

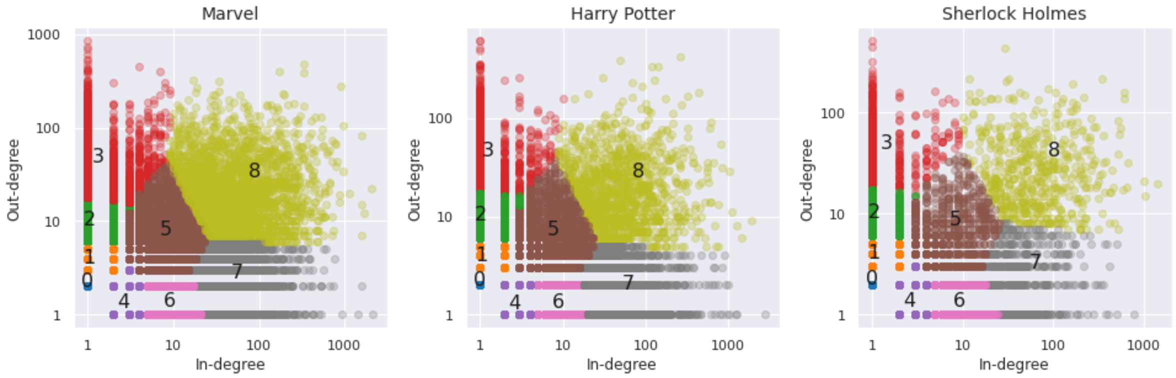


Figure 16 – Clusters resulting from the k-means algorithm applied on the feedback network based on comments (user-author comment network) in the bidimensional plan of out-degree (Y axis) vs in-degree (X axis), for the three largest communities in the dataset.

We can characterize 9 clusters of users, according to their combination of in-degree and out-degree ranges in the feedback network:

- Cluster 8 (light green). **Superprosumers**: high in-and out-degree
- Cluster 7 (gray). **Superproducers**: high in-degree, low out-degree
- Cluster 6 (pink). **Prosumers**: fair in-degree and out-degree
- Cluster 5 (brown). **Producers**: fair in-degree, low out-degree
- Cluster 4 (purple). **Occasional producers**: low activity level, higher in-degree than out-degree
- Cluster 3 (red). **Superconsumers**: high out-degree, low in-degree
- Cluster 2 (dark green). **Consumers**: fair out-degree, low in-degree
- Cluster 1 (orange). **Occasional consumers**:low activity level, higher out-degree than in-degree
- Cluster 0 (blue). **Sporadic consumers**:no out-degree, minimum in-degree

Running the algorithm for the bookmark network, with *k = 8*, we obtain the clusters shown in the figure below.

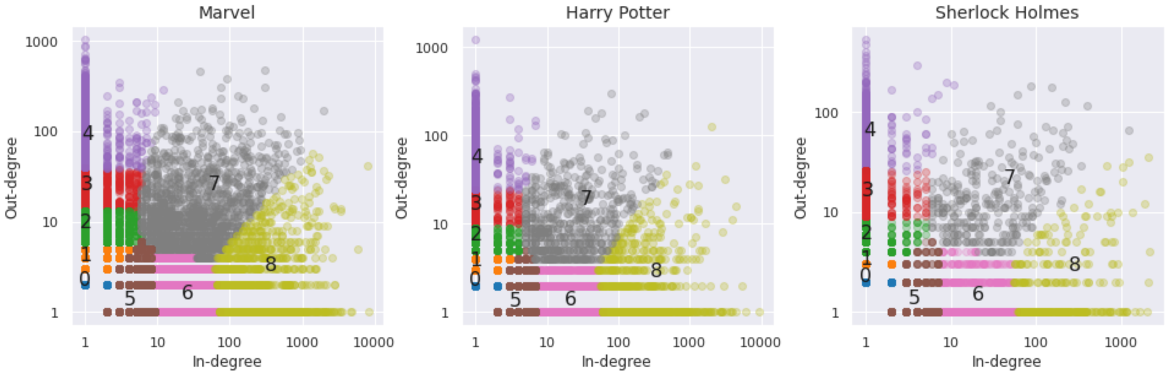


Figure 17 - Clusters resulting from the k-means algorithm applied on the feedback network based on bookmarks (user-author bookmark network) in the bidimensional plan of out-degree (Y axis) vs in-degree (X axis), for the three largest communities in the dataset.

The results for the bookmark user-author network present many similarities with the ones of the comment network; in particular we see again a cluster (in this case, cluster 7) of superprosumers, and two clusters of superproducers (cluster 8) and superconsumers (cluster 4).

2.7.3 Cluster profiles characterization

To characterize each and get an idea of how many users are in each of them, and which is their profile and behaviour, we look at aggregated statistics, with the mean values within each cluster.

Table 15 shows which is the size of each cluster in number of users, what is their in-degree and out-degree, how many works they posted, which is the length of their comments, how many days they have been active, and when they wrote their first and last comment.

	Cluster	Cluster size	In-degree	Out-degree	Works	Posted comments	Comment length	Activity days	First comment	Last comment
Marvel	#0	48689	0.00	1.00	0.24	11.24	220.34	583.20	2017-08-07	2019-03-14
	#1	26855	0.04	2.84	0.54	35.59	202.17	1087.44	2016-12-11	2019-12-04
	#2	9670	0.06	9.87	0.68	119.66	189.15	1481.89	2016-06-13	2020-07-04
	#3	3579	0.38	47.16	1.38	619.51	171.88	1844.17	2015-12-04	2020-12-22
	#4	12849	1.59	0.20	4.42	17.64	164.08	677.06	2017-06-29	2019-05-08
	#5	4452	7.28	6.19	11.07	143.74	200.74	1481.58	2016-06-19	2020-07-10
	#6	6471	7.93	0.23	7.85	53.86	169.69	955.44	2017-03-05	2019-10-17
	#7	3139	53.05	1.37	19.31	334.12	183.43	1449.29	2016-05-27	2020-05-15
	#8	2489	89.42	28.55	39.70	1212.32	193.69	1911.87	2015-10-29	2021-01-22
Harry Potter	#0	46247	0.02	1.00	0.24	10.45	251.22	517.42	2018-06-25	2019-11-25
	#1	19437	0.03	2.57	0.38	30.15	228.55	962.79	2017-11-30	2020-07-21
	#2	7991	0.04	7.93	0.51	94.72	215.45	1247.59	2017-07-25	2020-12-24
	#3	2793	0.22	38.96	0.78	506.47	187.26	1512.66	2017-02-09	2021-04-02
	#4	10264	1.87	0.10	4.26	15.20	188.58	613.97	2018-03-13	2019-11-18
	#5	2380	7.62	5.60	10.42	126.86	232.54	1226.56	2017-08-15	2020-12-24
	#6	4421	10.09	0.20	7.91	59.51	196.89	942.86	2017-11-05	2020-06-06
	#7	1859	73.01	0.94	16.91	342.90	217.64	1331.80	2017-03-26	2020-11-17
	#8	1166	81.95	25.72	34.22	988.16	210.68	1470.49	2017-04-04	2021-04-14
Sherlock Holmes	#0	13908	0.00	1.00	0.28	10.42	226.20	555.44	2016-01-01	2017-07-10
	#1	5574	0.04	2.57	0.60	31.82	208.57	1087.22	2015-03-07	2018-02-27
	#2	2375	0.06	8.43	0.92	105.89	197.68	1541.40	2014-10-08	2018-12-28
	#3	803	0.45	45.15	2.42	627.10	187.08	1980.50	2014-09-06	2020-02-08
	#4	3213	1.55	0.20	4.68	16.29	168.28	822.58	2015-06-21	2017-09-21
	#5	962	7.39	6.59	13.59	160.61	201.14	1778.34	2014-09-20	2019-08-04
	#6	1354	8.17	0.21	9.05	59.54	169.35	1203.93	2015-01-27	2018-05-15
	#7	549	61.53	1.81	24.56	461.37	188.88	2055.85	2013-12-31	2019-08-18
	#8	545	100.26	35.86	53.34	1689.48	186.30	2379.28	2014-05-08	2020-11-12

Table 15 - Aggregated statistics by cluster for the feedback network based on comments (directed unweighted user-author comment network). All the values, except the cluster size, are intended as the average of the values over each cluster.

We believe it is especially relevant to identify users in the 4 clusters in the upper-right side of the graph: first of all, superprosumers, and then superproducers, superconsumers, and prosumers. These users are responsible for most of the content and interaction in the platform; understanding their different profiles and roles can help to identify relevant users for specific purposes, and to drive the growth and productivity of a community. However, we remark that a community is somehow an ecosystem where all the profiles are important and also the larger clusters of users with little involvement have an important function.

Indeed, we can see that the clusters corresponding to higher levels of activity are much smaller in size; cluster 8, that covers a large part of the graph, has 10 times less users than cluster 1, which covers a small portion of the graph, but a very dense one, close to zero. This is an effect of the skewed distribution of activity in the communities.

We observe that “superprosumers” tend to have a higher in-degree than out-degree, which could be expected given the in-degree curve is more skewed, e.g. presents a higher inequality; while out-degree has some boundary due to the limited number of actions a user can perform, in-degree does not have a boundary apart from the size of the community, as it depends on the actions of other users who express interest in a user’s work. Superprosumers have published on average over 30 works (over 50 in the Sherlock Holmes community) and posted about 1000 comments or more. They have been active on average for 5 years (Marvel), 4 years (Harry Potter) and 6 years (Sherlock Holmes); this is in all cases more than the other groups of users even if the difference in many cases is not so large, pointing that time is only one relevant factor for becoming a superprosumer, but not the determinant one.

	Cluster	Cluster size	In-degree	Out-degree	Works	Bookmarks left	Activity days	First bookmark	Last bookmark
Marvel	#0	28999	0.01	1.00	0.75	29.86	962.69	2017-05-28	2020-01-16
	#1	23382	0.02	2.70	0.81	63.64	1264.94	2016-12-25	2020-06-12
	#2	15234	0.04	7.58	0.86	139.93	1476.17	2016-09-13	2020-09-28
	#3	9405	0.06	21.65	0.86	318.42	1666.38	2016-06-01	2020-12-23
	#4	4095	0.08	81.09	0.90	1016.61	1868.45	2016-02-04	2021-03-18
	#5	6161	3.16	0.21	10.68	16.78	793.66	2016-12-30	2019-03-04
	#6	3926	24.64	0.38	19.09	21.77	979.92	2016-06-10	2019-02-15
	#7	1486	60.67	21.45	29.25	292.04	1755.47	2015-11-25	2020-09-14
	#8	1581	362.13	1.86	48.84	48.53	1347.62	2015-09-16	2019-05-26
Harry Potter	#0	26856	0.00	1.00	0.58	25.54	875.17	2018-01-26	2020-06-19
	#1	16090	0.01	2.36	0.58	52.91	1138.51	2017-09-07	2020-10-19
	#2	12976	0.02	5.76	0.55	107.18	1327.88	2017-06-01	2021-01-19
	#3	6969	0.03	15.45	0.54	241.14	1483.61	2017-02-27	2021-03-22
	#4	2807	0.04	52.90	0.57	704.38	1583.99	2017-01-08	2021-05-11
	#5	3273	2.62	0.17	11.13	13.89	652.56	2018-02-02	2019-11-17
	#6	2338	20.46	0.21	18.20	15.92	778.80	2017-08-20	2019-10-08
	#7	573	48.53	16.33	22.26	230.98	1428.94	2017-02-12	2021-01-11
	#8	957	336.74	1.11	46.46	33.77	1183.26	2016-08-23	2019-11-19
Sherlock Holmes	#0	9286	0.01	1.00	0.94	21.78	954.66	2016-03-08	2018-10-19
	#1	4830	0.01	2.34	0.98	46.97	1244.84	2015-10-12	2019-03-09
	#2	3571	0.04	5.75	1.33	94.74	1473.53	2015-06-07	2019-06-19
	#3	1974	0.06	16.21	1.33	213.37	1611.86	2015-04-09	2019-09-07
	#4	808	0.13	63.58	1.45	707.72	1826.86	2015-03-29	2020-03-29
	#5	1212	2.92	0.20	12.73	15.76	812.38	2015-01-14	2017-04-05
	#6	714	21.84	0.34	25.34	20.53	1017.26	2014-08-13	2017-05-26
	#7	296	45.14	18.14	39.27	223.37	1793.13	2014-08-14	2019-07-12
	#8	334	293.53	1.84	57.16	45.63	1446.91	2013-08-21	2017-08-07

Table 16 - aggregated statistics by cluster for the feedback network based on bookmarks (directed unweighted user-author bookmark network). All the values, except the size, are intended as the average of the values over each cluster.

With respect to the clusters for the comment network (Table 16), here we see that the users in the "superprosumer" cluster are less central, and the highest in-degree is obtained by users in the "superproducer" clusters.

Again, the most active groups of users tend to have longer periods of activity, but the difference is even less striking than for comments, and presents some exceptions, demonstrating that time is not a determinant factor.

3. Analysis of a fandom wiki community (Fandom.com)

3.1 Platform description

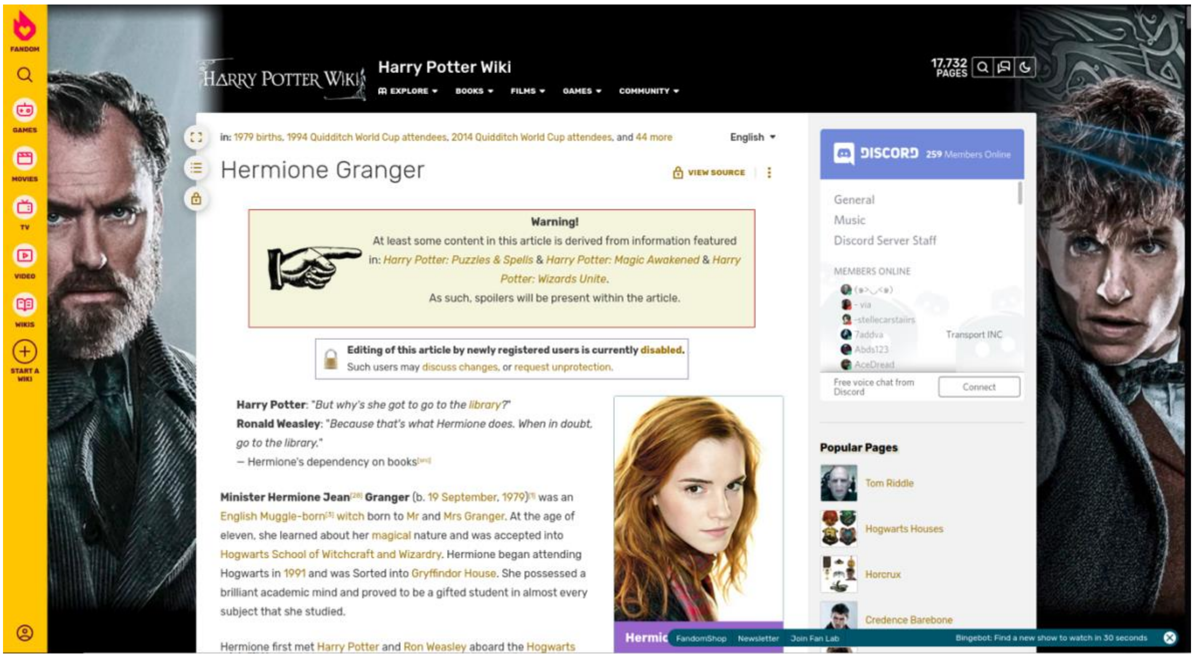
Fandom,[5] also known as Wikia before October 2016, is a web-hosting service that hosts wikis dedicated to entertainment. Fandom hosts wiki websites using MediaWiki, the same open-source wiki software used by Wikipedia.

Fandom, Inc., the company offering the service, was co-founded in 2004 by Jimmy Wales, also co-founder of Wikipedia, and by 2006 hosted approximately 1,500 wikis in 48 languages. Over time, Fandom has incorporated formerly independent wikis such as Uncyclopedia, a parody of the encyclopaedia Wikipedia, and WoWWiki, a website dedicated to the videogame “World of Warcraft.”

Fandom, Inc. derives its income from advertising and sold content, publishing most user-provided text under copyleft licenses. The company also runs the associated Fandom editorial project, offering pop-culture and gaming news.

In Fandom, each wiki can be seen as a kind of Wikipedia around a specific fictional universe, and pages are created to document characters, episodes, places, dates, concepts, or any other kind of element related to the fictional universe. As an example, Figure 18 shows the entry on Hermione, a popular character in Harry Potter books and films.

[5] Fandom official website, https://www.fandom.com, accessed on 2022-02-22.



in: 1979 births, 1994 Quidditch World Cup attendees, 2014 Quidditch World Cup attendees, and 44 more

Hermione Granger

Warning!
At least some content in this article is derived from information featured in: *Harry Potter: Puzzles & Spells* & *Harry Potter: Magic Awakened* & *Harry Potter: Wizards Unite*.
As such, spoilers will be present within the article.

Editing of this article by newly registered users is currently disabled.
Such users may discuss changes, or request unprotection.

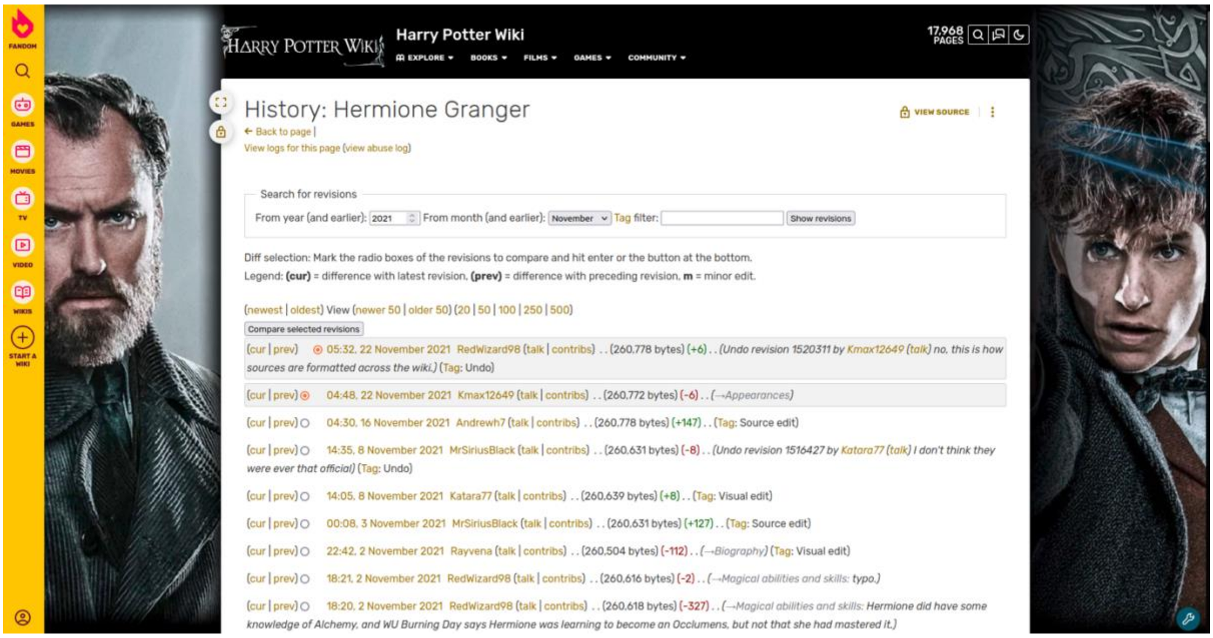
Harry Potter: "*But why's she got to go to the library?*"
Ronald Weasley: "*Because that's what Hermione does. When in doubt, go to the library.*"
— Hermione's dependency on books[src]

Minister Hermione Jean[28] **Granger** (b. 19 September, 1979)[1] was an English Muggle-born[3] witch born to Mr and Mrs Granger. At the age of eleven, she learned about her magical nature and was accepted into Hogwarts School of Witchcraft and Wizardry. Hermione began attending Hogwarts in 1991 and was Sorted into Gryffindor House. She possessed a brilliant academic mind and proved to be a gifted student in almost every subject that she studied.

Hermione first met Harry Potter and Ron Weasley aboard the Hogwarts

Figure 18 - Example wiki page from the Harry Potter Wiki on Fandom.com, the entry on "Hermione Granger", a popular character from the Harry Potter fictional universe.

The pages are edited by the users; for each page the edit history reporting the log of all activity is available. Figure 19 shows a snapshot of the edit history of the page dedicated to Hermione Granger.



History: Hermione Granger

← Back to page |
View logs for this page (view abuse log)

Search for revisions
From year (and earlier): 2021 From month (and earlier): November Tag filter: Show revisions

Diff selection: Mark the radio boxes of the revisions to compare and hit enter or the button at the bottom.
Legend: **(cur)** = difference with latest revision, **(prev)** = difference with preceding revision, **m** = minor edit.

(newest | oldest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)
Compare selected revisions

(cur | prev) ● 05:32, 22 November 2021 RedWizard98 (talk | contribs) . . (260,778 bytes) (+6) . . *(Undo revision 1520311 by Kmax12649 (talk) no, this is how sources are formatted across the wiki.)* (Tag: Undo)

(cur | prev) ● 04:48, 22 November 2021 Kmax12649 (talk | contribs) . . (260,772 bytes) (-6) . . *(→Appearances)*

(cur | prev) ○ 04:30, 16 November 2021 Andrewh7 (talk | contribs) . . (260,778 bytes) (+147) . . (Tag: Source edit)

(cur | prev) ○ 14:35, 8 November 2021 MrSiriusBlack (talk | contribs) . . (260,631 bytes) (-8) . . *(Undo revision 1516427 by Katara77 (talk) I don't think they were ever that official)* (Tag: Undo)

(cur | prev) ○ 14:05, 8 November 2021 Katara77 (talk | contribs) . . (260,639 bytes) (+8) . . (Tag: Visual edit)

(cur | prev) ○ 00:08, 3 November 2021 MrSiriusBlack (talk | contribs) . . (260,631 bytes) (+127) . . (Tag: Source edit)

(cur | prev) ○ 22:42, 2 November 2021 Rayvena (talk | contribs) . . (260,504 bytes) (-112) . . *(→Biography)* (Tag: Visual edit)

(cur | prev) ○ 18:21, 2 November 2021 RedWizard98 (talk | contribs) . . (260,616 bytes) (-2) . . *(→Magical abilities and skills: typo.)*

(cur | prev) ○ 18:20, 2 November 2021 RedWizard98 (talk | contribs) . . (260,618 bytes) (-327) . . *(→Magical abilities and skills: Hermione did have some knowledge of Alchemy, and WU Burning Day says Hermione was learning to become an Occlumens, but not that she had mastered it.)*

Figure 19 - Example edit history page from the Harry Potter Wiki on Fandom.com, for the entry on "Hermione Granger."

The pages in the wiki are divided into different namespaces. Namespace 0, or the main template, includes the entries that represent the most visible and visited part of the wiki. Other namespaces include "Talk", for wiki pages that can be used to discuss about entries, "User;" for the personal page of each user, "User talk", for pages that are used as a personal in-box for each user, or "Category" for defining categories of pages.

3.2 Basic dataset statistics

In this section, we present some basic statistics regarding Fandom wikis. For this study we considered 8 different communities, available at the following addresses

- **Marvel:** https://marvel.fandom.com
- **Harry Potter:** https://harrypotter.fandom.com/
- **Sherlock Holmes:** https://bakerstreet.fandom.com/
- **Lord of the Rings:** https://lotr.fandom.com/
- **Percy Jackson:** https://riordan.fandom.com/
- **Twilight:** https://twilightsaga.fandom.com/
- **Warriors:** https://warriors.fandom.com/

A bot is an automated or semi-automated software tool that a user can employ to carry out repetitive tasks such as correcting typos. Bots can make large amounts of edit, but they can also be disruptive if used incorrectly. Most wikis adopt policies on the usage of bots. One common rule is that bots need to be used with a separate account, i.e., not with the operator's main account, but with an account having the string “Bot” at the end of its name. For example if the operator is called "AUser" the policy encourage the use of the name "AUserBot" for their bot. With the name bot, we also indicate bot accounts.

In the following, we will present results both including and excluding bots (i.e., accounts whose name ends with the string “bot”). In general, we are interested in the contributions made by real (human) users, but bot edits are still part of the content of a wiki, and they can give important indications on the overall activity of its community.

The table below reports the size of each community in number of distinct users – without counting bots –, the overall number of edits (revisions), and the date of the first and last edit in our dataset for each wiki.

Wiki	Revisions*	Distinct users*	Date of first edit	Date of last edit
Marvel	3,831,537	17,624	2005-03-18	2021-07-15
Harry Potter	979,783	22,851	2005-07-05	2021-07-15
Sherlock Holmes	40,882	1,049	2003-03-07	2021-07-14
Lord of the Rings	175,602	5,010	2005-03-08	2021-04-04
Percy Jackson	293,198	9,142	2006-11-17	2021-07-15
Twilight	183,096	4,976	2007-12-19	2021-01-31
Warriors	585,599	10,027	2006-06-21	2021-02-01

Table 17 - Basic statistics for the 8 Fandom community selected. The asterisk (*) indicates that edits made by bot and bot accounts are not counted.

3.3 Community dynamics

3.3.1 Activity distribution

To investigate collective dynamics of co-creation in the wiki, we first look at the distribution of work among users and across pages. Below we present the distributions of users and pages by number of edits made and received, respectively.

User distribution by number of edits

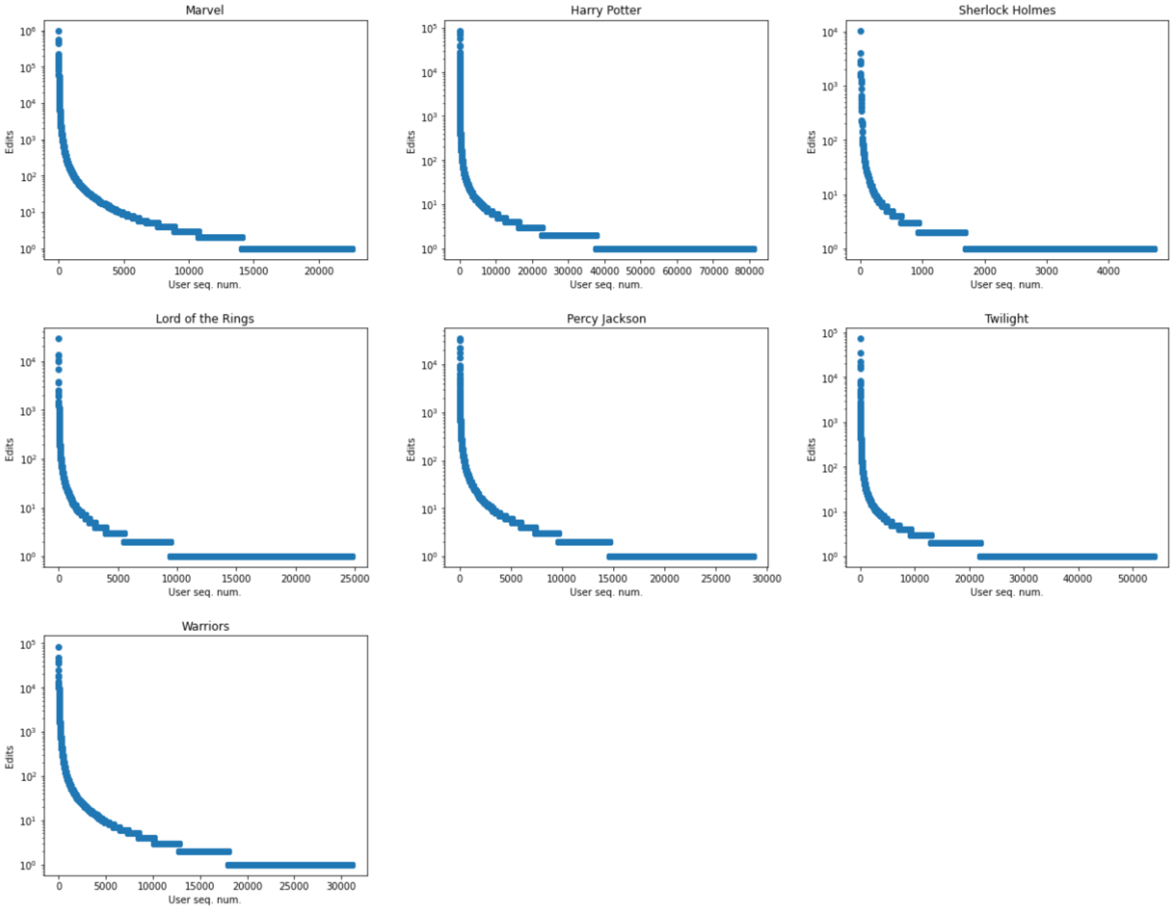


Figure 20 - Distribution of number users by number of edits, shown on a logarithmic axis, for each wiki.

As common in this kind of communities, we can see a heavy-tailed distribution: a few users get to perform tens of thousands of edits, and even hundreds of thousands in the largest communities, and many users perform very few edits.

Pages by number of edits

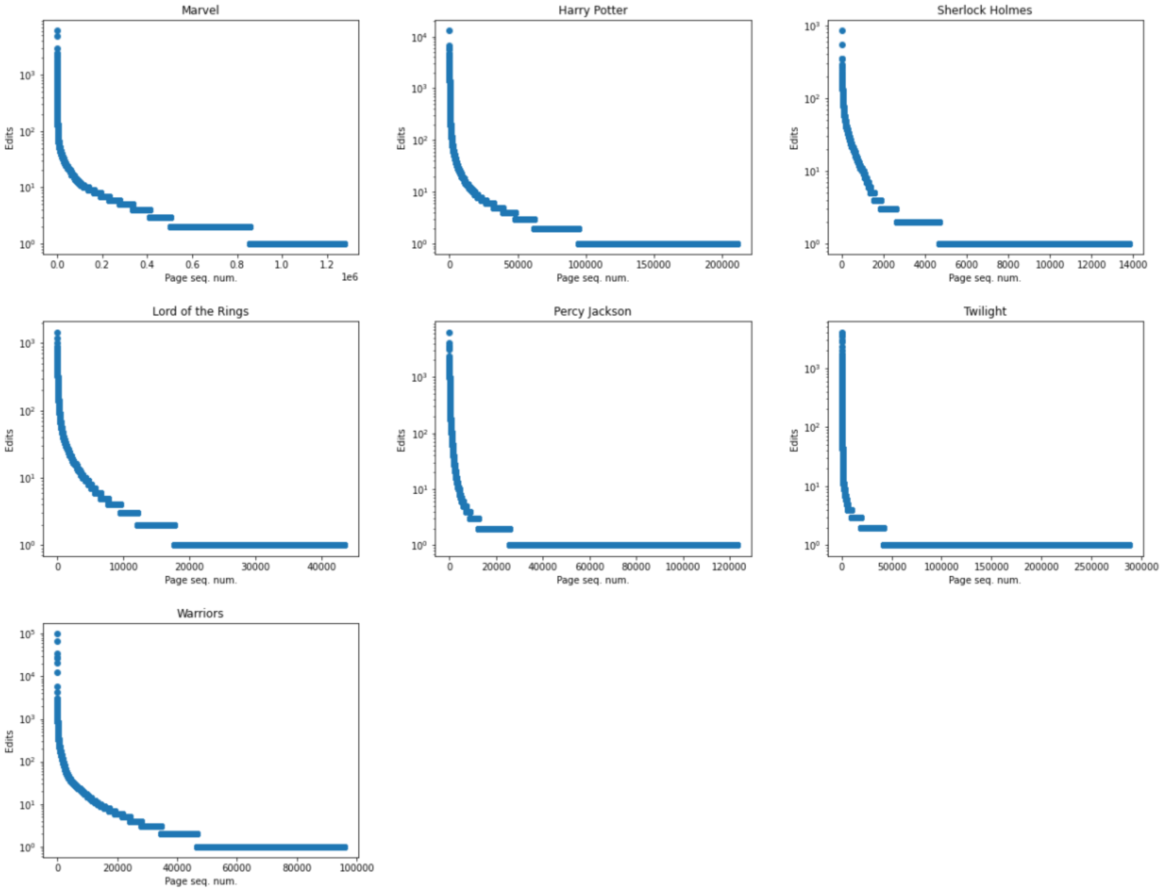


Figure 21 - Distribution of number of pages by number of edits for each wiki.

Also, for pages we find a skewed distribution, although to a lesser extent, with some pages attracting thousands or tens of thousands of edits, and many pages receiving few edits.

3.3.2 Activity types

To better understand community dynamics and organization of work in each community, we look at different kinds of activity in each wiki. We do this by looking at edits in different namespaces, i.e., different kinds of pages. The most visible part of the wiki is the main namespace, or namespace 0, that includes all the articles with the information shown to the reader. Other namespaces, less visible to the readers but accessible to anybody, are used for other aims such as coordination, discussion, personal communications between editors.

For simplifying the analysis, we group namespaces in a few categories re-adapting to our case the division proposed by Welser et al (2011) for Wikipedia:

- **Content** (namespaces 0, 6): main namespace, articles or entries which constitute the most visible part of the wiki, i.e., the information shown to the readers.
- **Content Talk** (namespaces 1, 7): talk pages, i.e., special pages used for discussion. Each talk page is associated to a wiki page from namespace 0 and is a space where users can discuss about the corresponding entry.
- **User and User Talk** (namespaces 2, 3): spaces for the users and user conversations; each user may have their own user page, with a description of their profile, activity and interests, and user talk page, where they can receive messages from other users.
- **Fandom wiki** (namespaces 4, 5): namespaces devoted to other technical aspects (e.g., templates, or pages through which content is automatically converted into a predefined format, such as infoboxes presenting structured information about entries of a certain category) or to coordination (e.g., for discussing community norms and policies, promoting projects, or electing administrators)
- **Other** (all the other namespaces).

Looking at how edits are distributed across namespaces is a proxy for how activity is distributed across these different kinds of activity in the wiki.

The figures below show the proportion of edits in different namespaces in each wiki. The graphs are obtained excluding bots; however, they would be quite similar also including bots, as the proportion changes only slightly.

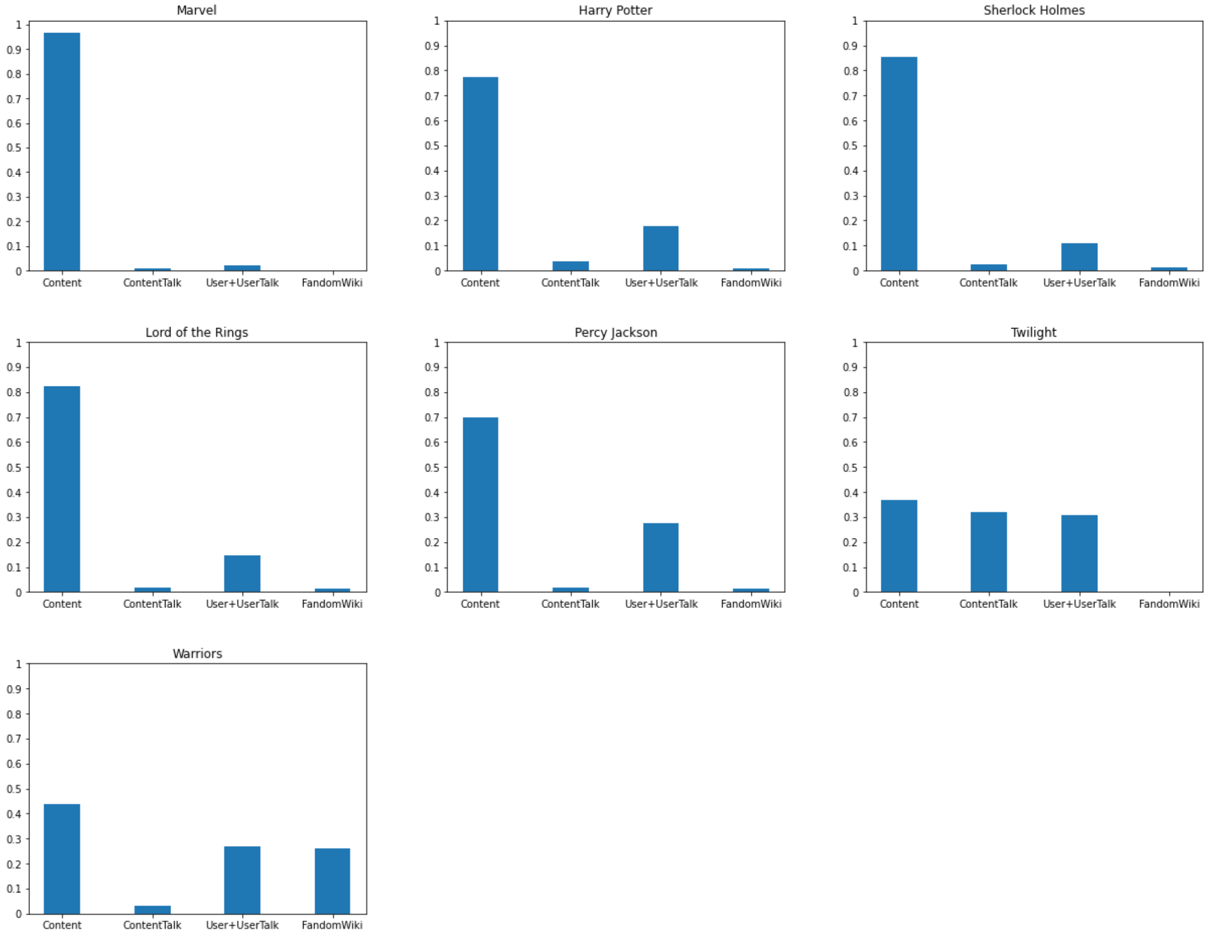


Figure 22 - Distribution of edits by namespace (without considering bot edits), showing the proportion with respect to the total number of edits.

Results are similar when we include or exclude bots and are quite different across wikis: in some communities, the large majority of edits are made in the Content namespace; in the case of Marvel, the proportion is above 95%. This indicates that, in these wikis, most of the effort of the editors is just devoted to writing content that readers will consult, with more or less activity happening in other namespaces. In Marvel, the very high proportion of edits in the Content namespace may be justified by the very high level of activity reported in Table 17 (over 3 million edits) made by a community of users that is in the same order of magnitude as other wikis' communities.

In other wikis, the activity in secondary namespaces has a much higher relative importance: especially in Twilight and Warriors we see that the amount of activity in the User and UserTalk namespaces is comparable to that observed in the Content namespace; moreover, in Twilight we have a comparable amount of activity also in the ContentTalk namespace, indicating discussion about content, and in Warriors in the FandomWiki namespace, indicating community spaces.

3.3.3 Activity evolution

In this section we look at how activity in each community evolves over time as the wiki grows, distinguishing edits in different namespaces to see the evolution of the effort in different kinds of activity. In the figures below, we present the time evolution of the number of monthly edits by namespace for each community, including and excluding bot edits.

Edits over time by namespace (all users)

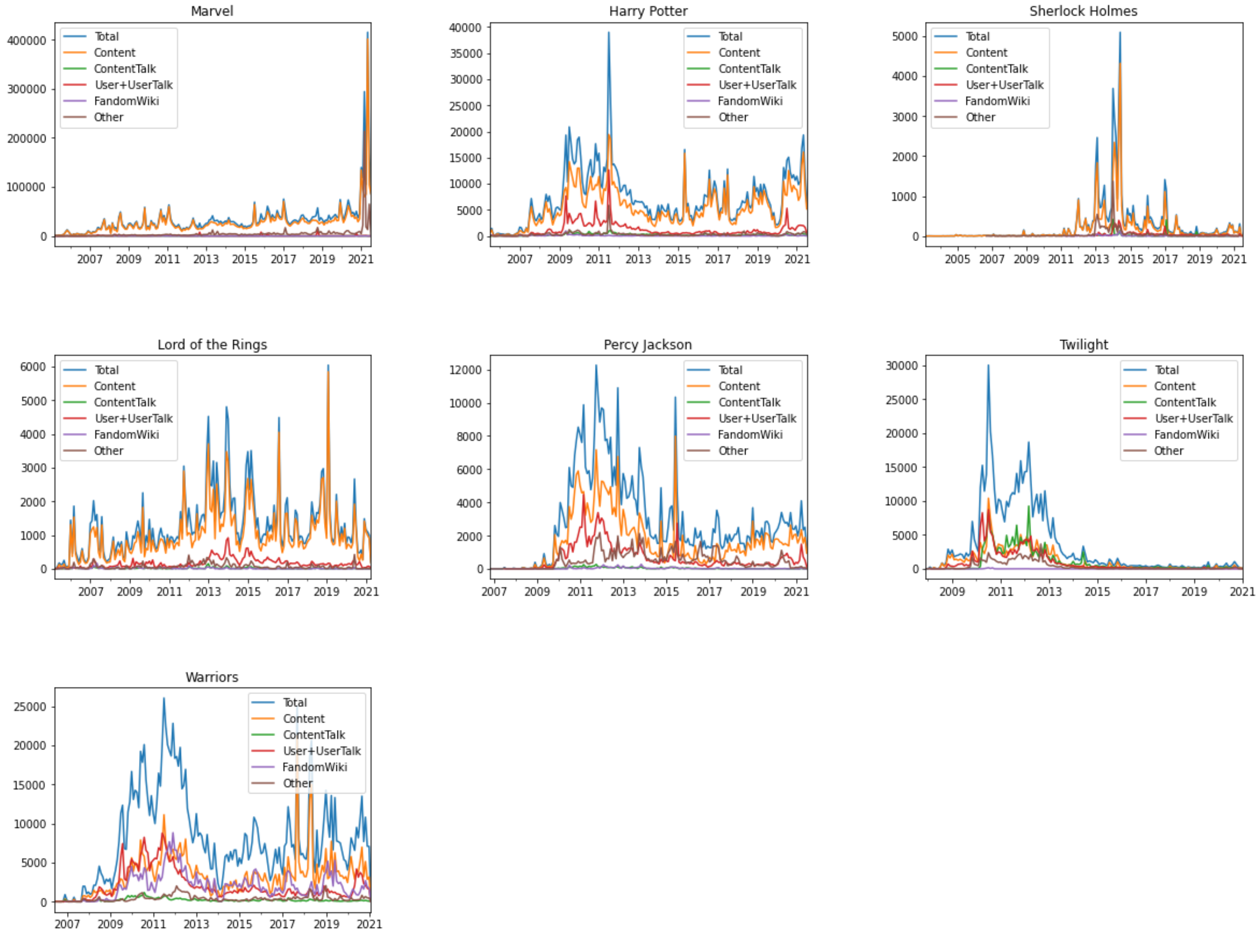


Figure 23 - Evolution over time of the number of edits per month by namespace for each community, considering also edits made by bots.

In the first figure (above), where bot edits, are included, we see some peaks of activity that are not present in the second figure (below), indicating that they are due to automated tasks performed by bots at a certain time, e.g., enacting some change on many pages of the same type at once. In particular, this is the case for Marvel wiki, which seems to have undergone a massive automatic editing of pages from different namespaces in 2021.

Edits over time by namespace (without bots)

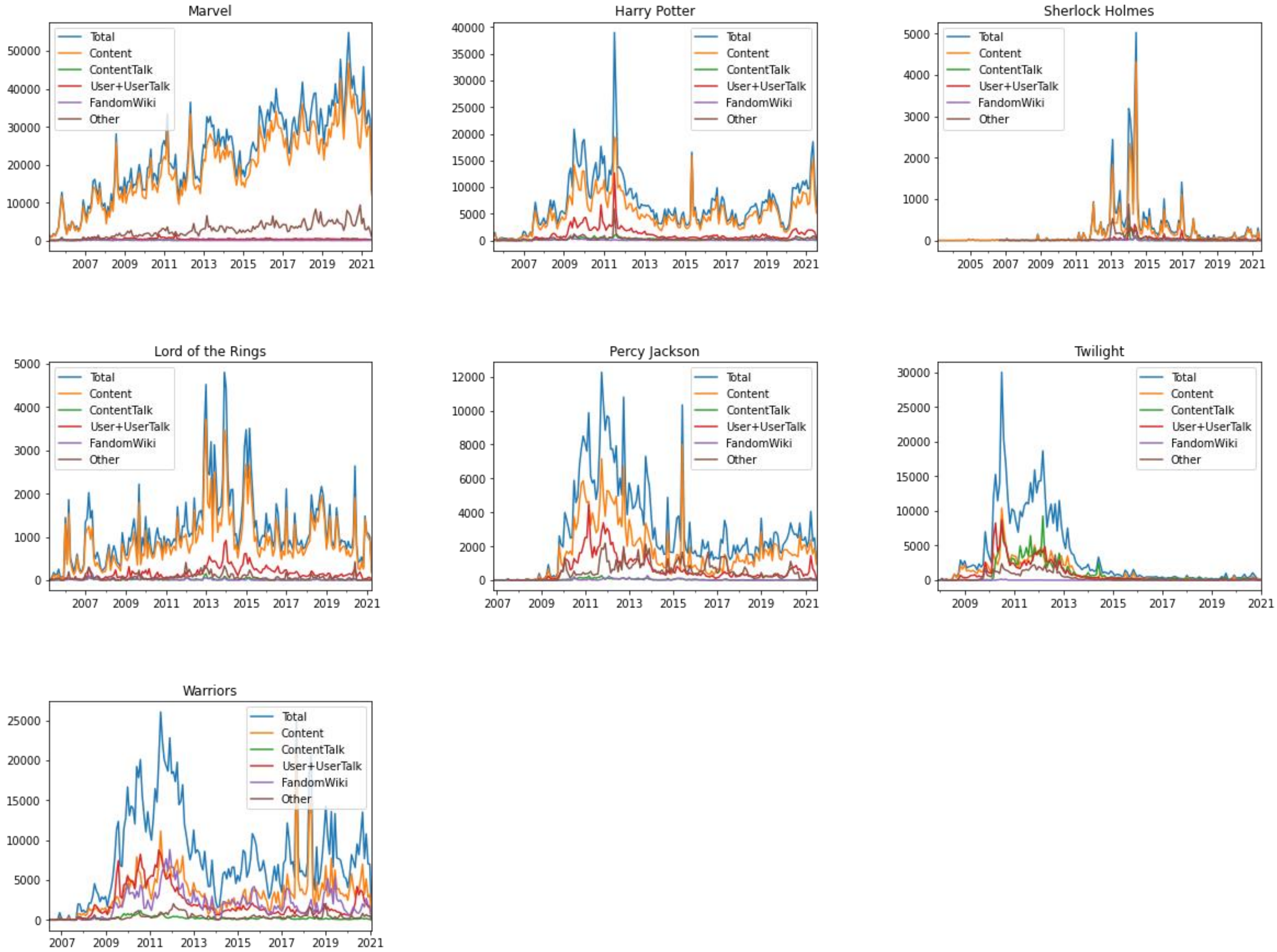


Figure 24 - Evolution over time of the number of edits per month by namespace for each community, without considering bot edits.

Therefore, we focus mostly on the second figure, Figure 24, where we can better see human activity. We observe that the wikis are in different phases. Marvel is the only community that seems to be still growing substantially in activity, following a relatively constant growth trend.

Harry Potter, Lord of the Rings, Percy Jackson and Warriors had an initial period of growth up to a peak of activity, then a decrease and then a stable or smooth growth pattern, showing that the communities are still well alive and active: we can hypothesize that the biggest effort for documenting most of the content has already been performed in the past. This would be in line with the low-hanging fruit hypothesis, according to which the topics that are easier or more straightforward to be described get increasingly covered, leaving less and less space for the creation of straightforward new content, and requiring a higher effort for further contributions. Such hypothesis, that has been proposed as one of the reasons for community stagnation or decline in Wikipedia (Gibbons et al, 2012; Collier & Bear, 2012) may be even more relevant in the Fandom context, where the scope of the topics to be covered is much more limited.

Sherlock Holmes and Twilight seem to have almost only residual activity, compared to activity in the past; we can think of various explanations for this. The fact that no new “canonical content” is being created for these fictional universes may explain why, once most of the topics corresponding to the “low-hanging fruits” have already been covered, only little activity is required. An interesting question for future work is to investigate whether the fans responsible for the creation of content in this wiki are now active on other platforms, or in other Fandom wikis corresponding to other fictional universes, livelier in terms of the current creation of new canonical and non-canonical content.

As expected from the previous section, the Content namespace receives most of the edits; it also has the highest peaks. However, we see that some peaks also occur for other namespaces, in particular for personal spaces and communications (User+User talk), and for discussion spaces (ContentTalk). The distribution of activity across namespaces seems to be mostly balanced over time, with periods of higher and lower activity reflected in a fairly even way across all the namespaces. We see remarkable levels of activity in namespaces User and UserTalk in most communities in some periods of time, that tend to coincide with the periods of maximum activity; interestingly, this seems to suggest that when there is more activity on a wiki, personal communication and interactions increase, not only (as it would be straightforward) in absolute terms, but also in proportion, suggesting that the need for personal communication grows in periods of higher activity.

3.4 Peaks of activity

In this section we cover a methodology for finding periods of augmented activities in each community. Here we decided to focus on periods of higher activity and of a certain duration, rather than on short spikes of activity. Therefore, we apply a rolling average and obtain a smoother version of the activity timeline for each wiki, then we look for peaks of a width of at least 6 months, representing longer periods of sustained activity in a community. For peak detection, we use the algorithm by Du et al (2006) implemented in the SciPy Python library. SciPy is a free and open-source Python library used for scientific computing.[6] In particular, the *find_peaks*[7] function takes an array of data, i.e., the succession of values over time and finds all local maxima by simple comparison of neighbouring values.

Edits in Content and in ContentTalk namespaces represent two parallel timelines that tend to be related to each other: in the former, users concurrently edit the articles co-creating content for the readers; in the latter, they discuss emerging issues, resolve conflict on the content on the articles, and coordinate the work for each specific article. To find periods of augmented activity in these two complementary dimensions at the same time, we consider both timelines, and look for their overlaps.

More concretely, for finding the windows of increased activity, we devised the following method:

1. Compute the rolling average of 12 months over the “Content “timeline.
2. Find the peaks with a width of at least 6 months.
3. Compute the rolling average of 12 months over the “Content Talk” timeline.
4. Find peaks with a width of at least 6 months.
5. Select the common peaks in both timelines (with an overlap window of 3 months on either side).
6. Select the top-10 pages by number of edits in the peak window.

Figure 25 presents an exemplification of the process for the Marvel wiki. In the figure we see the smoothed version of the timeline, i.e., we do not see the original data, but the rolling average on a window of 12 month. On these lines, we can see highlighted with a red background the time windows identified as peaks, with augmented levels of activity.

[6] SciPy official website, https://scipy.org/

[7] `scipy.signal.find_peaks`, https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find_peaks.html

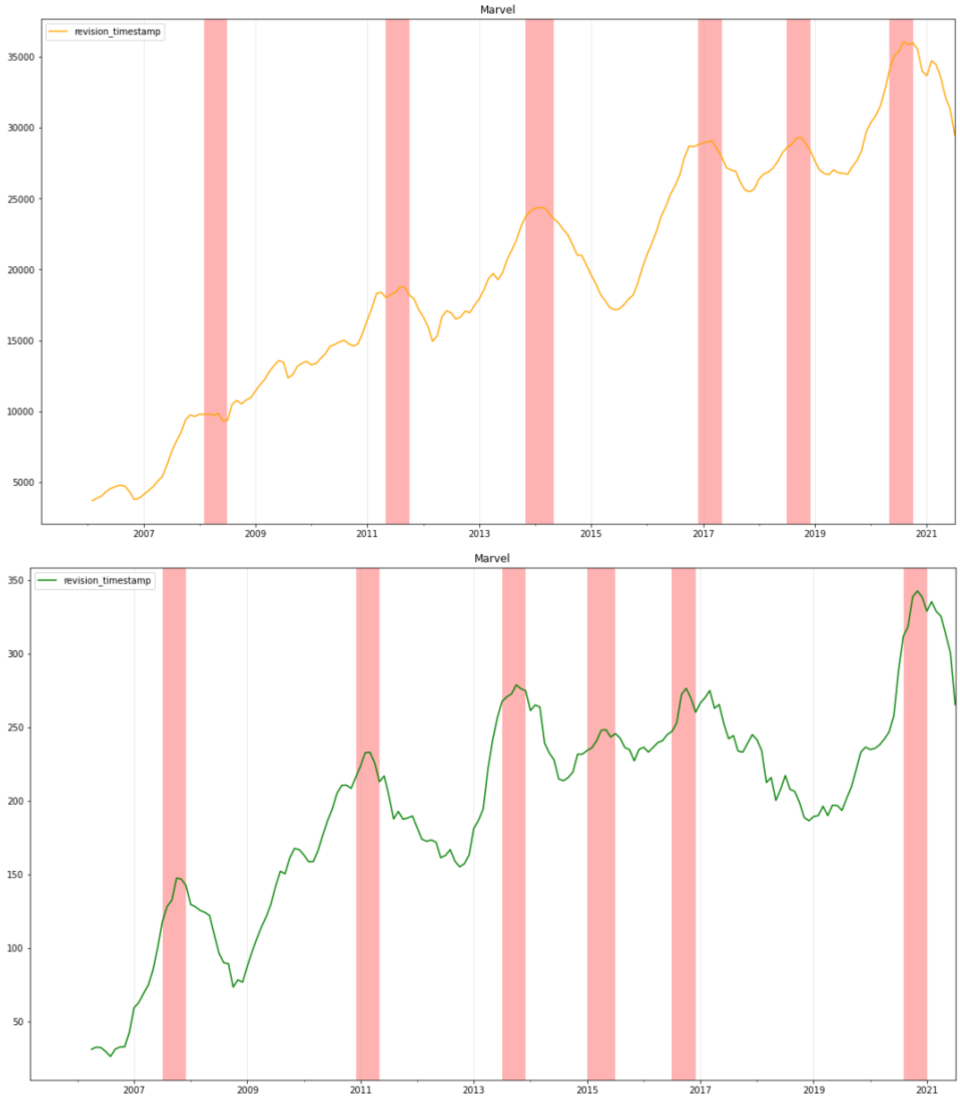


Figure 25 – Peak detection method. Example of the peaks found on Content pages (orange, up) and Content Talk (green, bottom) on the Marvel wiki.

To characterize each peak, we then select the most edited pages during the corresponding period. The tables shown below present the results with the major peaks detected, and the corresponding most edited articles during each peak, for the selected wikis. As these tables show, the most popular pages are related to the main characters of each fictional universe. Also when new movies or books are launched, the pages rising to the top of the list are the ones related to characters featured in the new publications.

Marvel

	Peak start date	Peak end date	Rank	Article title
1	2008-02-01	2008-07-30	1	Strength Scale
			2	Anthony Stark (Earth-199999)
			3	Iron Man (film)
			4	Secret Invasion
			5	Spider-Man
			6	Bartholomew Gallows (Earth-616)
			7	Iron Man Armor (Earth-199999)
			8	Jean Grey (Earth-616)
			9	Peter Parker
			10	Obadiah Stane (Earth-199999)
2	2011-05-03	2011-10-30	1	Bruce Banner (Earth-616)
			2	Peter Parker (Earth-616)
			3	Max Eisenhardt (Earth-616)
			4	Death of Spider-Man
			5	Thor (film)
			6	Miles Morales (Earth-1610)
			7	James Howlett (Earth-616)
			8	Henry McCoy (Earth-616)
			9	Steven Rogers (Earth-616)
			10	Thor Odinson (Earth-616)
3	2013-11-03	2014-05-02	1	Bruce Banner (Earth-616)
			2	Peter Parker (Earth-616)
			3	Thor Odinson (Earth-616)
			4	Scott Summers (Earth-616)/Gallery
			5	Otto Octavius (Earth-616)
			6	James Howlett (Earth-616)
			7	Captain America: The Winter Soldier
			8	Avengers (Earth-616)
			9	The Amazing Spider-Man 2 (film)
			10	Wade Wilson (Earth-616)
4	2016-12-01	2017-05-30	1	James Howlett (Earth-807128)
			2	Steven Rogers (Earth-616)
			3	Benjamin Reilly (Earth-616)
			4	Logan (film)
			5	Inhuman History
			6	Defenders Vol 1 64
			7	Defenders Vol 1 62
			8	Defenders Vol 1 63
			9	Peter Parker (Earth-616)

			10	Guardians of the Galaxy Vol. 2 (film)
5	2020-05-03	2020-10-30	1	Peter Parker (Earth-616)
			2	Bruce Banner (Earth-616)
			3	Triads (Earth-616)
			4	Character Index/Earth-Unknown-S
			5	Multiverse/Universe Listing
			6	Spider-Man 2099 Vol 3 4
			7	Anthony Stark (Earth-616)
			8	James Howlett (Earth-616)
			9	Empyre (Earth-616)
			10	Krakoa (Earth-616)

Table 18 - Top popular pages during the main peaks of activity detected on the Marvel wiki

Harry Potter

	Peak start date	Peak end date	Rank	Article title
1	2009-12-01	2010-05-30	1	Harry Potter
			2	Harry Potter and the Deathly Hallows: Part 2
			3	Hermione Granger
			4	Albus Dumbledore
			5	Ronald Weasley
			6	Severus Snape
			7	List of wizarding terms in translations of Harry Potter
			8	Filius Flitwick
			9	Sirius Black
			10	Bellatrix Lestrange
2	2011-07-03	2011-12-30	1	Harry Potter and the Deathly Hallows: Part 2
			2	Severus Snape
			3	LEGO Harry Potter: Years 5-7
			4	Battle of Hogwarts
			5	Bellatrix Lestrange
			6	Tom Riddle
			7	Harry Potter
			8	Minerva McGonagall
			9	Draco Malfoy
			10	Hermione Granger
3	2016-10-03	2017-04-01	1	Gellert Grindelwald
			2	Harry Potter
			3	Newton Scamander
			4	Chocolate Frog Card
			5	Credence Barebone
			6	Queenie Goldstein
			7	Obscurial
			8	Tom Riddle
			9	Fantastic Beasts and Where to Find Them (film)
			10	Porpentina Goldstein
4	2019-06-03	2019-11-30	1	Harry Potter: Hogwarts Mystery
			2	Tom Riddle
			3	Bellatrix Lestrange
			4	Albus Dumbledore
			5	Gellert Grindelwald
			6	Harry Potter
			7	Jacob's sibling
			8	Minerva McGonagall
			9	Severus Snape
			10	Gilderoy Lockhart

Table 19 - Popular pages detected during the main peaks of activity on the Harry Potter wiki.

Lord of the Rings

	Peak start date	Peak end date	Rank	Article title
1	2007-07-03	2007-12-30	1	The Return of the King (film)
			2	Witch-king of Angmar
			3	Aragorn II Elessar
			4	Battle of the Hornburg
			5	Gimli
			6	Goblin
			7	Battle of the Pelennor Fields
			8	Saruman
			9	The Lord of the Rings film trilogy
			10	Bilbo Baggins
2	2013-09-02	2014-03-01	1	Smaug
			2	Sauron
			3	Thranduil
			4	Tauriel
			5	Orcs
			6	Azog
			7	Dwarves
			8	The Hobbit: The Desolation of Smaug
			9	Fíli and Kíli
			10	Melkor
3	2015-05-03	2015-10-30	1	Bolg
			2	Hobbits
			3	Siege of Dale
			4	Sir Christopher Lee
			5	Battle of Five Armies
			6	One Ring
			7	Uin
			8	Nazgûl
			9	Azog
			10	Boromir
4	2018-12-01	2019-05-30	1	Tolkien (2019 film)
			2	Sauron
			3	Artamir
			4	Tom Bombadil
			5	Doors of Durin
			6	Nazgûl
			7	Fíli and Kíli
			8	The Fellowship of the Ring (film)
			9	Black Speech
			10	J.R.R. Tolkien

Table 20 - Top popular pages during the main peaks of activity detected on the Lord of the Rings wiki

Percy Jackson

	Peak start date	Peak end date	Rank	Article title
1	2012-05-03	2012-10-30	1	Percy Jackson
			2	Nico di Angelo
			3	Annabeth Chase
			4	Walt Stone
			5	Poseidon
			6	The Mark of Athena
			7	Camp Half-Blood
			8	The Serpent's Shadow
			9	Aphrodite
			10	Luke Castellan
2	2015-06-03	2015-11-30	1	Magnus Chase
			2	Samirah al-Abbas
			3	Hearthstone
			4	Percy Jackson
			5	Annabeth Chase
			6	Blitzen
			7	Zeus
			8	Sadie Kane
			9	Carter Kane
			10	Athena

Table 21 - Top popular pages during the main peaks of activity detected on the Percy Jackson wiki

Twilight

	Peak start date	Peak end date	Rank	Article title
1	2012-04-02	2012-09-29	1	Renesmee Cullen
			2	Bella Swan
			3	Breaking Dawn - Part 2
			4	Eclipse (film)
			5	Volturi
			6	Vampire
			7	Jacob Black
			8	Edward Cullen
			9	Alice Cullen
			10	Carlisle Cullen

Table 22 - Top popular pages during the main peaks of activity detected on the Twilight wiki

Warriors

	Peak start date	Peak end date	Rank	Article title
1	2015-12-02	2016-05-02	1	Alderheart
			2	Violetshine
			3	Jayfeather
			4	Prey
			5	Needletail
			6	Rileypool
			7	Firestar
			8	Mistakes in the Warriors Series
			9	The Apprentice's Quest
			10	Blue Whisker
2	2018-05-03	2018-10-02	1	Dovewing
			2	Medicine
			3	Firestar
			4	Jayfeather
			5	Fall of ShadowClan
			6	Scourge
			7	Crowfeather
			8	Bramblestar
			9	Briarlight
			10	Tawnypelt

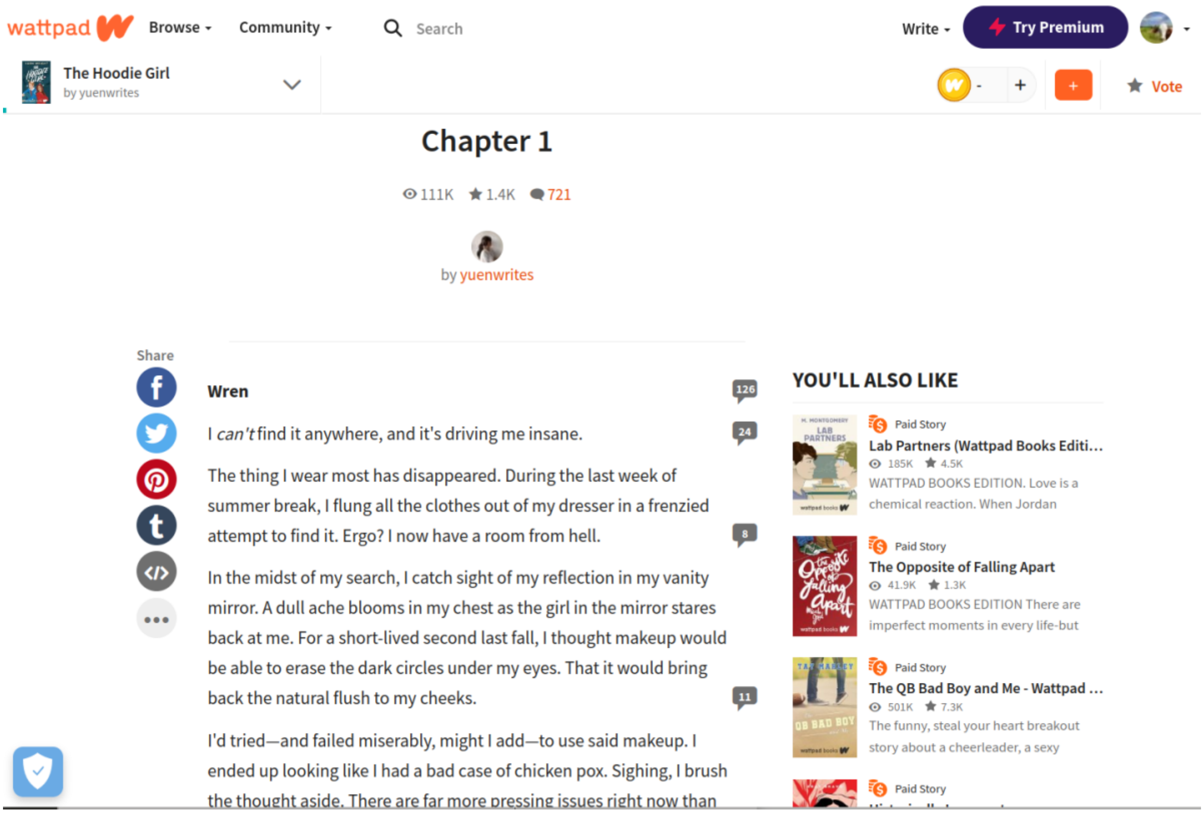
Table 23 - Top popular pages during the main peaks of activity detected on the Warriors wiki

4. Analysis of a social-reading community: Wattpad

4.1 Platform description

Wattpad is an online social reading platform intended for users to read and write original stories. The platform aims to create social communities around stories, allowing users to write and publish stories, or just read and comment stories generated by other users[8].

Comments can refer to specific paragraphs of a text, so that each paragraph has potentially a number of comments associated with it, as shown in the representing a screenshot from the popular Wattpad story "The Hoodie girl". Each comment can in turn receive replies from other users.


wattpad
Browse
Community
Search
Write
Try Premium
The Hoodie Girl
by yuenwrites
Vote
Chapter 1
111K
1.4K
721
by yuenwrites
Share
126
24
8
11
YOU'LL ALSO LIKE
Paid Story
Lab Partners (Wattpad Books Editi...
185K
4.5K
WATTPAD BOOKS EDITION. Love is a chemical reaction. When Jordan
Paid Story
The Opposite of Falling Apart
41.9K
1.3K
WATTPAD BOOKS EDITION There are imperfect moments in every life-but
Paid Story
The QB Bad Boy and Me - Wattpad ...
501K
7.3K
The funny, steal your heart breakout story about a cheerleader, a sexy
Paid Story

Wren

I *can't* find it anywhere, and it's driving me insane.

The thing I wear most has disappeared. During the last week of summer break, I flung all the clothes out of my dresser in a frenzied attempt to find it. Ergo? I now have a room from hell.

In the midst of my search, I catch sight of my reflection in my vanity mirror. A dull ache blooms in my chest as the girl in the mirror stares back at me. For a short-lived second last fall, I thought makeup would be able to erase the dark circles under my eyes. That it would bring back the natural flush to my cheeks.

I'd tried—and failed miserably, might I add—to use said makeup. I ended up looking like I had a bad case of chicken pox. Sighing, I brush the thought aside. There are far more pressing issues right now than

Figure 26 - Example screenshot from a popular book on Wattpad ("The Hoodie Girl").

[8] https://www.wattpad.com/writers/

In this section we present an analysis of the Wattpad dataset from Pianzola et al (2021). The dataset is formed by a selection of 15 books from two different categories: young adult literature created within the Wattpad platform (from now on simply “teen” or “young adult”) and classics from traditional literature imported into Wattpad (from now on simply “classics”). These two categories were selected by Pianzola et al (2021) as they can be considered representative of popular and prestigious literature respectively, according to common opinions in literary studies (Underwood, 2019). The two categories are very different in many aspects: the former is one of the most popular categories of “Wattpad native” literature, and may be seen in some way as a typical expression of the spontaneous creative culture of the platform, while the latter is made of classical books no longer covered by copyright, that are not “native” in Wattpad and have just been included in the social reading platform, where they can be commented paragraph by paragraph like the users' creations. This analysis aims to shed light on the opinion dynamics in the platforms, basically focusing on the interactions, namely comments and replies, that users have around the works as well as the sentiments and emotions of the comments.

4.2 Basic dataset statistics

The dataset is formed by a selection of 15 books from two different categories, classic and young-adult literature. The dataset contains different fields that provide information about the interactions of users with the book. The fields are described in

Table 24. The dataset contains in total 125,021,688 items. Each data item describes an interaction of a user with a book. More precisely with a particular part of a book, either chapter and/or paragraph. Each interaction can be a comment, that can be spontaneous, or a reply to a previous comment. In some cases, an interaction (data item) may be not a comment but another kind of action; in these cases, we discarded the corresponding data items.

Table 24 - Fields for each interaction in the Wattpad dataset

Wattpad dataset fields	
uid	Unique identifier (Integer)
bookID	Book identifier (Integer)
chapterID	Chapter identifier (Integer)
paragraphID	Paragraph identifier (Integer)
book	Book name (String)
chapter	Chapter name (String)
paragraph	Paragraph text (String)
username	Unique anonymized code by user (String)
date	Date interaction (String)
comment	Comment text (String)
reply	Reply indicator (Boolean)

In general, the books created in Wattpad have more comments on the platform than the classic literature books. *The Hoodie Girl* and *The Bad Boy's Girl* are the ones with more interactions, and we will focus an important part of this analysis in those books with a lot of interactions. On the other hand, books like *Anna Karenina* or *Emma* have very little interactions. This makes sense as classic literature is mostly read out of the Wattpad platform, while popular literature originally created in Wattpad may typically generate a large amount of activity and feedback on the platform. In Table 25 we summarize some numbers about comments by book.

Table 25 - Summary of comments and replies to comments for each interaction in the Wattpad dataset

Book	Comments	Replies
Alice's Adventures in Wonderland (1865)	2.528	660
Anna Karenina (1877)	1.075	253
Emma (1815)	1.785	405
I Sold Myself to the Devil for Vinyls... Pitiful I Know	1.797.839	374.614
Jane Eyre (1847)	4.922	1.048
My Brother's Best Friend	709.110	166.523
Pride and Prejudice (1813)	35.820	10.842
Romeo and Juliet (1597)	8.273	2.713
She's With Me	1.501.141	301.668
The Bad Boy Stole My Bra	46.617	7.768
The Bad Boy's Girl (Now Available as a Paperback and ebook)	2.591.067	541.416
The Bad Boy, Cupid & Me	1.006.862	174.244
The Cell Phone Swap	564.695	117.572
The Hoodie Girl	3.055.798	766.578
Wuthering Heights (1847)	5.660	2.385

Error! Reference source not found. shows the distribution of comments over time considering that we have data from end-2014 to mid-2018. We detect some peaks of interaction in “The Hoodie Girl”, we will analyse these peaks later.

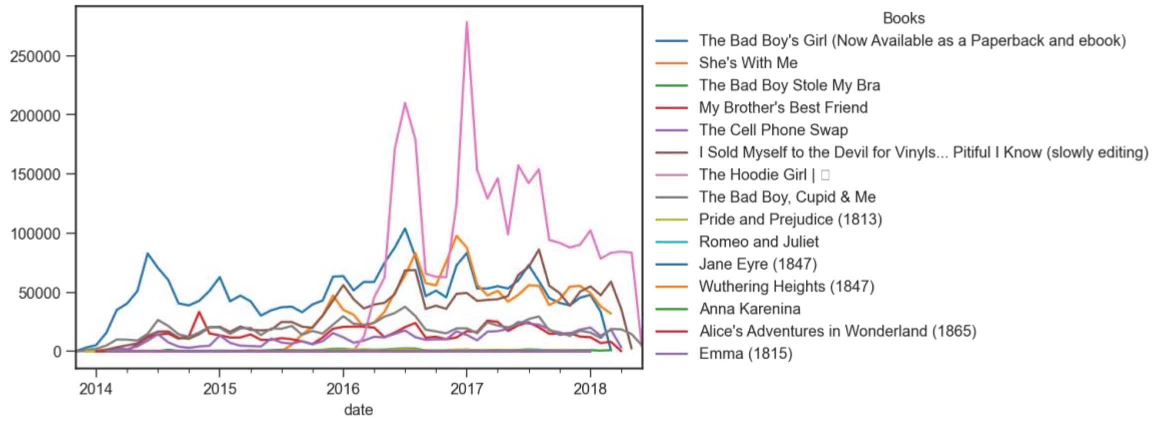


Figure 27- Evolution over time of comments and replies, aggregated by month, for each book.

4.3 Evolution of comments

4.3.1 Comments by chapter

In general, we see some clear pattern in the distribution of comments by chapter. Most of the books have a lot of comments in their initial chapters: *Emma, Alice's Adventures in Wonderland, Anna Karenina, Wuthering Heights, Romeo and Juliet, Pride and Prejudice, The Bad Boy, Cupid & Me, The Hoodie Girl* are some examples.

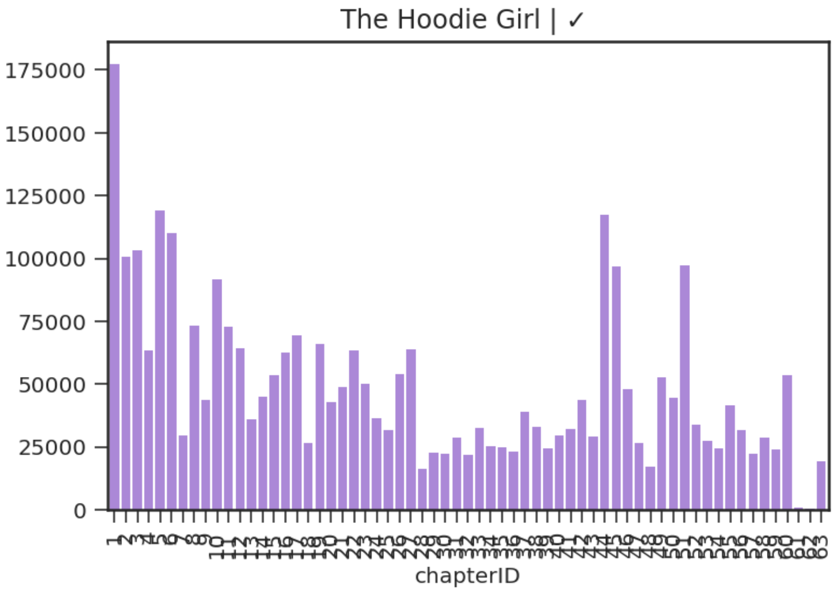


Figure 28 - Distribution of comments by chapter in The Hoodie Girl

Janey Eyre has the most of comments in Chapter I, although the peak is in Chapter XXIV in the middle of the book. *I sold myself to the Devil for Vinyl's...* and *The Bad Boy Stole My Bra* have the peak in the last chapters; probably the users are discussing about the ending of the book.

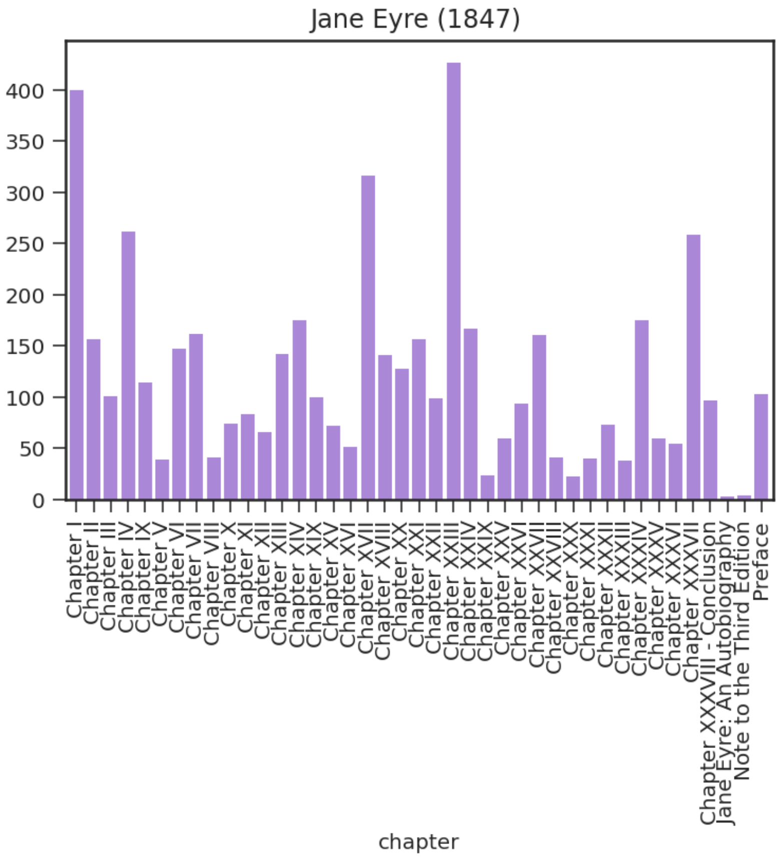


Figure 29 - Distribution of comments by chapter in Jane Eyre

The Cell Phone Swap, My Brother's Best Friend and The Bad Boy's Girl have most of the comments in chapters in the middle of the book. And *She's With Me* has some peaks in the beginning, in the middle and in the end of the book.

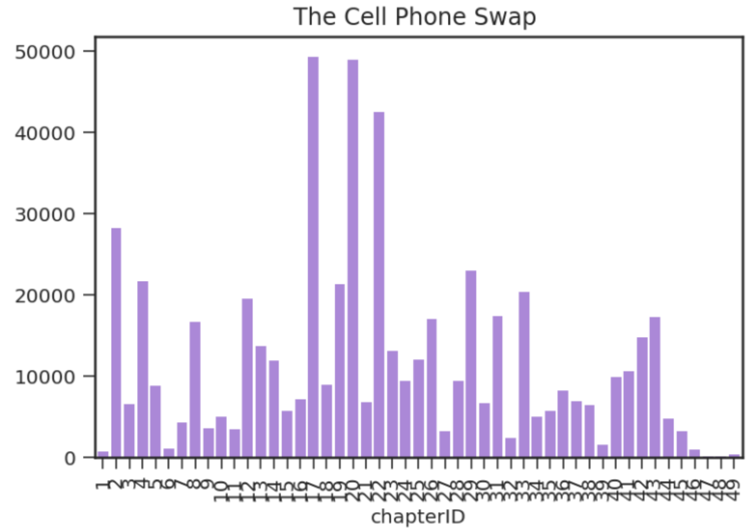


Figure 30 - Distribution of comments by chapter in The Cell Phone Swap

Clearly it seems that there is a substantial difference between the traditional literature books and the books created in the platform. We suppose that this difference in interactions is due to the nature of the works: Wattpad-native works are created serially, whereas other works are already complete at the time of publishing on the platform.

Next steps. *We could perform some extra analyses to better understand the discussion (positive or negative) on the chapters with most comments, or for the most relevant topics in each chapter.*

4.3.2 Comments by user

Within our sample there are some users, named *bf1e94e9*, *53c7453b* and *4f169451*, that accumulated a lot of comments across the books we examined. We carried out two different kinds of analyses across all books:

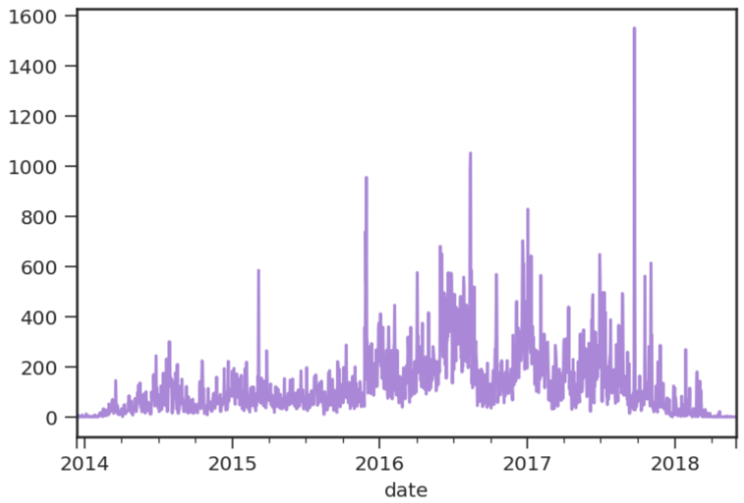


Figure 31 – Evolution of daily comments of author bf1e94e9, who has been commenting in all the books

The author *bf1e94e9* has contributed most of the days since 2014 (see Figure 31) with peaks of activity of more than 1,500 comments per day. The mean of comments is 122 per day and the superuser has commented all fifteen books on the database.

4.3.3 Comments by book

The Hoodie Girl is one of the books with more interactions in our dataset (more than 3,000,000). From our exploratory analysis we have detected that there are some peaks of interactions in early 2017. In general, most of the comments are concentrated in the first paragraph. It might be due to different reasons, maybe because the readers write general comments about the book in this first paragraph, maybe because the content of the paragraph for some reason is more susceptible to be commented or just because the first chapter / paragraph has potentially the more readers and more visibility than any other.

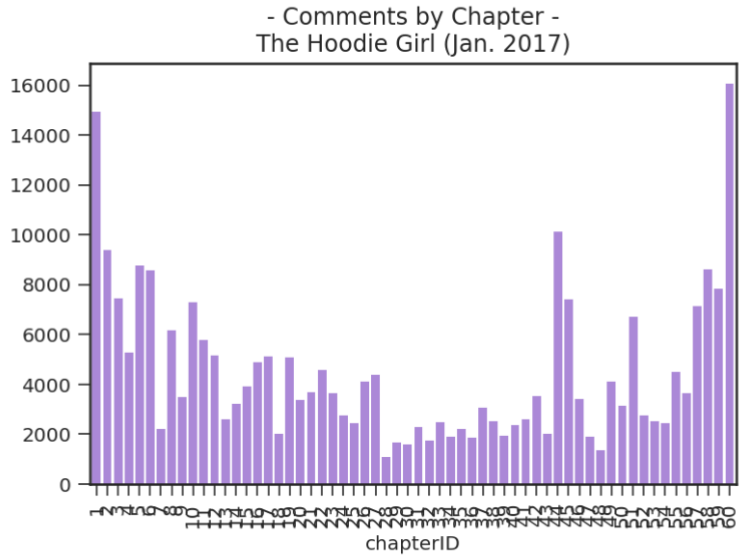


Figure 32 - Evolution of comments by chapter in The Hoodie Girl during the January 2017

From the distribution of user contributions in *The Hoodie Girl*, we can see how there are some users that generated most of the comments. Currently we cannot explain the existence of these superusers. We hypothesise that they might correspond to administrators or aggregated anonymous and/or guest users.

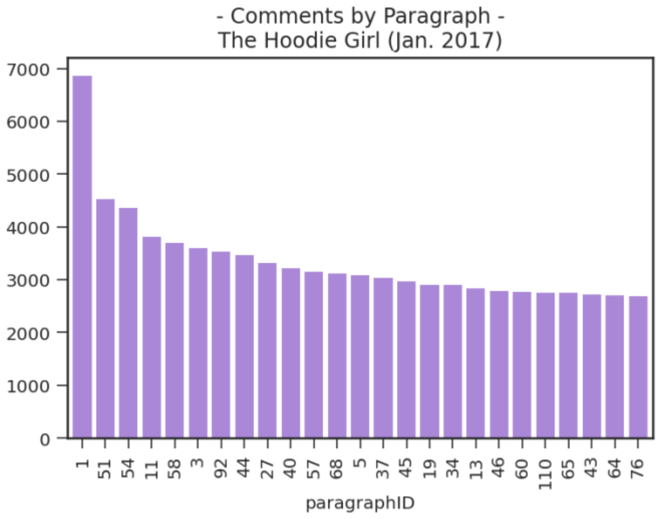


Figure 33 - Distribution of comments by paragraph in The Hoodie Girl during January 2017

4.4 Comments analysis

To analyse the content of the messages, we performed different approaches to text analysis that shed light in different aspects like sentiments or emotions. We use LIWC (Tausczik & Pennebaker, 2010) which reads a given text and counts the percentage of words that reflect different emotions, thinking styles, social concerns, and even parts of speech. And secondly, we use NRC[9] to measure the emotion intensity. Potentially, in further analysis, we could complete a comparative analysis using VADER[10] (Valence Aware Dictionary and sEntiment Reasoner), a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media which is fully open sourced and available in NLTK.

In general, we focus our analysis on pairwise comparison with word shift graphs (Gallagher, R. J., et al., 2021) between different works[11]. We compare the pair of works with most interaction from young literature and classic literature respectively. The most popular works of young literature are: *The Hoodie Girl* and *The Bad Boy's Girl*, and the most popular in classic literature *are Pride and Prejudice* and *Jane Eyre*.

4.4.1 Linguistic, psychological, and social processes (LIWC)

To analyse linguistic, psychological, and social processes in the comments to a book, we use LIWC (Tausczik & Pennebaker, 2010), an established tool for emotion and language analysis, based on a dictionary that associates frequent words to 80 different linguistic categories (e.g., positive emotion, function words, social words, cognitive process, etc.).

First, we present the textual analysis on the two young literature books. We compute the Shannon entropy shifts which measure the surprise or unpredictability of the words; the entropy of the comments for the two books is practically the same (5.20 vs. 5.22), *Bad Boy's Girl* comments are slightly more unpredictable. If we focus on the top 10 words, we can observe from the pairwise plot that for *The Hoodie Girl* (compared to *The Bad Boy's Girl)* the first person (I) is the most surprising word category, followed by Work and Percept. Percept indicates

[9] http://saifmohammad.com/WebPages/lexicons.html

[10] https://www.nltk.org/_modules/nltk/sentiment/vader.html

[11] https://shifterator.readthedocs.io/en/latest/index.html

perceptual process like seeing, hearing, and feeling. We observe in the pairwise comparation that See is the most surprising perception. Conversely, the most unpredictable words in *The Bad Boy's Girl* (in comparation with *The Hoodie Girl*) are social words (family, friends, female, or male referents), in terms of personal pronouns here the most relevant are the third and second person singular.

Regarding the classics, the Shannon entropy of both book is very similar, 4.83 and 4.84. What we can observe very clear difference in words. In *Pride and Prejudice*, we find comments on relativity (time, space, motion etc.), drives (affiliation, power, risk, etc.) and positive emotions. Regarding *Jane Eyre* the words with higher entropy shift are about time orientation (present) and, in terms of grammar, verbs. Also worthy of note the presence of sexuality and death.

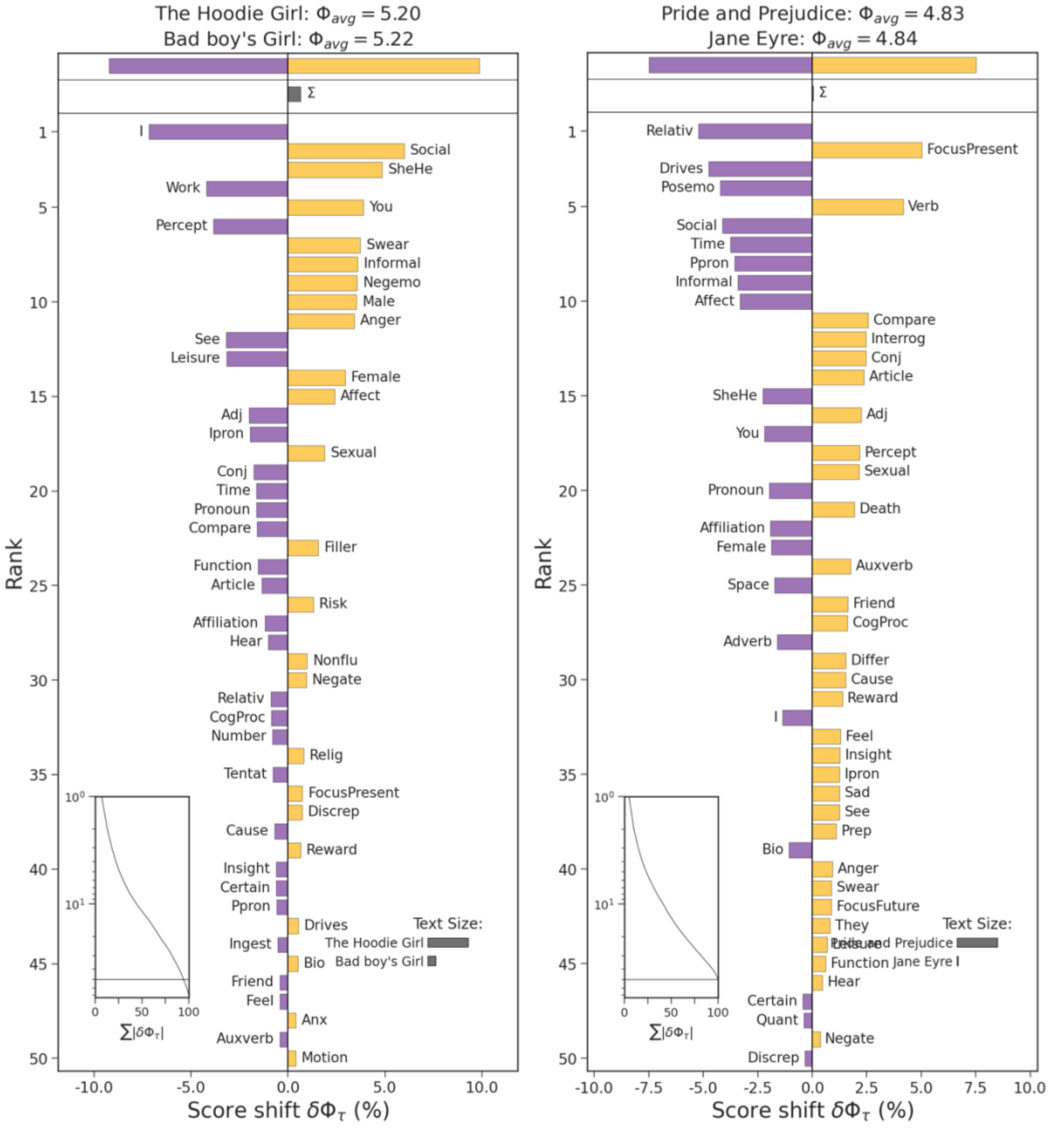


Figure 34 - Pairwise comparation of Shannon entropy shifts on LIWC dimensions for The Hoodie Girl and Bad boy's Girl (left) and Pride and Prejudice and Jane Eyre (right)

4.4.2 NRC-Emotion

The NRC Emotion Lexicon (Mohammad, S. M., & Turney, P. D, 2013) is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive).

The annotations were manually created via crowdsourcing. We applied emotions lexicon to *The Hoodie Girl* and *Bad Boy's Girl* and represented it in Figure 35, one can see the comparison of two emotions: anger and joy. The word shifts show the top 25 words contributing to the difference in emotion: on the right column of each graph, the words that contribute to the first book (The Hoodie Girl) have a higher score in joy/anger than the second (Bad Boy's Girl), while words in the left column of each graph on the contrary contribute to make the first book have fewer expressions of joy and anger than the second. Bars at the top show the overall difference and the effect of each type of word contribution on that difference.

In general, we can say that in *Bad Boy's Girl* there is stronger emotional expression, with both more positive and more negative words. For instance, regarding **joy** we can see *laughing* (2nd) and *perfect* (3rd) as being more frequent and having more impact than *beautiful* (5th) or *favorite* (13th), which are instead more frequent in The Hoodie Girl. If we look at the comparison for **anger,** we observe how terms like *hate* (2nd) and *bitch* (3rd), that are in the top of the list, contribute a lot on anger.

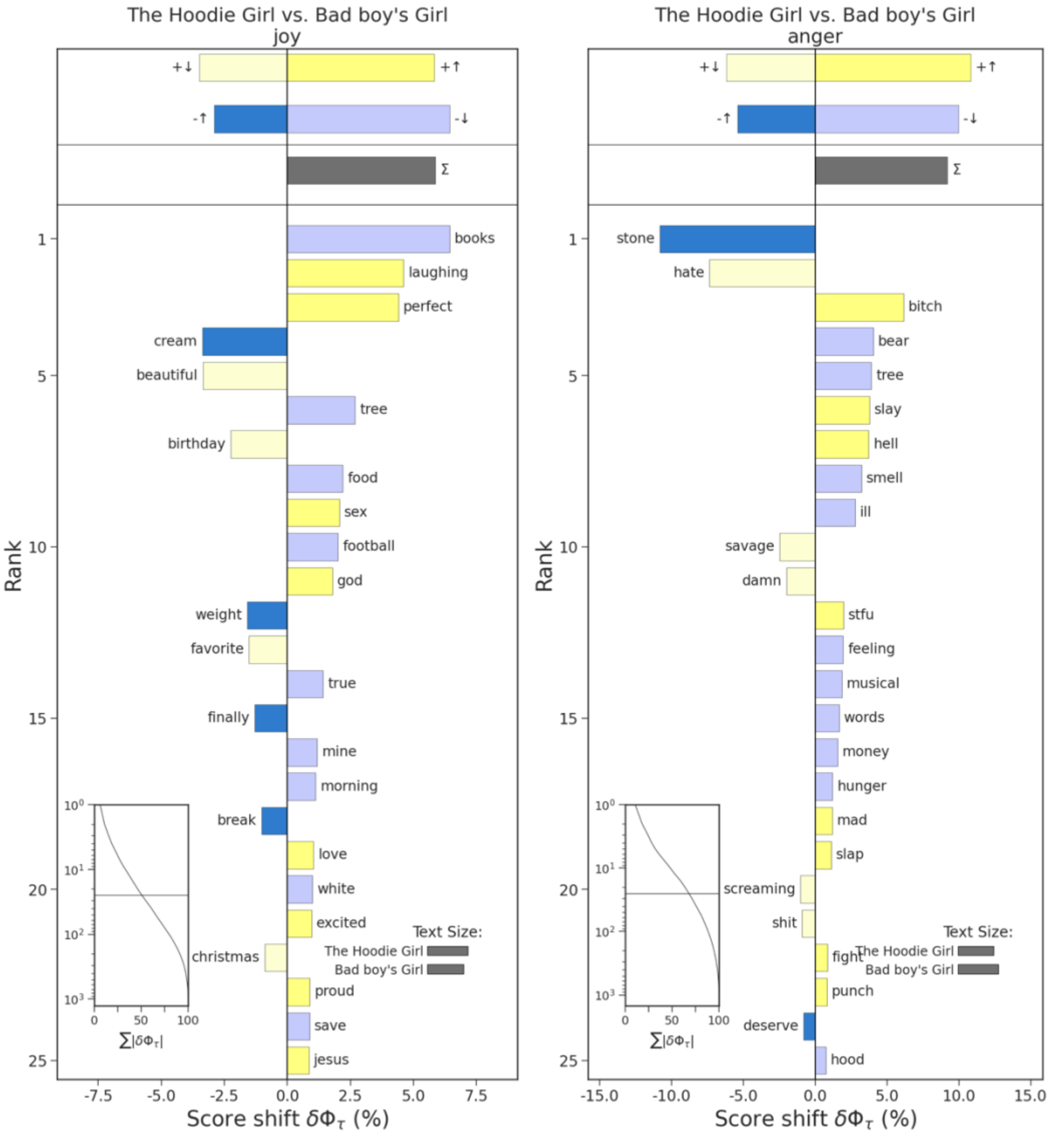


Figure 35 - Pairwise comparation of emotions, joy (left) and anger (right) for the books The Hoodie Girl and Bad Boy's Girl as a word shift graph.

This analysis has been extended to all the emotions included in the dictionary: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust (see Annex). NRC lexicon also could provide the analysis of affects, based on valence, arousal, and dominance which has not been implemented in this report.

5. Conclusions and future work

In this document, we have presented the results of our analysis of data from existing prosumer communities, setting the bases for the development of the Prosumer Intelligence Toolkit.

We have focused on data from three successful prosumer platforms in which users co-create content, for which we managed to access digital traces of user interactions from thousands of users. With these three platforms we have explored three different contexts in which prosumers show their creative potential: a fanfiction community where users create works inspired by existing fictional universes and review one another's work (AO3); a fandom wiki, where users create content collaborating on editing wiki pages to document any element related to a fictional universe; and Wattpad, where users create stories (online books) and read and comment on one another's work.

5.1 Summary of findings

In **AO3**, given the availability of a large volume of fine-grained data from user interactions, we have performed a more in-depth analysis, in two main directions.

First, we have analysed **popularity dynamics**, studying how content gets popular over time, and we have developed a model to predict works that will become very popular in the near future, based on previous history. We have further re-adapted the model to apply it to tags and predict topics that will become popular. Thus, we respond to the most recurrent concern we collected from publishers and stakeholders, that of predicting popular content and topics. This may help to reveal trends in the interests of readers and writers, and to identify valuable content to be considered for publishing: ideas and elements developed and tested on one popular fanfiction platform could presumably be applied to other prosumer communities with similar characteristics.

Moreover, we have investigated **social interactions** between AO3 users, modelling them as graphs, to study structural properties of the social networks resulting from different kinds of interactions (replies, or feedback by comments or bookmarks) and to study the centrality and the role of the users in the community. In particular, we have characterized each user by the combination of their centrality as a producer (feedback received as an author) and as an active consumer (feedback given as a reader) in the network; we have performed clustering to identify different profiles of prosumers along these two dimensions. We found consistent clusters across the major communities, which seems to give some robustness to the different emerging profiles, among which the ones we dubbed *superproducers*, *superconsumers*, and *superprosumers,* as the users who have the highest levels of centrality in one of the two dimensions, or in both (in the case of the latter). We believe this "map" of prosumer roles may be helpful to understand the composition of a community and to identify relevant users for specific aims. We have further offered a characterization of the users in each of these categories, based on their aggregated statistics. The availability of demographic attributes or other relevant characteristics of the users would help to enrich this model, and the knowledge we have about the different profiles and kinds of prosumers.

In **Fandom**, we have focused on **collective dynamics**, studying how activity on different tasks and spaces evolves over time and in different phases of community growth, and for different communities. We have further investigated peaks of activity and their nature, identifying pages undergoing the highest levels of activity in the corresponding periods of time.

In **Wattpad**, the popular social reading platform, we have studied the dynamics and characterized the **language and emotion** of a sample of very popular books from two very diverse categories: teen literature, and classics. We have looked at the distribution of comments and found a tendency to have more activity on the first and last chapters of a book, and we have shown how language in the feedback around each book can be characterized through Shannon divergence, and through different tools for emotional analysis. We have also shown how emotions in the feedback on different books can be compared, highlighting the words that are more important to make this difference, for their marked difference in frequency between the two books. As we have only recently got access to this dataset, and we had little time to perform this analysis, in this deliverable we have presented a preliminary exploration of the data, that we plan to extend in several directions.

5.2 Limitations and lines for future research

The work presents some limitations, mainly due to the limited availability of data in relation to some needs expressed by project partners and stakeholders. As discussed in the Introduction, we identified several kinds of data inaccessible to us: demographic data, complete records of activity attached to unique users, navigation behaviour, click and reading patterns on platforms, and social media data from closed platforms such as TikTok or Instagram. Access to such data would open up promising directions for enriching our analysis and metrics (as clearly testified by interviews with publishers); however, its feasibility is not straightforward.

In the case of collaboration with a company or organization running a platform, or of the development a new platform, it might become possible, with the permission of the users, to collect demographic data, navigation patterns, and behaviour data across the whole platform. The latter could also be achieved on pre-existing platforms using different scraping criteria focused on users, although this could raise privacy issues, as discussed above.

Other than developments associated with data access, there are several potential developments that emerged from conversations with partners and stakeholders, and from the analyses performed and results obtained so far.

Regarding our popularity prediction model, other features could improve our current approach; i.e., some machine learning model could be developed to detect content that will become popular that also includes elements such as the profile and previous record of activity of the author, or features automatically extracted from the text such as topics, emotions, characters and entities, and linguistic styles, quality, and diversity.

Regarding social interactions, in this work we have modelled each community as a social network; it would be interesting to model the conversation around each work as a social network (reply network) to characterize the patterns of discussion around it. This way, it would be possible to identify the kind of discussions around different works, and to investigate the role of the author in the discussion on their creation.

Regarding community analysis, we have looked at peaks of activity of a certain time width, in order to identify periods with higher and sustained levels of activity. Other methods from previous literature on peak detection in social media (Lehmann et al, 2012) and specifically on Wikipedia (Kaltenbrunner and Laniado, 2012) could be applied to identify more punctual peaks which would help to investigate the relation with specific external or internal community events that provoke rapid spikes of activity in a community.

Furthermore, regarding emotion analysis, we have presented some preliminary results that we plan to extend by using other methods for sentiment analysis, and by investigating further aspects like the evolution of the emotions over time in the feedback on a given work, and the comparison of the comments written by the author of a work with respect to the comments written by other users.

Finally, it would be interesting to compare social interaction and controversy metrics with emotion metrics, to investigate the relationship between emotional expression, controversy, conflict, and other aspects of online interactions in the context of fanfiction and prosumer communities.

5.3 Next steps: first ideas for dashboard development

The development of the Prosumer Intelligence Toolkit will be guided by different iterations of input and feedback received from stakeholders and potential end-users, to make sure it is aligned with their needs, and with the values of the communities (Smith et al, 2020; Miquel-Ribé & Laniado, 2021); however, we have made a first exercise of imagining possible content for the dashboards to be developed. Although this is just a first attempt at thinking about these next steps, we believe it is worth reporting it here, as we have already started to discuss with the consortium about it, in particular in the plenary session held in Barcelona on December 1st, 2021.

We have thought of four kinds of dashboards, focusing respectively on content, authors, users, and communities. As an important note, the content of the user dashboards should be carefully evaluated in terms of privacy issues: even in cases where user data are publicly accessible, computing metrics on them and showing indicators on individual users may be problematic. So, this should be kept in mind and checked in each specific case to make sure the dashboard contents are in line with ethical principles, community values and platforms' terms of service.

Content-centred dashboards

Aim: identifying and characterizing trending topics, content and tags that will become popular.

Dashboards with rankings of works, tags or topics, by different metrics:

- Most popular
- Most trending
- Most controversial

Additional variables shown per work:

- ✓ Language and emotion features in work's text
- ✓ Language and emotion features in comments
- ✓ Discussion network metrics
- ✓ Author involvement in discussion network

Timelines per work/tag/topic:

- → Feedback received over time
- → Possibility to compare different works/tag/topics

Author-centred dashboards

Aim: identifying and characterizing relevant authors.

Dashboards with user ranking by different author metrics:

- Most prolific → Most works authored/contributed
- Author relevance → Most central in the author feedback network (different centrality measures)
- Most trending → authors of trending works
- Most controversial → authors of controversial works

Additional variables shown per author

- ✓ Language and emotion features in text
- ✓ Language and emotion features in comments
- ✓ Centrality in discussion network of their own works

Timelines per author:

- → Works published, feedback received and centrality over time

User-centred dashboards

Aim: identifying and characterizing relevant users with different roles in the community.

Dashboards with user ranking by different activity metrics:

- Most active (giving feedback) → Out-degree in the author feedback network
- Discussion relevance → Most central in the discussion network (different centrality measures)

Additional variables shown per user

- ✓ Time spent in the community
- ✓ Role in the community or prosumer profile (cluster in the in-degree / out-degree plan)
- ✓ Author metrics
- ✓ Main interests / tags
- ✓ Activity in other communities

Timelines per user:

- → Activity and centrality over time

Community-centred dashboards

Aim: understanding and monitoring collective dynamics of co-creation in a community.

Dashboards with timelines on different aspects of community dynamics:

- → Amount of activity over time
- → Different kinds of user actions / namespaces
- → On different topics (tags)

- → Controversy over time
- → Distribution of authorship and activity over time
- → Proportion of users concentrating the 80% of the authorship / actions / received feedback
- → Gini coefficient for authorship / actions / received feedback
- → Social network metrics over time

As discussed above, this has to be intended only as a first thought about the dashboard development, building on the work done until now, and on the input collected so far; further interviews, focus groups and workshops from WP2 will drive the design and development of the dashboards in the Prosumer Intelligence Toolkit, making sure it is aligned with the end users' needs and goals (within the limits already identified).

6. References

Collier, B., & Bear, J. (2012). Conflict, criticism, or confidence: An empirical examination of the gender gap in Wikipedia contributions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work* (pp. 383-392).

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. InterJournal, Complex Systems, 1695(5), 1-9.

Dalton, K. L. (2012). Searching the archive of our own: The usefulness of the tagging structure (Doctoral dissertation, The University of Wisconsin-Milwaukee).

Du, P., Kibbe, W. A., & Lin, S. M. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *bioinformatics*, *22*(17), 2059-2065.

Fiesler, C., Morrison, S., & Bruckman, A. S. (2016). An archive of their own: A case study of feminist HCI and values in design. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 2574-2585).

Fiesler, C. (2018). Owning the Servers: A Design Fiction Exploring the Transformation of Fandom into 'Our Own'. *Transformative Works and Cultures*, *28*.

Gallagher, R. J., Frank, M. R., Mitchell, L., Schwartz, A. J., Reagan, A. J., Danforth, C. M., & Dodds, P. S. (2021). Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts. EPJ Data Science, 10(1), 4.

Gibbons, A., Vetrano, D., & Biancani, S. (2012). Wikipedia: Nowhere to grow. Retrieved July, 6, 2013.

Kaltenbrunner, A., & Laniado, D. (2012). There is no deadline: time evolution of Wikipedia discussions. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration* (pp. 1-10).

Lang, B., Dolan, R., Kemper, J., & Northey, G. (2020). Prosumers in times of crisis: definition, archetypes and implications. *Journal of Service Management*.

Lehmann, J., Gonçalves, B., Ramasco, J. J., & Cattuto, C. (2012). Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web* (pp. 251-260).

Miquel-Ribé, M., & Laniado, D. (2021). The Wikipedia Diversity Observatory: helping communities to bridge content gaps through interactive interfaces. *Journal of Internet Services and Applications*, *12*(1), 1-25.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. Computational intelligence, 29(3), 436-465.

Pianzola, F., Rebora, S., & Lauer, G. (2020). Wattpad as a resource for literary studies. Quantitative and qualitative examples of the importance of digital social reading and readers' comments in the margins. *PloS one*, *15*(1), e0226708.

Price, L., & Robinson, L. (2021). Tag analysis as a tool for investigating information behaviour: comparing fan-tagging on Tumblr, Archive of Our Own and Etsy. *Journal of Documentation*.

Riley, O. (2015). Archive of Our Own and the Gift Culture of Fanfiction. Retrieved: https://conservancy.umn.edu/bitstream/handle/11299/175558/Archive%20of%20Our%20Own%20and%20the%20Gift%20Culture%20of%20Fanfiction.pdf

Smith, C. E., Yu, B., Srivastava, A., Halfaker, A., Terveen, L., & Zhu, H. (2020). Keeping community in the loop: Understanding wikipedia stakeholder values for machine learning-based systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).

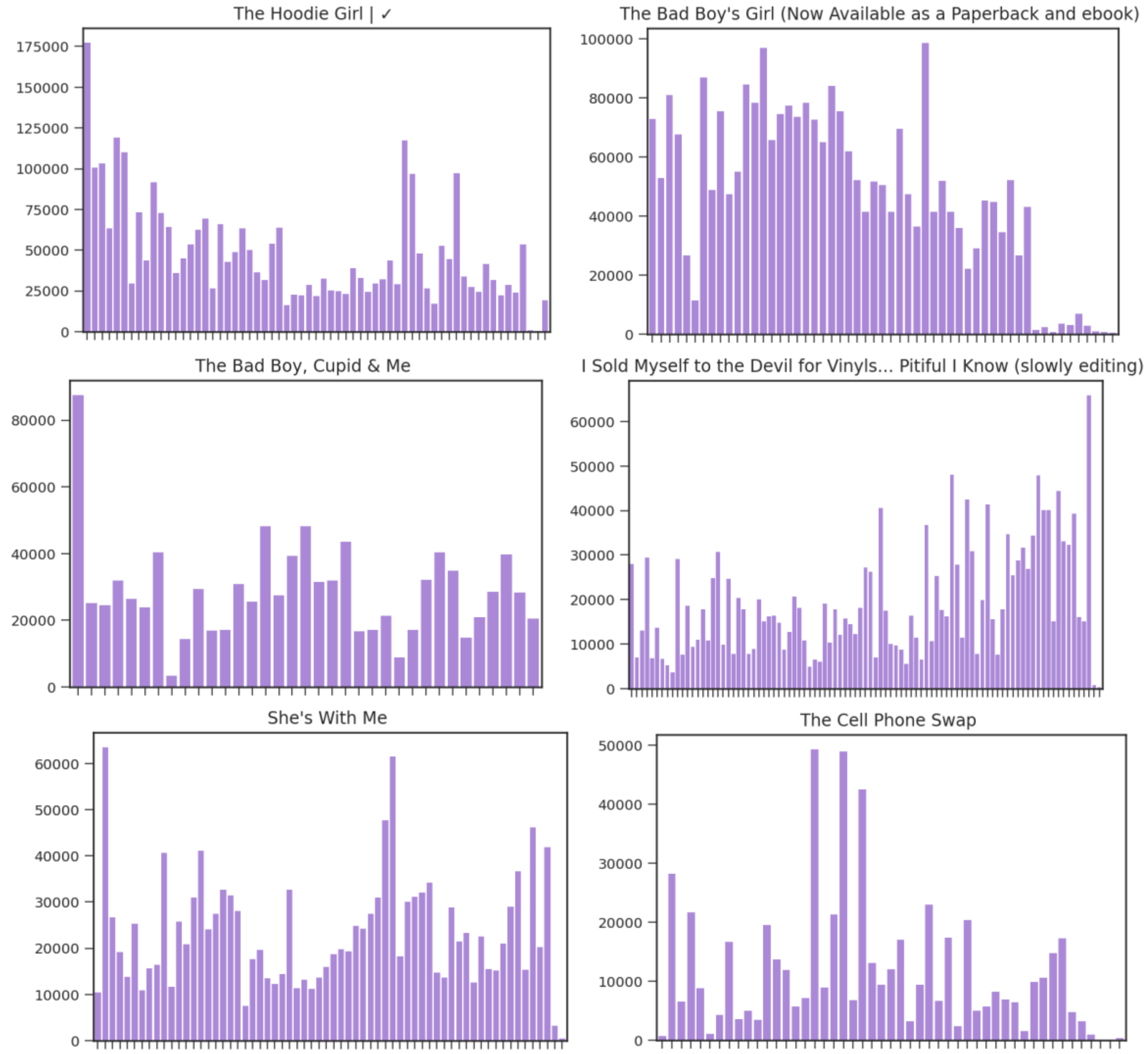
Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24-54.

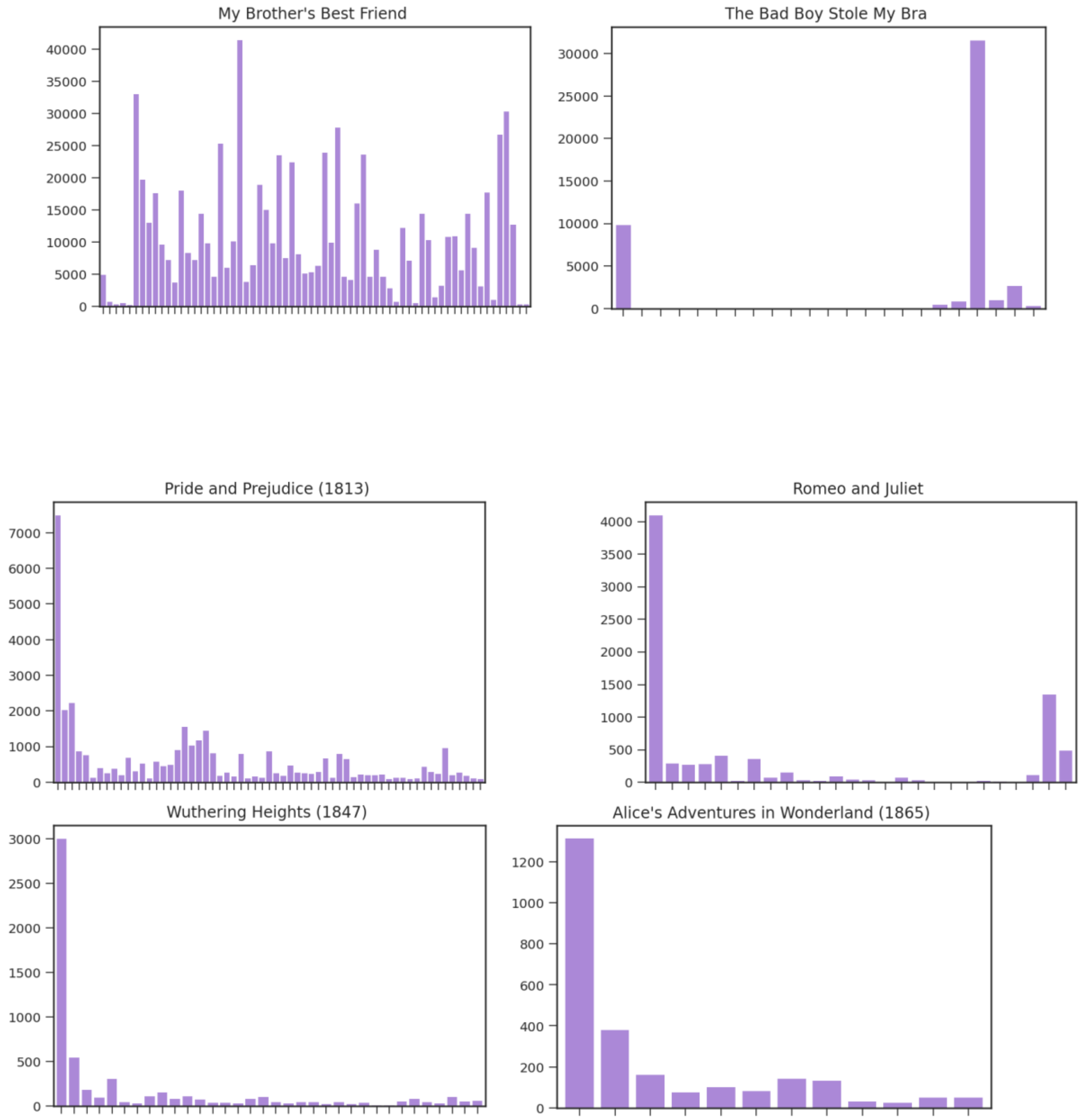
Wasserman, S., and Faust, K. (1994). Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.

Zeng, A., Gualdi, S., Medo, M., & Zhang, Y. C. (2013). Trend prediction in temporal bipartite networks: the case of Movielens, Netflix, and Digg. *Advances in Complex Systems*, *16*(04n05), 1350024.

Annex: Wattpad

Distributions of comments




My Brother's Best Friend
40000
35000
30000
25000
20000
15000
10000
5000
0
The Bad Boy Stole My Bra
30000
25000
20000
15000
10000
5000
0
Pride and Prejudice (1813)
7000
6000
5000
4000
3000
2000
1000
0
Romeo and Juliet
4000
3500
3000
2500
2000
1500
1000
500
0
Wuthering Heights (1847)
3000
2500
2000
1500
1000
500
0
Alice's Adventures in Wonderland (1865)
1200
1000
800
600
400
200
0

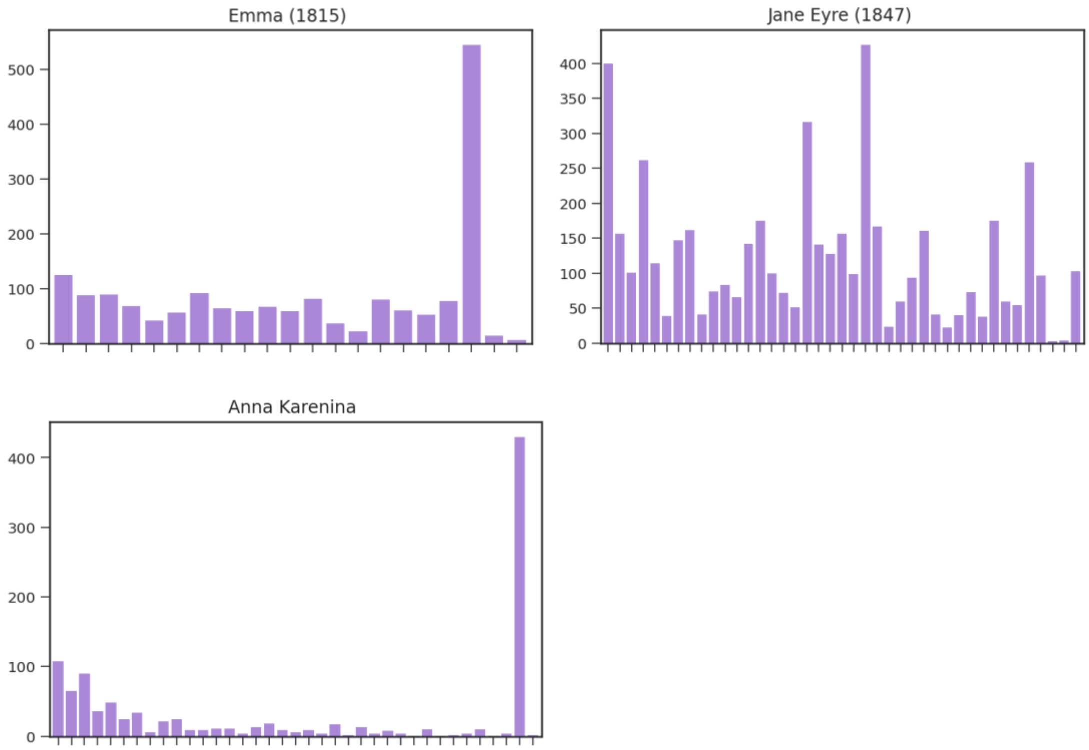
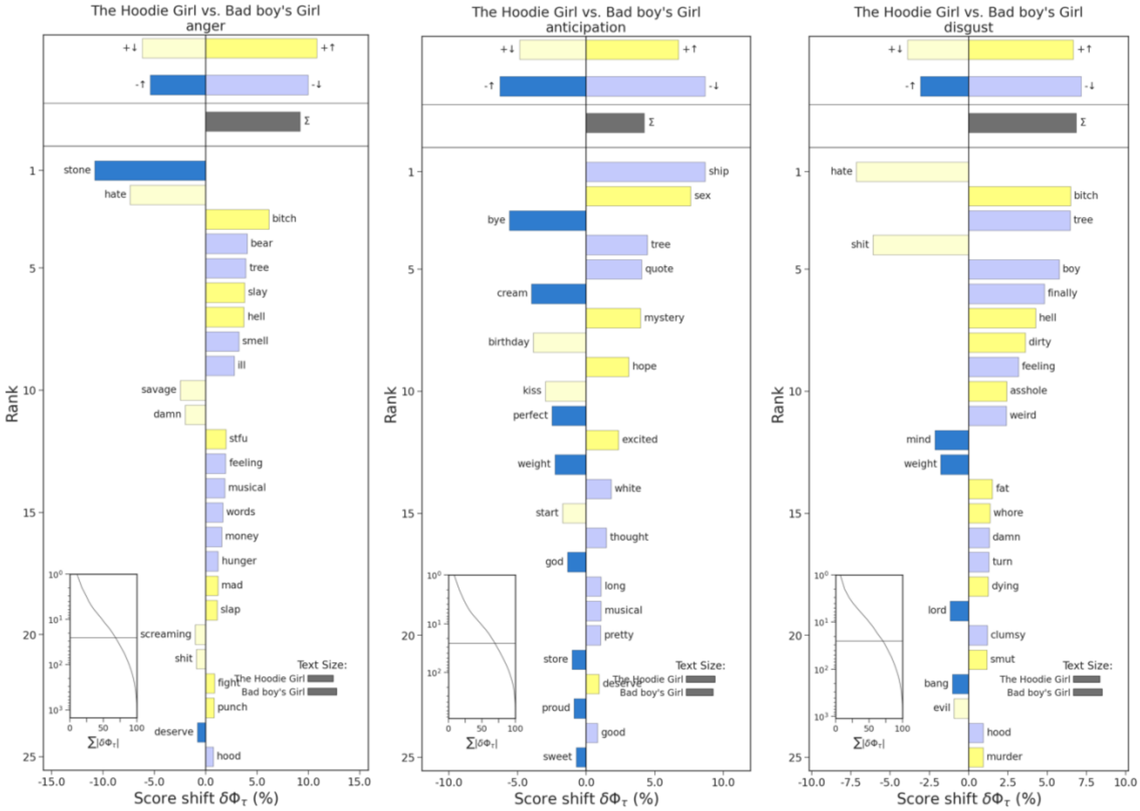
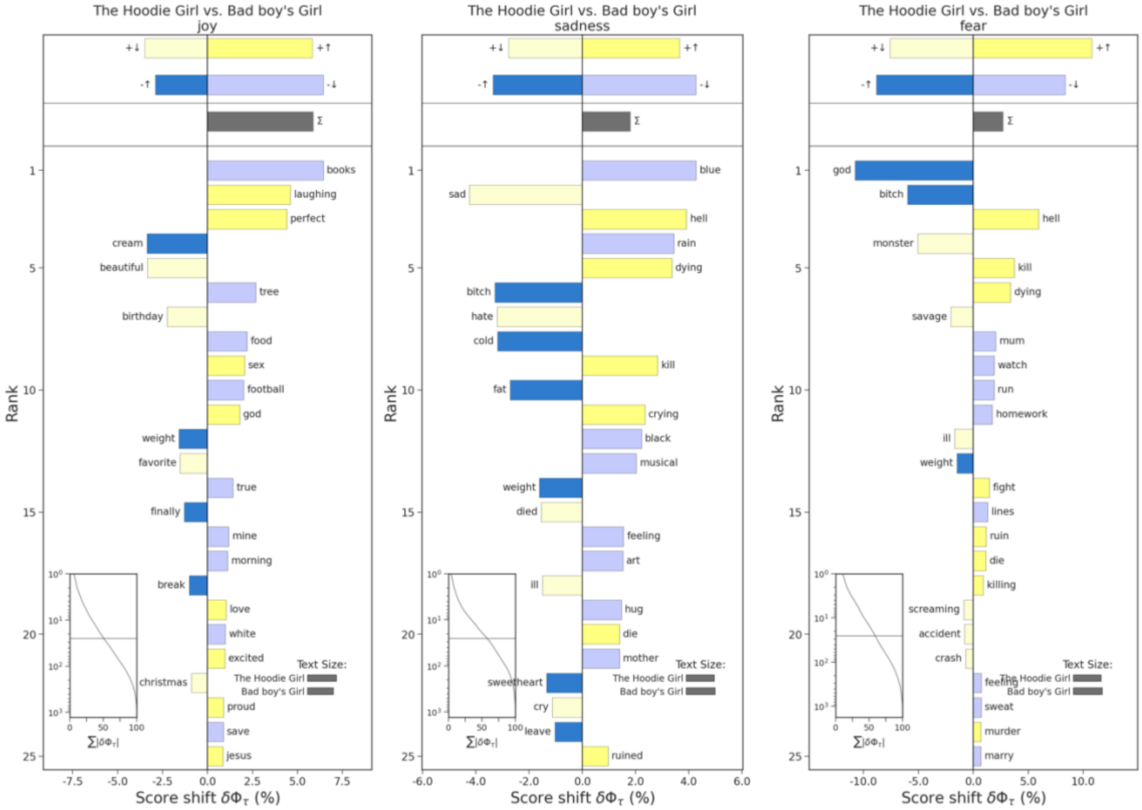


Figure 36 - Distribution of comments by chapter for all the books

Pairwise comparation using NRC lexicon




The Hoodie Girl vs. Bad boy's Girl
joy
+↓
+↑
-↑
-↓
Σ
books
laughing
perfect
cream
beautiful
tree
birthday
food
sex
football
god
weight
favorite
true
finally
mine
morning
break
love
white
excited
christmas
proud
save
jesus
Text Size:
The Hoodie Girl
Bad boy's Girl
Rank
Score shift $\delta\Phi_\tau$ (%)
$\sum|\delta\Phi_\tau|$
The Hoodie Girl vs. Bad boy's Girl
sadness
blue
sad
hell
rain
dying
bitch
hate
cold
kill
fat
crying
black
musical
weight
died
feeling
art
ill
hug
die
mother
sweetheart
cry
leave
ruined
The Hoodie Girl vs. Bad boy's Girl
fear
god
bitch
hell
monster
kill
dying
savage
mum
watch
run
homework
ill
weight
fight
lines
ruin
die
killing
screaming
accident
crash
feeling
sweat
murder
marry

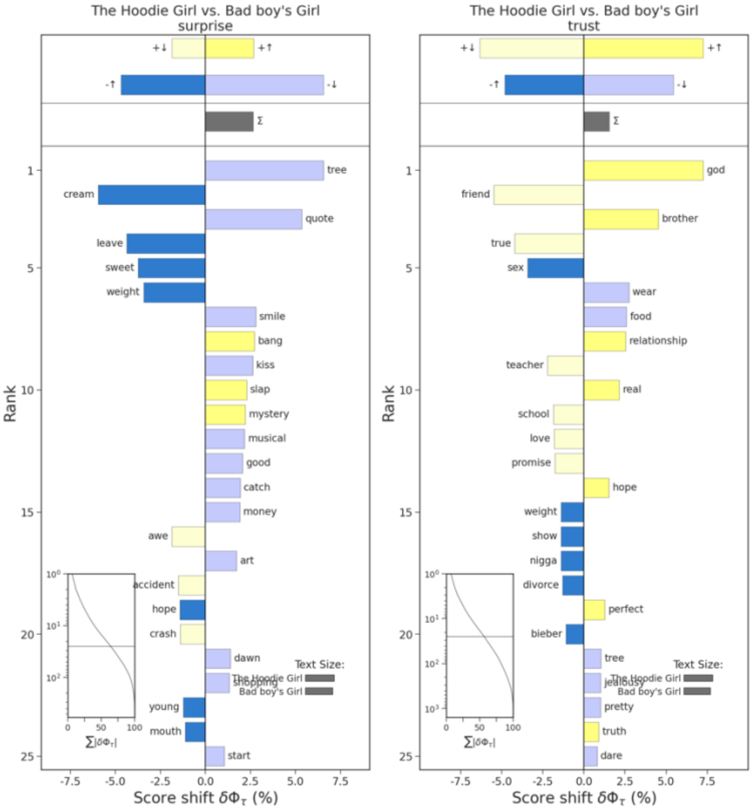


Figure 37 - Pairwise comparation NRC emotions between The Hoodie Girl and Bad Boy's Girl