Specific behaviours in Wikipedia talk pages: some insights from extreme cases

Ludovic Tanguy, Céline Poudat, Lydia-Mai Ho-Dac

Specific behaviours in Wikipedia talk pages: some insights from extreme cases

Ludovic Tanguy[(1)], Céline Poudat[(2)], Lydia-Mai Ho-Dac[(1)]

(1) CLLE: CNRS & University of Toulouse, France
(2) BCL: CNRS & University of Nice Côte d'Azur, France
Email: ludovic.tanguy@univ-tlse2.fr, celine.poudat@univ-cotedazur.fr, lydia-mai.ho-dac@univ-tlse2.fr

Abstract

Based on a dataset of 3.4 million threads from English Wikipedia talk pages, we specifically focus on extreme cases. We propose a qualitative analysis of the most prolific message authors, the longest threads in terms of messages, contributors and durations, as well as the longest monologues (single-user threads). These case studies allow us to identify a number of behaviours that can significantly differ from the typical discussions between Wikipedians. If some threads do not have a real dialogic status (polls, monologues, logbooks and diaries), some of them push online communication to its limits across time. These sometimes unexpected behaviours can help us get a more precise understanding of this unique source of computer-mediated communication data.

Keywords: Wikipedia talk pages, online interaction, extreme behaviours

1. Introduction

The study presented in this paper is part of a larger project that explores various dimensions of Wikipedia talk pages. Talk pages have been extensively studied as they provide a unique means to examine the dynamics of interaction between Wikipedians (Laniado et al. 2011). They also serve as a valuable source of computer-mediated communication data which is abundant, multilingual and freely accessible, making them suitable for large-scale studies on generic online interactions (Gomez et al. 2011, Lügen & Herzberg 2019). The main practices in Wikipedia talk pages have already been studied and described with a focus on the topics discussed (Schneider et al. 2010) or local interaction patterns (Kopf 2022).

The case study presented here focuses on marginal, or even extreme behaviours in Wikipedia talk pages. We have selected a number of outlier cases that exhibit unexpected characteristics at the thread or user levels. These include highly prolific users, excessively long threads (in terms of duration, number of posts or users involved) and monologues. We assume that the analysis of such extreme cases can help to better understand expected and unexpected interactions between Wikipedians. This will also allow us to highlight practices which are generally neglected although they may be found in more typical configurations.

2. Dataset: English Wikipedia talk pages

We base our study on a dataset, which consists of threads extracted from the August 2019 dump of Wikipedia. At that time the English version of Wikipedia contained 14,856,106 article pages and 7,903,148 talk pages, including archives. Among these, only 2,025,888 contained at least one posting with at least 2 words.

It is worth noting that talk pages on Wikipedia are produced on the same infrastructure as the articles, using *wikicode* formatting. This means that a talk page is fully editable by any user and that its layout and organisation can be freely modified, in spite of strong recommendations from the Wikipedia community. Talk pages typically feature a section-based structure, with each section representing a distinct discussion having its own heading and clear boundaries. Individual messages are organised along a tree structure which follows the example of the more traditional online discussion platforms. However, the wikicode allows freeform editing which may lead to unusual structures in discussion threads, such as the re-sectioning of existing talk pages (used for archival purposes for example), the writing of non-contiguous answers to a previous long message (similar to emails), or postings appearing in a non-chronological order. This situation has dire consequences on the parsing of Wikipedia talk pages, which requires additional efforts to identify the network of interactions.

Despite these challenges, we segmented each talk page into sections, with each section representing a thread. Each thread was segmented into posts (or comments or messages) following an heuristic based on signatures and indentations. The whole structure was then converted into XML format following the TEI-CMC guidelines, so that each post is associated with its author's name and date. Finally, threads containing a post written by a bot were discarded. In the end our corpus contains 3,385,583 threads and 8,873,620 messages (Ho-Dac, to appear).

Table 1 gives an overview of the dataset characteristics that were considered relevant for identifying extreme behaviours. The large differences between means and medians suggest highly skewed distributions with numerous outliers for each variable. In the following section we focus on the outlier cases corresponding to the highest values for each variable in the table.

Feature	Maximum	Median	Mean
Number of posts per user	25,078	1	20.06
Number of posts per thread	651	1	2.62
Number of users involved	97	1	1.85
Duration of threads with 2 or more posts (N=1,688,939**)**	16.6 years	5.3 days	260 days
Longest duration between 2 posts in the thread	16.1 years	4.1 days	233 days
Number of posts per single user thread (N=1,812,457**)**	150	1	1.08

Table 1: Overview of features used to identify extreme behaviours

3. Extreme behaviours

While identifying outliers is a common initial step in data analysis, its primary objective is to remove atypical individuals which can skew the study of the central tendencies. Here, although we initially targeted outliers in order to exclude them from the dataset and facilitate discourse analysis studies, the qualitative analysis of these outliers allows us to identify behaviours that are made possible by the Wikipedia device, and that may even be typical of Wikipedia interactions.

3.1 The most prolific message authors

Our first investigation targets Wikipedia users who have produced a significant number of posts on talk pages. In our dataset, we found a total of 499,137 different usernames in the signatures of all talk pages (without including the bots or the unregistered users who are only identified by their IP addresses). As expected, the number of posts per user follows a Zipfian distribution, meaning that while a majority of users have written a single comment, a few Wikipedians are the authors of a very large number of messages. The user ranking first posted 25,078 messages, the user ranking #10 14,281, and the user ranking #100 5,900.

To compare message-posting behaviour with actual Wikipedia editing activity, we gathered data on the number of edits (i.e. the modifications made on any page of the Wikipedia, including posts in any kind of talk page) and the number of posts in the article talk pages for the 1000 most productive Wikipedia editors as indicated in the official leaderboard[1] (as of July 2019), shown in Figure 1. We measured a weak positive correlation (ρ=0.09) between the number of edits and the number of messages. As an example, the most active editor of the English Wikipedia (Steven Pruitt, who was responsible for more than 3 million edits in 2019, and over 5 million in 2023) has never participated in a discussion in any article talk page (although he did post some messages in a few users' personal talk pages, not included in our dataset).

[1] https://en.wikipedia.org/w/index.php?title=Wikipedia:WBE

Similarly, several of the most prolific authors on the articles talk pages rarely modify the articles themselves, limiting their role to commenting or proofreading the text written by others, or to enforcing Wikipedia policy and rules through discussion.

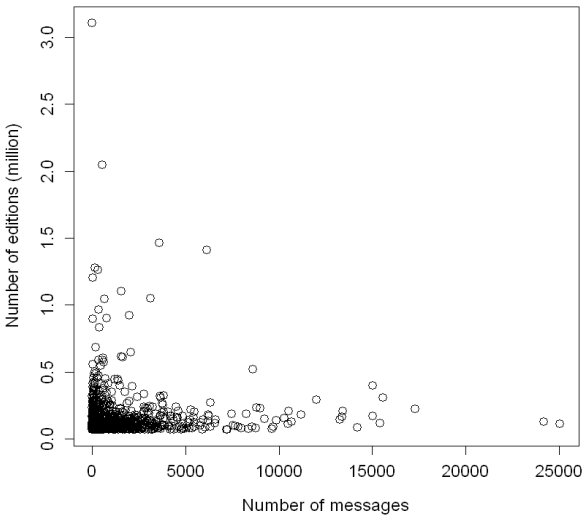


Figure 1: Number of editions versus number of messages for the 1000 most productive Wikipedia editors

These first observations would clearly show that taking part in a Wikipedia discussion can to some extent be considered as a specific activity, decorrelated from article writing, at least for a subset of the Wikipedia users.

3.2 Threads with the highest numbers of posts

The second phenomenon we investigated is the number of posts per thread. If 53% of the threads consist of a single post, some of them contain several hundred posts.

We examined the 100 longest threads in our dataset (threads with more than 90 posts, up to 651). Surprisingly, these very long threads rarely imply a large number of participants (median of 14 different users) and they can even be written by a single user (this particular category is examined more closely in §3.5).

If we only consider their organisation and structure, these long threads can be classified as follows:

- 68 of the 100 examined threads can be qualified as *standard discussions*. Indeed, these threads follow the conventional organisation where users exchange their views and arguments, following a tree-like structure where the replies and reactions to previous posts are indicated through cumulative indentations. However, due to the extensive size and depth of the threads, indentation can hinder their readability. To address this, some users (most of the time participants to the discussion) sometimes use the flexibility of the talk pages (based on the same wikicode used for article pages) to organise them into sections. When appropriate, subtopics can be identified and used to start a new nested thread in a subsection, while remaining in the same section and therefore related to

the same topic. When not, *arbitrary breaks* are introduced to reset the indent level when it becomes too deep[2].

- 26 of them are *polls* or *series of polls*. In these threads a user collects the position or opinion of others on specific topics. As such, every single vote by the polled users counts for a message. The length of these threads can be attributed to the high number of participants (up to 97), multiple related polls grouped together (with the same users posting a message for each subtopic), or one or more nested threads developing inside the poll. For example, a user may explain his or her position, eliciting reactions from others. These threads are further described in §3.3.
- 6 are long *lists*, the items of which are expressed as separate messages, and are initially posted by the same user. As these threads only marginally contain posts by different users we study in more detail this specific type of thread in §3.5.

To summarise, our findings indicate that only two thirds of the 100 longest threads can be classified as discussions, highlighting the diverse uses of talk pages.

3.3 Threads with the most users

The 100 threads with the highest numbers of different participants are all polls or series of polls. Polls are a common practice in Wikipedia talk pages as they represent the pursuit of consensus (Kopf 2022). Polls can cover various decisions related to the article page, such as article deletion, merging with another related article, changing the article's title, deleting a whole section, choosing between different pictures etc. These polls may be created after inconclusive discussions or as a first intent when dealing with a new issue. The questions asked can be binary (support/oppose a suggestion) or open-ended (propose a new title, picture etc.). As we focus here on the number of different users, our sample is limited to threads with a single poll.

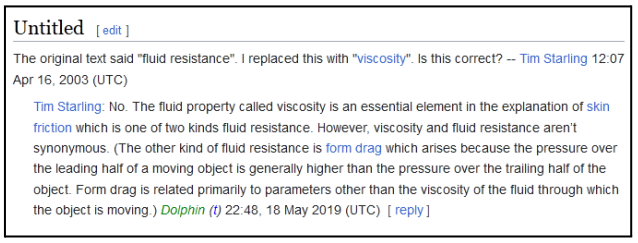
Due to the flexibility of the underlying wikicode, polls may be organised in two different ways. Messages can be in chronological order, with each user expressing her opinion in sequences. Alternatively, messages can be grouped based on their position, so that all messages, users and arguments in support or opposing the initial proposition are in the same section.[3]

Some of the polls are both spontaneous and local, and can be organised inside a discussion: they are qualified as straw polls. Others are qualified as Request for Comments (RfC) and follow a more sophisticated organisation. RfC polls are indexed in the Wikipedia space and therefore receive much more attention. This increased attention can lead to some problems when high stakes motivate certain users to manipulate the voting process with additional or fake accounts (*puppetry*), leading to their abandonment.[4] Several of our most massive threads show such cases that are explicitly flagged, but all expressed votes and comments remain available.

3.4 Longest-lasting threads

The temporal dynamics of Wikipedia discussions has been studied in (Kaltbrunner & Laniado 2012) but, as seen in Table 1, some threads can last more than 15 years, nearly the timespan of our dataset. In 2019, the 100 longest-lasting threads covered a duration of over 14.5 years. 8 of the threads we examined are false positives: the prolonged duration is merely a consequence of some messages being placed in a generic section of the talk page (labelled as "*Comments*" or similar). Therefore the messages simply do not constitute a discussion; but the 92 other cases are clear instances of communicating occurring over an extended period of time.

About 10% of these threads exhibit a continuous spread over a significant period, with regular postings and no extended periods of silence exceeding a couple of years. However, the majority of threads demonstrate a single notable jump across time, with a message being posted in response to a comment made over a decade ago, such as the example in Figure 2.


Untitled [edit]

The original text said "fluid resistance". I replaced this with "viscosity". Is this correct? -- Tim Starling 12:07 Apr 16, 2003 (UTC)

Tim Starling: No. The fluid property called viscosity is an essential element in the explanation of skin friction which is one of two kinds fluid resistance. However, viscosity and fluid resistance aren't synonymous. (The other kind of fluid resistance is form drag which arises because the pressure over the leading half of a moving object is generally higher than the pressure over the trailing half of the object. Form drag is related primarily to parameters other than the viscosity of the fluid through which the object is moving.) *Dolphin (t)* 22:48, 18 May 2019 (UTC) [reply]

Figure 2: sample thread with a 16-year gap
https://en.wikipedia.org/wiki/Talk:Charles-Augustin_de_Coulomb#Untitled

Surprisingly, most of these dialogues (72) contain no explicit mention of the temporal specificity. Users write their comments as if the message they are reacting to was posted just a few minutes ago. A wide range of dialogue acts can be observed in such situations: answering a simple factual question (as in Figure 2), providing a reference, commenting on a statement[5], etc. In a few of these cases however we found that the answerer addresses the author of the first message in the third person, which may seem unusual in online communications ("Related to why that was put by *an earlier editor*, the reason is [...]"[6], "I have to wonder what *this IP user* imagined [...]"). This may indicate that the more recent author acknowledges the fact that his interlocutor has long departed from the talk page and that the response is directed toward present

[2] https://en.wikipedia.org/wiki/Talk:Gamergate_(harassment_campaign)/Archive_12#KotakuInAction_moderators_misogynist/anti-feminist/interested_in_female_subjugation_porn

[3] https://en.wikipedia.org/wiki/Talk:Campaign_for_the_neologism_%22santorum%22/Archive_6#Proposal_to_rename,_redirect,_and_merge_content

[4] https://en.wikipedia.org/wiki/Talk:K._P._Yohannan#Keeping_the_controversy_Section_in_this_article

[5] https://en.wikipedia.org/wiki/Talk:T-shirt#Capitalisation

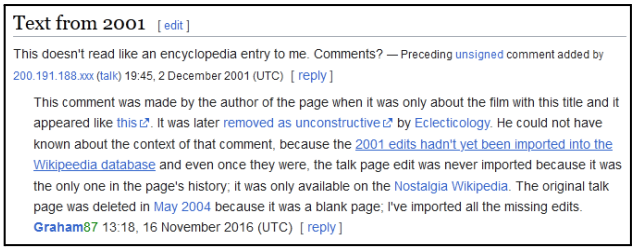
[6] https://en.wikipedia.org/wiki/Talk:Brondesbury#Place

and future readers. But this particular behaviour has to be studied more precisely; Herzberg & Lügen (to appear) studied the different ways a user addresses the author of a previous message, and found that a second person address occurs in less than 30% of replies.

If the late response is sometimes justified by a change in the world or an advancement of knowledge, it can also deal with atemporal topics. All these efforts to provide answers and additional information across time, even in the absence of the original participant, reflects the global dynamics and objective of the Wikipedia project.

In the remaining cases, users also take advantage of the flexibility of Wikipedia talk pages. Some users explicitly modify the timestamp of their message, pre-dating them to several years in the future to prevent their automatic archival. This is a move similar but somewhat more drastic to "bumping" a thread in online forums (i.e. adding empty messages to an existing thread to keep it visible).

In two cases we found what can be qualified as *talk page archaeology* (see example in Figure 3). A user re-posts an old message or discussion that had been deleted or lost in the restructuring of Wikipedia. The reason for this is apparently not to answer the initial question or to correct a statement, but simply to preserve a trace from previous efforts. This preservative attitude has even led to keeping the very first versions of Wikipedia accessible in *Nostalgia Wikipedia*[7].


Text from 2001 [edit]

This doesn't read like an encyclopedia entry to me. Comments? — Preceding unsigned comment added by 200.191.188.xxx (talk) 19:45, 2 December 2001 (UTC) [reply]

This comment was made by the author of the page when it was only about the film with this title and it appeared like this. It was later removed as unconstructive by Eclecticology. He could not have known about the context of that comment, because the 2001 edits hadn't yet been imported into the Wikipeedia database and even once they were, the talk page edit was never imported because it was the only one in the page's history; it was only available on the Nostalgia Wikipedia. The original talk page was deleted in May 2004 because it was a blank page; I've imported all the missing edits. Graham87 13:18, 16 November 2016 (UTC) [reply]

Figure 3: sample thread restoring a previous comment
https://en.wikipedia.org/wiki/Talk:Casablanca#Text_from_2001

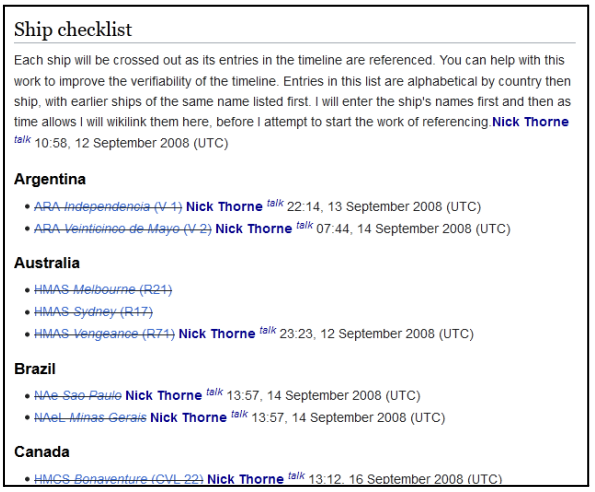
Although these temporal behaviours have not been formally described before, they confirm the specific position of the Wikipedia project as a global memory as expressed by Pentzold et al. (2017).

3.5 Longest single-user threads

Our last study focuses on single-user threads. In our dataset, 53% of all threads are authored by a single user, primarily due to them consisting of a single post. However, 6.9% of threads with 2 or more posts are written entirely by a single user. These "monologues" can grow to be quite extensive, reaching up to 150 messages. Similar to our previous analyses, we examined the 100 longest single-user threads (with 12 or more posts) and identified two main configurations.

A significant majority of these threads (88) are *lists*, as we had observed in some of the longest threads (§3.2). The messages within these threads can take the form of paragraphs that include comments, remarks or suggestions[8]. These cases typically result from a review of the article, or a series of proposals and suggestions for rewriting or expanding it. Of course, these items can sometimes receive comments or extensions in the form of nested messages by other users as noted in §3.3.

But long lists of another kind contain only simple informational elements relevant to the article, such as products, dates, characters, users… In most cases, the thread lacks an explicit communication goal and appears to function as a logbook or to-do list for the author. A thread of such "grocery list" type can include check marks or crossed out items, indicating that they have been processed (e.g. proofread, referenced, integrated into the article…). In only 12 cases of such lists we could find explicit invitations from the author to others to contribute by extending, commenting or correcting the items, although in our sample these remained unanswered.. Figure 4 shows such an explicit checklist with the author giving potential helping hands precise instructions.


Ship checklist

Each ship will be crossed out as its entries in the timeline are referenced. You can help with this work to improve the verifiability of the timeline. Entries in this list are alphabetical by country then ship, with earlier ships of the same name listed first. I will enter the ship's names first and then as time allows I will wikilink them here, before I attempt to start the work of referencing. Nick Thorne talk 10:58, 12 September 2008 (UTC)

Argentina

- ~~ARA *Independencia* (V-1)~~ Nick Thorne talk 22:14, 13 September 2008 (UTC)
- ~~ARA *Veintiocho de Mayo* (V-2)~~ Nick Thorne talk 07:44, 14 September 2008 (UTC)

Australia

- ~~HMAS *Melbourne* (R21)~~
- ~~HMAS *Sydney* (R17)~~
- ~~HMAS *Vengeance* (R71)~~ Nick Thorne talk 23:23, 12 September 2008 (UTC)

Brazil

- ~~NAe *Sao Paulo*~~ Nick Thorne talk 13:57, 14 September 2008 (UTC)
- ~~NAeL *Minas Gerais*~~ Nick Thorne talk 13:57, 14 September 2008 (UTC)

Canada

- ~~HMCS *Bonaventure* (CVL 22)~~ Nick Thorne talk 13:12, 16 September 2008 (UTC)

Figure 4: sample list thread (extract)
https://en.wikipedia.org/wiki/Talk:Timeline_for_aircraft_carrier_service/Archive_1#Ship_checklist

The 12 remaining long monologues contain heterogeneous posts, which can consist of larger text segments such as problem analyses, reviews, suggestions, hypotheses, reports of actions taken, steps in an investigation and more, to various combinations of such messages within the same thread[9]. In all cases these monologues lack explicit indicators of dialogue such as the use of second-person pronouns or explicit calls for reactions. Instead, they can be considered as some kind of diary, following a Wikipedian's work and thoughts on a topic, spread out over time.

[7] https://nostalgia.wikipedia.org/

[8] https://en.wikipedia.org/wiki/Talk:Timeline_of_the_Irish_War_of_Independence#Doubtful_edits

[9] https://en.wikipedia.org/wiki/Talk:CMB%20cold%20spot#Professor_Mersini_Radio_Broadcast

4. Conclusion

Our study of the outlier threads in a dataset of over 3 million discussions from the English Wikipedia talk pages has allowed us to identify several specific behaviours.

The flexibility of the platform plays a crucial role in enabling these behaviours, as users can reshape and reorganise the posts in ways which are not possible in the other online discussion environments. The ability users have to freely (re-)order messages in a thread facilitates the emergence of new forms such as organised polls, sectioned long threads and the use of threads as checklists. In some cases, these possibilities may induce a shift away from the central communicational goal of the talk pages, such as monologues and threads used as log books or diaries. However, interaction remains possible even in these cases.

Our observations of long-lasting discussions confirm the objective of the Wikipedia project to create a cultural monument and testament. Talk pages, as the main articles of the encyclopaedia, are considered permanent documents. Therefore, it is not a problem for a Wikipedian to respond to a message even 15 years later, with the response being primarily directed towards the community rather than the original user.

It was not our aim to investigate the specific topics or domains in which certain types of discussion take place. During our observations we did not identify any particular area of knowledge that would correlate with specific behaviours. However, it is evident that popular topics such as pop culture, sports and geopolitics tend to attract a larger number of participants. Nevertheless, impressive efforts to gather information from a single individual can be found across various subjects, including niche areas.

On the methodological front, our approach needs further completion by exploring the extent to which these phenomena appear in less extreme cases. Preliminary surveys have shown, for instance, that polls and single-author lists appear at much smaller scales (2-3 voters, a few items in a list, short monologues) and, therefore, occur more frequently.

This naturally calls for further investigations, including a more systematic corpus search of local configurations in order to estimate the frequency of these behaviours, and to enable cross-lingual comparisons. It should be noted, however, that Wikipedia talk pages cannot be regarded as typical CMC data without taking these specificities into account.

5. References

Gómez, V., Kappen, H. J., & Kaltenbrunner, A. (2011). Modeling the structure and evolution of discussion cascades. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia* (pp. 181-190).

Kaltenbrunner, A., & Laniado, D. (2012). There is no deadline: time evolution of Wikipedia discussions. In *Proceedings of the 8th Annual International Symposium on Wikis and Open Collaboration* (pp. 1-10).

Kopf, S. (2022). *A Discursive Perspective on Wikipedia: More than an Encyclopaedia?* Palgrave Macmillan.

Herzberg, L. & Lügen, H. (to appear). Investigating reply relations on Wikipedia talk pages to reconstruct interactional strategies of Wikipedia authors. In C. Poudat, H. Lüngen & L. Herzberg (Eds.), *Investigating Wikipedia: linguistic corpus building, exploration and analyses*. John Benjamins.

Ho-Dac, L.-M. (to appear). Building a comparable corpus of online discussions in Wikipedia: the EFG WikiCorpus. In C. Poudat, H. Lüngen & L. Herzberg (Eds.), *Investigating Wikipedia: linguistic corpus building, exploration and analyses*. John Benjamins.

Laniado, D., Tasso, R., Volkovich, Y., & Kaltenbrunner, A. (2011). When the Wikipedians talk: Network and tree structure of Wikipedia discussion pages. In *Fifth international AAAI conference on weblogs and social media*.

Lüngen, H. & Herzberg, L. (2019): Types and annotation of reply relations in computer-mediated communication. *European Journal of Applied Linguistics* 7 (2). Berlin/Boston: de Gruyter, 2019. S. 305-331.

Mehler, A., Gleim, R., Lücking, A., Uslu, T., & Stegbauer, C. (2018). On the Self-similarity of Wikipedia Talks: a Combined Discourse-analytical and Quantitative Approach. *Glottometrics*, *40*, 1-45.

Pentzold, C., Weltevrede, E., Mauri, M., Laniado, D., Kaltenbrunner, A., & Borra, E. (2017). Digging Wikipedia: The online encyclopedia as a digital cultural heritage gateway and site. *Journal on Computing and Cultural Heritage (JOCCH)*, *10*(1), 1-19.

Schneider, J., Passant, A., & Breslin, J. G. (2010). A content analysis: How Wikipedia talk pages are used. In *Proceedings of the 2nd International Conference of Web Science* (pp. 1-7).