Tor Users Contributing to Wikipedia: Just Like Everybody Else?

Chau Tran*, Kaylea Champion,† Andrea Forte,‡ Benjamin Mako Hill,§ Rachel Greenstadt¶

Abstract User-generated content on sites like Wikipedia is produced by volunteers who not only create content, but invest time and effort in quality control by reviewing others’ contributions. Because of a perception that privacy enhancing tools are a source of vandalism, spam, and abuse, many user-generated sites like Wikipedia block contributions from anonymity-seeking editors who use proxies like Tor. While blocking anonymity seeking editors is perceived to be effective in stemming abuse, collateral damage in the form of unrealized valuable contributions from anonymity seeking editors is typically invisible. Although Wikipedia has taken steps to block contributions from Tor users since as early as 2005, we demonstrate that these blocks have been imperfect and that tens of thousands of attempts to edit on Wikipedia through Tor have been successful. We draw upon several data sources to measure and describe the history of Tor editing on Wikipedia over time and to compare contributions of Tor users to other groups of Wikipedia users. Our analysis suggests that the Tor users who manage to slip through Wikipedia’s ban contribute content that is similar in quality to unregistered Wikipedia contributors and to the initial contributions of registered users.

INTRODUCTION

When a Wikipedia reader using the Tor Browser notices a stylistic error or a missing fact and clicks the “Edit” button to fix it, they are shown a message like the one reproduced in Figure 1. Wikipedia informs Tor users who attempt
to edit an article that they, like other people using systems to protect their privacy, have been preemptively blocked from contributing. Wikipedia is far from alone in their decision to block participation from anonymity seeking users. Although service providers vary in their approaches McDonald, Hill, Greenstadt, and Forte, 2019, policies like Wikipedia’s are common across a range of user-generated content websites.

Figure 1: Screenshot of the page a user is shown when they attempt to edit the Wikipedia article on “Privacy” while using Tor.

Tor is an anonymity network that allows people to browse the Internet without being tracked. Previous research has shown how Wikipedia’s policy prohibiting edits by Tor users stems from a fear of abuse and vandalism McDonald et al., 2019. On the other hand, it stands to reason that not all would-be contributors from Tor are abusive. Wikipedia can ill afford to lose contributors. Research has shown that the number of Wikipedia contributors has declined steadily since it peaked in 2007, and this decline is partially attributed to the increasing rejection of good-faith contributions Halfaker, Geiger, Morgan, and Riedl, 2013. Moreover, Wikipedia editors, especially those from underrepresented or marginalized groups or who wish to edit about stigmatized topics face significant privacy concerns Forte, Andalibi, and Greenstadt, 2017. There is also evidence that concerns about government surveillance in particular have had chilling effects on people’s willingness to even view articles on certain topics Penney, 2016.

In this work, we present a method of measuring the value of contributions made by the privacy-seeking community. We focus on the users of a single service, Wikipedia, and a single anonymity protecting technology, Tor, to understand what is lost when a user-generated content site systematically blocks contributions from users of privacy-enhancing technologies. We make use of the fact that Wikipedia’s mechanism of blocking Tor users has been im-
perfect. We found and extracted a total of 11,363 edits on English Wikipedia made by Tor users between 2007 and 2018 as well as 12,576 edits made to the German, French, Spanish, and Russian Wikipedia editions.

In the paper that follows, we explain how some Tor users managed to slip through Wikipedia’s ban and describe our process for constructing our dataset of Tor edits. We use this dataset to compare Tor editors (i.e., people using Tor to edit), with three different control sets of time-matched edits from other Wikipedia contributor populations:

1. **IP editors** people editing from non-Tor IP addresses, also referred to by Wikipedia editors as “anons,”
2. **First-time editors** people logged into accounts making their first edit, and
3. **Registered editors** people logged into accounts with more than one edit.

Using several different measures of quality, we find that the contributions of Tor editors are similar in quality to those of IP editors and first-time editors. In an exploratory analysis, we fit topic models to Wikipedia articles and find intriguing differences and similarities between the kinds of topics that Tor users and other Wikipedia editors contribute to. Additionally, we examine Tor users’ behavior across different language editions on Wikipedia, where their contributions are reviewed by communities and administrators with different attitudes towards Tor use and anonymity and find distinct patterns over time. We conclude with a discussion of how sites like Wikipedia might let the estimated 2,000,000 daily Tor users\(^1\) edit freely and how doing so might benefit both Wikipedia and society.

**Related Work**

Most people seek out anonymity online at some time or another.\(^2\) Their reasons for doing so range from seeking help and support Andalibi, Haimson, De Choudhury, and Forte, 2016, contributing to online projects Forte et al., 2017, seeking information, pursuing hobbies, and illegal activities like file sharing Kang, Brown, and Kiesler, 2013. Despite the fact that many of the activities anonymity seekers pursue are typical of other Internet users, people using Tor and other anonymity-seekers have a history of being treated as second-class citizens on the Web. This differential treatment is not specific to Wikipedia. Many websites systematically block traffic coming from Tor with

---

\(^1\)https://metrics.torproject.org/userstats-relay-country.html

\(^2\)http://www.pewinternet.org/2013/09/05/anonymity-privacy-and-security-online/
multiple levels of access limitation.\textsuperscript{3} According to Khattak et al. at least 1.3 million IP addresses blocked Tor at the TCP/IP level as of 2015, and “3.67\% of the top 1,000 Alexa sites are blocking people using computers running known Tor exit-node IP addresses” Khattak et al., 2016.

Understanding how anonymity-seeking Internet users get blocked can inform efforts both to help these users retain the right to access the Web like everyone else, and to mitigate undesirable behaviors that website hosts aim to inhibit, like spam. Khattak explained that blocks of anonymity-seeking users can be easily achieved by constructing a Tor-node blacklist to store all currently active IP addresses of exit nodes, then deny their request from servers Khattak et al., 2016. When exit nodes from the Tor network have multiple IP addresses and not all of them are used for outgoing connections, this blocking mechanism can prove ineffective, leading to both overblocking and under-blocking Khattak et al., 2016.

Another approach to blocking Tor relies on the “consensus” data.\textsuperscript{4} Blocking using this approach involves blacklisting all the IP addresses that volunteer to relay outgoing traffic, have the exit flag assigned to them by the directory authorities, and have at least two ports out of three that are 80, 443, and 6667. This approach also results in overblocking because not all the IP addresses in the pool actually get picked to be the exit address. In both cases, any non-Tor Internet users that share an IP address with a Tor exit node cannot access the blocking website. Additionally, both approaches inflict collateral damage that prevents unintended targets of the ban from reaching the websites they want.

Websites do not block anonymity tools like Tor without reason. For example, research has shown that online anonymity is sometimes associated with toxic behaviors that are hard to control Lapidot-Lefler and Barak, 2012. According to a report by Distill networks,\textsuperscript{5} 48\% of requests coming from Tor are considered malicious, consisting of spamming, scanning, and content scraping activities. Another report made by Sqreen,\textsuperscript{6} an application protection service, claims that “a user coming from Tor is between 6 and 8 times more likely to perform an attack” on their website, such as path scanning and SQL/NoSQL injection. Tor exit node operators often receive complaints of “copyright infringement, reported hacking attempts, IRC bot network con-

\textsuperscript{3}https://trac.torproject.org/projects/tor/wiki/org/doc/ListOfServicesBlockingTor
\textsuperscript{4}https://metrics.torproject.org/collector/archive/relay-descriptors/consensuses/
\textsuperscript{5}urlhttps://resources.distilnetworks.com/all-blog-posts/cloudflare-vs-tor-is-ip-blocking-causing-more-harm-than-good
\textsuperscript{6}https://blog.sqreen.io/tor-the-good-the-bad-and-the-ugly/
trols, and web page defacements” McCoy, Bauer, Grunwald, Kohno, and Sicker, 2008, the most frequent complaints of Tor users’ abusive behavior are about DCMA violations which make up 99.74% of the approximately 3 million email complaints sent to exit operators from Torservers.net Singh et al., 2017.

On the other hand, anonymity can confer important benefits, not just for the individual seeking anonymity, but also for the collective good of online communities Kang et al., 2013; Omernick and Sood, 2013. The use of pseudonyms in collaborative learning has been demonstrated to improve equity, participation rates, and creative thinking in a group of students Chester and Gwynne, n.d. On social media sites, De Choudhury et al. found that anonymity helps users discuss topics that are stigmatized Choudhury and De, 2014. Studies of anonymous behaviors on Quora have found that the anonymous contributions on the site are no worse than answers given by registered users and the only significant difference is that “with anonymous answers, social appreciation correlated with the answer’s length” Mathew, Dutt, Maity, Goyal, and Mukherjee, 2018. Furthermore, Mani et al.’s study of the domains visited by Tor users showed that 80% of the websites visited by Tor users are in the Alexa top one million, giving further evidence that Tor users are similar to the overall Internet population Mani, Wilson-Brown, Jansen, Johnson, and Sherr, 2018.

The tradeoffs between anonymity’s benefits and threats have been investigated and discussed from many perspectives, but the question of what value anonymous contributions might have in contexts where they are not allowed is difficult to answer. How does one estimate the value of something that is not happening? By examining the relatively small number of Tor edits that slipped through Wikipedia’s block between 2008-2018, we can begin to do just that. In the next sections, we explain the context of our data collection and analysis as well as the methods we used to identify a dataset of Wikipedia edits from Tor.

**Empirical Context**

*Tor*

The Tor network consists of volunteer-run servers that allow users to connect to the Internet without revealing their IP address. Rather than users making a direct connection to a destination website, Tor routes traffic through a series of relays that conceal the origin and route of the traffic. Within Tor, each relay
only knows the immediate sender and the next receiver of the data, but not the complete path that the data packet will take. The destination receives only the final relay in the route (called the ‘exit node’), not the Tor users’ original IP address. The list of all Tor nodes is published so that Tor clients can pick relays for their circuits. This public list also allows observers to determine whether or not a given IP address is a Tor exit node at the time when they are receiving traffic. Wikipedia and other websites use these lists of exit nodes to restrict traffic from the Tor network.

**Wikipedia**

As one of the largest peer-production websites, Wikipedia receives vast numbers of contributions every day. As of February 2019, the English language Wikipedia “develops at a rate of 1.8 edits per second,” with approximately 136,608 active editors, counting only those who have registered and performed at least one action—such as generating a new article or making or reverting an edit—in the past 30 days. When a registered Wikipedia editor changes something on a page, their username is credited with that edit. Wikipedia also allows people to contribute without asking them to sign up or log in. In these cases, the contributor’s IP address is credited with the change.

However, with low barriers to participation come threats of “vandalism” and poor-quality editing. Within Wikipedia, vandalism refers to the act of deliberately degrading the quality of an article either by removing part of the existing work, or adding inaccurate or damaging content. For example, adding swear words or racial slurs and deleting good content are common forms of vandalism. The Wikipedia community invests heavily to minimize and mitigate vandalism. Using a combination of scripted bots and human judgments, the Wikipedia community has developed banning mechanisms to avoid repeated attempts from individuals determined to sabotage the community’s work. For example, if someone is detected making an attempt to vandalize an article, their account’s privilege to edit on Wikipedia might be halted and the IP address of their device could be banned. Of course, this does not stop more tech-savvy saboteurs from using methods to change their online identities and continue damaging acts.

The earliest public record of Wikipedia blocking Tor exit nodes is in Oc-

---

tor 2005 following discussion about the role of anonymous proxies.\textsuperscript{10} In the archives of Wikipedia’s public mailing lists, the origins of mistrust toward IP editors and anonymous proxies was evident from the early years of the project when as a small number of editors worked to establish and maintain the quality of a growing corpus of articles. In posts from 2002\textsuperscript{11} and 2004\textsuperscript{12} by Wikipedia’s founder, it is clear that users without accounts were treated differently and that anonymous proxies were viewed as creating problems for the Wikipedia community. Recent research demonstrates why service providers that support open collaboration projects like Wikipedia generally welcome different forms of anonymous contributions—mainly because it lowers the barriers to participation. However, service provider perceptions of what anonymity is good for differs from the experiences of contributors who seek privacy, particularly those who use Tor (McDonald et al., 2019).

Although Wikipedia participants have discussed unblocking many times, the Tor network remains blocked.

**TOR EDITS TO WIKIPEDIA**

*Identifying Tor edits*

Tor editors appear just like contributors from other non-registered editors on Wikipedia: attributed to an IP address. To identify edits as coming from Tor, we first used a complete history database dump of the English Wikipedia\textsuperscript{13} and obtained metadata of all revisions made on Wikipedia up to March 1, 2018. This metadata included revision ID, revision date, editor’s username or IP address, article ID or title, and article “namespace” (a piece of metadata used to categorize types of pages on Wikipedia). We also used a Python library called *mwreverts* to detect whether or not a revision was subsequently reverted by someone else, and whether or not it was a revert action itself, undoing someone else’s work.

The Tor metrics site maintains the list of volunteer exit nodes.\textsuperscript{14} As the name suggests, the exit list consists of “known exits and corresponding exit IP addresses available in a specific format.” Exit list data goes back to February 22, 2010, and is updated and archived every hour. Each archive has details

\textsuperscript{11}https://lists.wikimedia.org/pipermail/wikien-l/2002-November/000087.html
\textsuperscript{12}https://lists.wikimedia.org/pipermail/wikien-l/2004-February/010659.html
\textsuperscript{13}https://dumps.wikimedia.org/
\textsuperscript{14}https://metrics.torproject.org/collector.html
of exit nodes available at the time the list was produced. Most websites that restrict access from Tor, including Wikipedia, have relied on this list. After consulting with the Tor metrics team, we learned that this information does not give us a complete picture.

Before a node is picked to be an exit node, the Tor network uses their dedicated servers to compile a list of all volunteer Tor relays to examine the node’s status to determine whether or not it meets the requirements necessary to function as part of the network. These dedicated servers are called directory authorities, and they are also in charge of making the available and eligible relays reach a consensus to form a network. Once a consensus is reached, the exit nodes become effective at the time indicated by the directory authorities. This often happens hours before they are recorded on the exit list. This gap, combined with the fact that the archived record of the consensus data goes back to October 27, 2007,¹⁵ led us to combine both lists of exit nodes to construct a list all exit nodes and the specific time periods in which they were active. Using all the data collected, we crosschecked and identified any Wikipedia revision that came from an IP address during a time it was listed as an exit node on the Tor network, in either the exit list or the consensus data. Because some exit nodes recorded in the consensus data might end up

¹⁵https://metrics.torproject.org/collector/archive/relay-descriptors/consensuses/
not being used as part of the Tor network, we queried the timestamps of our newfound revisions on the Tor relay search tool to verify that the IP addresses were indeed active exit nodes around the same time. We extracted and found a total of 11,363 edits on English Wikipedia made by Tor users, between 2007 to 2018. We also found Tor edits in other languages, which we will discuss in later sections.

Figure 2 displays the number of Tor edits to English Wikipedia per month over time. The spikes in the graph tell us that the edits are distributed unevenly. This suggests that there were occasions Wikipedia failed to block certain exit nodes, so streams of revisions were able to slip through systematically before the Wikipedia community took notice and blocked them again. This happened at least five times before late 2013, when the edit trend finally died down and failed to rise back up again.

How Wikipedia blocked Tor over time

To better understand why Tor users were able to edit on Wikipedia at certain times but not others, we sought to understand Wikipedia’s Tor blocking mechanisms. Upon investigation, we found that there are two ways of banning Tor nodes: (1) using the TorBlock extension for MediaWiki, the software that runs Wikipedia, or (2) by explicitly blacklisting exit node IP addresses in Wikipedia’s server. In 2008, Wikipedia started using the TorBlock extension. TorBlock is a script that “automatically applies restrictions to Tor exit node’s access to the wiki’s front-door server.” This extension preemptively limits access from all active Tor nodes by pulling the current exit list published by Tor described in §. The obvious benefit of using TorBlock is that only current Tor exit nodes are blacklisted by the script to prevent account creation and editing on Wikipedia. As soon as IP addresses stop volunteering as Tor exit nodes, they are restored to full access by TorBlock. However, as stated by an admin named Shirik, the TorBlock extension did not seem to work well initially and also went down occasionally. As a result, Wikipedia administrators issued bans manually.

Using the publicly available data Wikipedia keeps on bans, we traced the blocking of known Tor IPs from 2007 to 2018. Wikipedia’s block log provides details about the timestamp of each block action, the enforcer’s username,

\[\text{16}https://www.mediawiki.org/wiki/Extension:TorBlock\]
\[\text{17}https://en.wikipedia.org/wiki/User:TorNodeBot\]
Table 1: Block actions against Tor IP addresses

<table>
<thead>
<tr>
<th>Block actions</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block actions against all Tor IPs</td>
<td>45130</td>
</tr>
<tr>
<td>Block actions against Tor IPs with at least one edit</td>
<td>4964</td>
</tr>
<tr>
<td>Number of Tor IPs blocked</td>
<td>32947</td>
</tr>
<tr>
<td>Number of Tor IPs with at least 1 edit blocked</td>
<td>2148</td>
</tr>
<tr>
<td>Block actions explicitly due to vandalism</td>
<td>532</td>
</tr>
</tbody>
</table>

Unsurprisingly, most of IPs in this list are described as being blocked simply because they are Tor exit nodes. This is expected because the Wikipedia community had already agreed to ban contributors from using Tor. Table 1 gives us the overview of the ban actions against Tor IP addresses over the course of 11 years. There were a total of 45,130 ban actions against IP addresses that were used as Tor exit nodes during this period. Roughly 11% of these bans were against Tor IPs that successfully made at least one edit, which only accounts for 6.5% of all addresses blocked. This provides evidence that many IP addresses were preemptively banned by Wikipedia before anyone had used them to edit. We found that fewer than 1% of the block actions explicitly state that they are due to vandalism.

Our data on block actions show that, initially, Tor blocks were mainly handled by administrators. Additionally, the duration of blocks from 2008 to 2009 were typically long in duration (from one year up to five years). This nearly always resulted in overblocking as IP addresses usually spend a very limited amount of time volunteering as Tor exit nodes. Banning these IPs for long durations prevented these addresses from editing on Wikipedia even when they were no longer Tor nodes. From 2010 to early 2014, Wikipedia started employing bots to automatically spot and block Tor nodes. During this period, typical ban durations reduced significantly to only two weeks. Tor users could once again edit from the banned nodes after two weeks. Although many exit nodes are only active for a portion of this block period, some large nodes are active for much longer. As a result, we found many nodes were blocked multiple times with multiple edits made between bans. Additionally, using a security loophole that the TorBlock extension failed to spot, Tor users frequently slipped through the IP checker. These events may explain the spikes we see in Figure 2: a sharp drop in the number of Tor ed-

18https://en.wikipedia.org/w/index.php?title=Special%3ALog&type=block
its from 2008 to 2009, and the frequent spike in edits from 2010 to 2013. A Wikipedia administrator named Shirik explained that the TorBlock tool only checks for the current list of Tor nodes, but when some of them are shut off abruptly, their server descriptors are no longer published on the exit list. If nodes are used as Tor nodes after, they do not reappear on the list for some time and the TorBlock extension would fail to notice. As a result, Shirik wrote an automated tool named TorNodeBot to spot and block on any Tor node that has access to editing Wikipedia. TorNodeBot was active from 2010 to 2014 and is recorded to have successfully blocked 32,123 IP addresses during this period. The deactivation of TorNodeBot in early 2014, along with the significant drop of Tor edits and banning actions against Tor nodes, suggests that the TorBlock extension started working as intended at this point in time.

COMPARING TOR EDITORS TO OTHERS ON ENGLISH WIKIPEDIA

Were these attempts to block Tor keeping out good contributions over the years? To answer this question, we take a closer look at the revisions themselves.

In doing so, we focus on Wikipedia’s article pages. Although its article pages are the most visible, Wikipedia contains many other pages devoted to discussion, coordination, user profiles, policy, and more. With the exception of § which considers contributions across all pages in Wikipedia, our analysis is restricted to edits made to the content of articles pages (also known as “namespace 0”). We focus our analysis to article pages for two reasons. First, article production is the primary work of the Wikipedia community and hence contributions here have the potential to be of the greatest value. Second, the nature of article contributions lend themselves to large-scale computational analysis better than discussions about policy, or social interactions, etc., which require substantial interpretation in order to be assessed for value.

Comparison sets

In order to assess the value of contributions, we used several measures that were developed within the Wikipedia community and by social computing researchers with the goal of supporting editors in maintaining the quality of the encyclopedia. A set of measures was drawn from a group of machine learning tools called ORES, which classifies edits in terms of the likelihood that they

---

are “good faith” and “damaging” Halfaker, Geiger, Morgan, Sarabadani, and Wight, 2018. A good faith edit is one which seems to be made with the intention of building up a quality encyclopedia—even though mistakes may have been made in the contribution. A damaging edit is one which is not acceptable for the standards of the encyclopedia and should be reverted. Damaging edits may have been made with good intentions, but are ultimately unwelcome.

In addition to our dataset of Tor edits, we developed datasets from three comparison groups: IP editors, first-time registered editors, and registered editors. IP editors are those users who edit Wikipedia without making an account, and hence their edits are credited to their IP addresses, but who are not using Tor. The second group includes registered editors making their first contribution. The third comparison group includes more experienced registered users, who have made more than one edit using the same account name before. For each of these populations, we cannot know if the people editing have other accounts or if they have contributed from other IP addresses.

We randomly pick the same number of revisions for each group such that they are time-matched with the original dataset. We do so by determining the number of edits made each month by Tor users, then randomly picking the same number of edits made by each user group within the same month. Our analysis includes only edits made before January 1, 2014 because the number of edits made from Tor shrinks drastically by the end of 2013 due to an improvement to the TorBlock extension.

Measuring contribution quality using reversion rates

Our first indicator of edit quality is whether an edit has been reverted. Reverting is often regarded as the main way to respond to low quality contributions and vandalism Kittur, Suh, Pendleton, and Chi, 2007; therefore, reversion rates can give us insight into how the efforts of an editor or group of editors are perceived by the Wikipedia community. We examine the reversion rate
of each set of edits, both in general and over time. Figure 3 plots how the reversion rate of articles pages changes over time for each groups of editors. Looking at the reversion percentage of Tor edits, we can see that the rate of change of these reversion percentages is higher than other groups, whose rates remain relatively. The trends of the reversion rate for Tor edits is bursty but appears to rise over time. This suggests that when there is an influx of Tor edits slipping through Wikipedia’s ban mechanism, they are also more likely to be reverted. All of the other groups exhibit a slight downward trend over time. Overall, 41.12% of Tor edits are reverted, while 29.33% of IP edits, 33.53% of first-time registered edits, and 5.5% of registered edits are.

Research has shown that several factors can affect the likelihood of an edit being reverted. One factor is whether or not an edit is itself a revert of someone else’s work. Buriol et al. examined the distribution of article reverts and found out that even though revert is a tool aimed to fight vandalism, reverting is also frequently used in “edit wars” when “editors who disagree about the content of a page repeatedly override each other’s contributions,”\textsuperscript{21} changing the content of the page back and forth between versions Buriol, Castillo, Donato, Leonardi, and Millozzi, 2006. In November 2004, the Wikipedia community issued a guideline known as the three-revert rule (3RR), which prohibits an editor from performing “more than three reverts, in whole or in part, whether involving the same or different material, on a single page within a 24-hour period.” Anyone who violates this rule is at risk of being blocked by Wikipedia administrators. Because this rule creates an incentive for editors who want to make multiple reverts to do so from an IP address rather than an account, Tor might be a useful tool for circumventing this rule.

To identify evidence of edit wars, we examined the revision history of Tor edits in chronological order. While the three-revert rule states that an editor cannot make more than three reverts on a single page within a 24-hour period, it is easy to see that Tor users can simply change to a different exit address to avoid being flagged by automated tools and getting banned. As a result, we found it reasonable to treat any series of more than two reverts made on a single page within a 24-hour period from IP addresses associated with Tor exit nodes as a violation of the three-revert rule. Among 1,577 Tor revert actions, we detected 59 3RR violations with a total of 364 revert actions made across 42 different articles. This suggests that roughly a quarter of revert actions made by Tor users are made to circumvent the rule against edit warring. While

\textsuperscript{21}https://en.wikipedia.org/wiki/Wikipedia:Edit_warring
Table 2: Users’ overall behavior of using the revert power and their effectiveness

<table>
<thead>
<tr>
<th>Group</th>
<th>Revert actions</th>
<th>Reverts kept</th>
<th>Non-reverts kept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tor users</td>
<td>1577 (13.88%)</td>
<td>399 (25.30%)</td>
<td>3512 (36.42%)</td>
</tr>
<tr>
<td>IP users</td>
<td>411 (3.61%)</td>
<td>178 (43.30%)</td>
<td>3245 (29.64%)</td>
</tr>
<tr>
<td>First-time registered users</td>
<td>448 (3.94%)</td>
<td>172 (38.39%)</td>
<td>6550 (33.5%)</td>
</tr>
<tr>
<td>Registered users</td>
<td>1327 (11.67%)</td>
<td>1049 (79.05%)</td>
<td>9483 (94.5%)</td>
</tr>
</tbody>
</table>

the edit wars in our dataset rarely lasted more than several days, and most of these violations did not last long before their IP addresses were banned by the administrators, this analysis provides signs that Tor editors abused the power to revert while undetected.

Because our comparison datasets are random samples of edits, they are unlikely to include repeated edits by the same users to the same article. As a result, we cannot easily compare other groups to Tor editors in terms of user participation in edit wars. However, we can conduct a related comparison of whether or not reverts by users in our dataset are reverted. A study of user contributions by Javanmardi et al. in 2009 Javanmardi, Ganjisaffar, Lopes, and Baldi, 2009 showed that IP editors’ contributions were twice as likely to be reverted, and that registered users were almost three times more likely to revert another user as IP editors. Our results in Table 2 suggest that Tor editors are more likely to revert someone else’s work than other types of editors are. As measured by reverts, contributors by registered editors are substantially higher quality than other groups across the board. We also found that Tor editors reverts were much more prone to be nullified. When we exclude reverted reverts and only consider normal revisions made by each group of editors, we find that Tor edits are not significantly more likely to be reverted when compared to IP editors and first-time registered editors.

Measuring contribution quality using ORES

Reversion rate is not a perfect measure of edit quality. Revisions made by Tor users might be reverted because they violated a community rule, regardless of how constructive their attempts were. Another measure of contribution quality in Wikipedia comes from ORES—a machine learning system operated by the Wikimedia Foundation to automatically categorize the quality of both articles and individual contributions on Wikipedia Halfaker et al., 2018.22 The ORES project sources edit quality judgments from the Wikipedia community

22https://www.mediawiki.org/wiki/ORES
to train their machine learning model using 24 different features for English Wikipedia, and with some modified features for other languages Q. V. Dang and Ignat, 2016.

While no gold standard set of features for assessing the quality of work on Wikipedia exists Q. Dang and Ignat, 2016, ORES draws features from previous research Warncke-Wang, Ayukaev, Hecht, and Terveen, 2015; Warncke-Wang, Cosley, and Riedl, 2013. Features include the presence of “bad words,” informal language, whether words appear in a dictionary, repeated characters, white space, uppercase letters, and so on. Other features are related to the amount of text, references, and external links, added or removed in a revision. Other, features use metadata about articles including the presence of edit summaries as well as features drawn from the contributors of revisions. The specific list of features differs by language and a full list is available in the publicly available ORES source code.23 Because ORES uses features related to users’ rights and correlated with experience, ORES scores are systematically biased toward treating edits by IP editors and inexperienced users as of lower quality (Halfaker et al., 2018). As a result, our estimates of quality for edits by users without accounts, including Tor editors, are systematically biased downward compared to our estimates for registered users. As a result, our ORES-based estimates of Tor edit quality are very likely conservative.

Our work focuses on the two ORES models that assess edit quality: the damaging model and the goodfaith model. These ORES models are trained using data collected from experienced Wikipedia editors in crowd-sourced labeling campaigns. These two models focus on two similar-sounding but distinct concepts. Damaging edits cause damage to the article as judged by Wikipedia editors which can occur regardless of the intention of the contributor. Good faith edits are those that Wikipedia editors perceive as reflecting the work of a well-meaning editor. For example, consider an editor who wishes to provide footnotes and reference links for many key terms in an article but fails to follow the guidelines and templates required by the Wikipedia community. While the editor in this scenario is acting in good faith, their contribution may nonetheless be unwanted and judged as damaging. We considered both good faith and damaging models in our analysis of Tor edits.

In the top panel of Figure 4, we use non-parametric LOESS (Cleveland, 1979) curves to visualize how the average prediction that an edit is damaging changes over time. LOESS plots are a visualization tool that use low-

---

23https://github.com/wikimedia/editquality/tree/master/editquality/feature_lists
order polynomial regression on each datapoint to calculate a smoothed fit line which describes the data as a weighted moving average. The grey bands represent standard errors around LOESS estimates over time. Because of the bursty nature of the Tor edit volume, our data is not distributed evenly over time. The bottom colored area of each subplot shows the density of data over time—solid colors indicate more data and lighter colors indicate less. These plots show that edits by Tor editors are similar to those by IP editors and those by first-time editors in terms of quality. The top half panel of Figure 4 predicts whether or not an edit made by a user in one of our groups is with good faith and tells a similar story. Tor users’ behavior is similar to the IP editors and the first-time editors groups. Compared with the registered editors, it is worse.
To provide a statistical test of the differences between the groups shown in Figure 4, we fit two logistic regression models on our measures of quality with the editor group as our independent variable. Table 3 reports the coefficients from this analysis with 95% confidence intervals in brackets. We estimated a model using fixed effect controls for month and found the results substantively unchanged.

The good faith analysis shown in the first column suggests that the difference between Tor users and IP users in this model is not statistically significant. Our model predicts that 71.1% of edits made by Tor users, the baseline group, are made in good faith. First-time editors are predicted to be 77.4% good faith, a relatively minor difference of 5.7% from IP and Tor editors. Our model predicts that registered editors are substantially higher in their good faith prediction of 97.7%.

Our model of damaging edits in the second column of Table 3 predicts that 41.3% of Tor edits will be damaging. The model predicts that 39.9% of IP editors will be damaging. And although this difference is statistically significant, it is not substantively so. Edits from first-time editors have a 33.8% probability of being damaging (only 7.5% lower). Contributions by registered editors are estimated to have only a 4.4% chance of being damaging. Both models suggest that Tor edits are of similar quality to other editors without accounts and to registered contributors making their first edits.

**Topic Modeling**

Although average quality may be similar, Tor editors may differ systematically from other editors in terms of what they choose to edit. Moreover, knowing what kind of topics Tor users edit might provide insight into Tor users’ reasons for seeking anonymity and the value of their contributions. For example, Tor users might pay more attention to matters that are sensitive and controversial. The Wikipedia category system seems like it might be useful in this regard. Unfortunately, it is an incredibly granular human-curated graph that is poorly suited to the construction of coarse comparisons across broad selections of articles Thornton and McDonald, 2012. Topic modeling may assist such an exploration by offering clusters of keywords interpretable as general semantic topics present in a collection of documents. One of the most popular topic modeling techniques is called *Latent Dirichlet Allocation* (LDA)—a generative probabilistic model for collections of discrete data such as text corpora developed by Blei et al. in 2003 Blei, Ng, and Jordan, 2003. Ma-
Table 3: Logistic Regression Results: Prediction that an Edit was Good Faith, Damaging; Tor Editors are the Reference Category.

<table>
<thead>
<tr>
<th>Good Faith</th>
<th>Damaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.90*</td>
</tr>
<tr>
<td></td>
<td>[0.85; 0.95]</td>
</tr>
<tr>
<td>IP Editors</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>[-0.03; 0.10]</td>
</tr>
<tr>
<td>First Edits</td>
<td>0.33*</td>
</tr>
<tr>
<td></td>
<td>[0.26; 0.40]</td>
</tr>
<tr>
<td>Reg Edits</td>
<td>2.87*</td>
</tr>
<tr>
<td></td>
<td>[2.73; 3.00]</td>
</tr>
</tbody>
</table>

| AIC        | 28709.73  | 37843.85  |
| BIC        | 28744.07  | 37878.20  |
| Log Likelihood | -14350.86 | -18917.93 |
| Deviance   | 9627.94   | 10486.61  |
| Num. obs.  | 39585     | 39594     |

\* 0 outside confidence interval. Ranges are 95% confidence interval.

*chinese Learning for Language Toolkit* (MALLET) McCallum, 2002 provides a widely used way to use LDA. While topic models are not known for their stability, they should still be useful for the purpose of comparing articles edited by our four groups of users. Given a list of documents and a number of topics, MALLET estimates a set of probability distributions of topics over the vocabulary of unique words. With these probability distributions and a further inspection of the keywords MALLET outputs, we can have a reasonable estimation of the topics that Tor editors and other groups of editors pay attention to.

Using our datasets of edits, we identify all the articles edited by each group. Next, we mine all textual content of these articles and then process them through MALLET to produce keywords and their probability distributions. There is no optimal number of topics for all situations, and the choice of the number of topic must be selected. We fit LDA topic models with 20 topics because doing so resulted in interpretable topics that seemed to reflect broad non-overlapping categories of encyclopedia articles. All other parameters needed for the LDA algorithm are run with default values in MALLET. Even with only 20 topics, we found some topics appeared semantically similar. For example, the topic with keywords ["league", "cup", "goals", "club", "team", "season", "stadium", "football", "world", "match", "clubs", "years", "won", "final", "goal", "united", "scored", "caps", "champions", "played"] and
Table 4: Topic interpretation based on clusters of keywords

<table>
<thead>
<tr>
<th>Topics</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soccer</td>
<td>league cup goals club team season stadium football match world clubs years united won final goal scored played caps win</td>
</tr>
<tr>
<td>Sports</td>
<td>score team world match championship win open round title seed won wrestling event time champion final defeated mexico san lost</td>
</tr>
<tr>
<td>Biology</td>
<td>species food water found made large sea fish animals common red island white small called years north animal south long</td>
</tr>
<tr>
<td>Drama TV</td>
<td>back time family episode father death life man mother house season series make home son end find friend friends story</td>
</tr>
<tr>
<td>Military</td>
<td>war army military forces force battle british general air killed ship attack united u.s states troops police german soviet command</td>
</tr>
<tr>
<td>Locations</td>
<td>city area county park north river south west town station street population london state east district located road built national</td>
</tr>
<tr>
<td>Male Biographies</td>
<td>american john born james william george robert actor english player david british united york thomas michael henry charles years richard</td>
</tr>
<tr>
<td>Health</td>
<td>health disease people treatment medical found research study sexual human blood risk effects cells children studies include symptoms brain age</td>
</tr>
<tr>
<td>Music</td>
<td>album music song band released single songs tour rock chart albums number records live guitar video year top label love</td>
</tr>
<tr>
<td>Technology</td>
<td>utc system data software windows talk users internet support information version wikipedia computer network systems page mobile user content web</td>
</tr>
<tr>
<td>Physics</td>
<td>energy water system light power time form space surface high called number process heat large field mass theory temperature gas</td>
</tr>
<tr>
<td>Transportation</td>
<td>air airport aircraft company car engine international flight airlines system service cars speed model year production line design vehicles vehicle</td>
</tr>
<tr>
<td>American Football</td>
<td>season game team games player football league record teams year won coach played bowl players win championship points nfl career</td>
</tr>
<tr>
<td>Religion</td>
<td>church book century god work life world early press history published society religious time people books modern jewish women christian</td>
</tr>
<tr>
<td>Movies</td>
<td>film series show television award season episode awards role films episodes movie year september time production released actor comedy channel</td>
</tr>
<tr>
<td>Video Games</td>
<td>game games series released character comics characters japanese player japan world version time video players story battle team original unknown</td>
</tr>
<tr>
<td>Europe</td>
<td>french france century german russian empire europe european roman republic population italian language greek king germany italy russia world spanish</td>
</tr>
<tr>
<td>Asia</td>
<td>india indian chinese china pakistan tamil sri muslim khan islamic ali state dynasty islam hindu government south temple asia muslims</td>
</tr>
<tr>
<td>Education</td>
<td>school university college students education high state schools campus research science national student year center program institute medical public arts</td>
</tr>
<tr>
<td>Politics</td>
<td>states government united state party law president national public u.s court political rights act people election years international economic federal</td>
</tr>
</tbody>
</table>
Table 5: Top 5 topics for each dataset

<table>
<thead>
<tr>
<th>Tor Users</th>
<th>IP Users</th>
<th>First-time Reg. Users</th>
<th>Registered Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politics</td>
<td>Music</td>
<td>Music</td>
<td>Locations</td>
</tr>
<tr>
<td>Technology</td>
<td>Movies and TV</td>
<td>Locations</td>
<td>Music</td>
</tr>
<tr>
<td>Locations</td>
<td>Music</td>
<td>Education</td>
<td>Politics</td>
</tr>
<tr>
<td>Movies and TV</td>
<td>Politics</td>
<td>Sports</td>
<td>Movies and TV</td>
</tr>
<tr>
<td>Religion</td>
<td>Sports</td>
<td></td>
<td>Sports</td>
</tr>
</tbody>
</table>

the topic with keywords ["score", "team", "world", "match", "championship", "win", "open", "round", "title", "seed", "won", "wrestling", "event", "time", "champion", "final", "defeated", "mexico", "san", "lost"] both heavily suggest topics about sports. However, further inspection of the articles most likely to belong to each group shows that the first topic focuses more heavily on soccer while the second topic is more general and includes terms about wrestling and other sports. After fitting LDA topics models with MALLET, we manually interpreted each cluster of words and created an appropriate topic header. Table 4 shows the mapping of keyword collections to topic header.

To gain insight into the degree to which the distribution of topics differs across groups, we compared the proportions of Tor edits in each topic to each of our comparison groups. A Pearson’s linear correlation between topic members of Tor edits and IP edits was not significant ($\rho = 0.41; df = 18; p \leq 0.073$). However the correlation between Tor and first-time edits was moderate in size and statistically significant ($\rho = 0.72; df = 18; p \leq 0.00039$), as was the correlation between Tor and registered edits ($\rho = 0.53; df = 18; p \leq .016$). A non-parametric Spearman’s rank-order correlation shows a similar pattern of results although the correlation between Tor and registered edits was attenuated and not statistically significant. These results suggest that Tor edits, although distinct from other groups of edits, are most similar in their topic selections to first time editors—but that important differences exist among all three groups. These findings parallel our results from examination of edit quality.

Figure 5 displays the distribution of topics from 4,037 article pages that Tors users edited using a gradient where more prevalent topics are darker and less prevalent topics are lighter. While there are many horizontal bands of similar shade where the topics edited by our different sets of users are similar, we can also see some key differences. While we only drew a very small portion of the population of edits for other editor groups, the comparison
is enough to highlight the similarities between of Tor editors’ interests and other groups of editors. Table 5 compares the top 5 topics that each group most focused most on. We find that Tor editors share many common interests with other editors, and that topics such as Movies and TV and Locations are among the most popular across all groups. We see proportionally fewer contributions from Tor editors in the Sports, Soccer, and American Football topics. Compared with other kinds of users, Tor editors are more likely to contribute to articles corresponding to Politics, Technology, and Religion. Interestingly, these topics are of somewhat controversial nature. Sensitive or stigmatized topics might attract Wikipedia editors interested in using tools like Tor to conceal their identity Forte et al., 2017.

OTHER LANGUAGE WIKIPEDIA EDITIONS

Wikipedia exists in hundreds of languages. These language editions are supported by different communities of administrators and users with varying guidelines and rules. In this section, we examine the behaviors of Tor users across five of the largest and most active Wikipedia language versions. We examined all Wikipedia language editions that ORES edit quality assessment models exist for in order to identify the top 5 languages in terms of number of edits made through Tor. We found 6,020 Tor edits made on German Wikipedia (dewiki), 1,632 edits made on French Wikipedia (frwiki), 3,795 edits made on Russian Wikipedia (ruwiki), and 1,129 made on Spanish Wikipedia (eswiki), from 2007 to 2018. These correspond to 5 of the top 6 Wikipedia language editions in terms of number of edits in February 2019. With the exception of Spanish, these languages are the largest languages in countries that are among the top 10 in terms of Tor relay usage.

Figure 6 shows the LOESS smoothed number of edits made through Tor over time (similar to Figure 2). Because there were virtually no edits made after 2013, we omit the edits after 2013. Once again, we see spikes of edits indicative of ineffective ban mechanisms from 2008 to 2013. While the edits in German and Russian follow a similar trend to English, the contributions of Tor users in French and Spanish increase over time before the abrupt drop in 2013. The unlikely growth of Tor edits over time in these two languages could mean that interest in using Tor was increasing in these countries.

---

24https://www.thebalancecareers.com/topics-to-avoid-discussing-at-work-526267
25https://tools.wmflabs.org/ores-support-checklist/
27https://metrics.torproject.org/userstats-relay-table.html
Figure 7 shows the trends for the average likelihood of damaging and good faith edits across all five language editions. To visualize the difference in behaviors of Tor editors across multiple languages, we again show LOESS curves side-by-side. Compared to English, German, French, Russian, and Spanish Wikipedia receive higher probabilities of damaging edits. Figure 7 also suggests that estimated levels of good faith edits in German Wikipedia are much higher than other languages. While Russian and Spanish decrease, English and French show an increase in good faith edits from Tor, especially toward the end of the period of study.

LIMITATIONS

Our study is limited in many important ways. Because our study uses IP addresses and account names to identify editors, we cannot know exactly how usernames and IP addresses map onto people. Indeed, users may even be choosing different levels of identifiability because of the kinds of edits they wish to make. For example, a registered Wikipedia editor may use Tor for certain activities. Indeed, Wikipedia editors sometimes explicitly comment on the use of Tor for the creation of “sockpuppet” accounts and may create an account for the purpose of making a single edit. Exploratory qualitative investigation of the Tor edits in our sample suggests these behaviors do occur. We hope to use future work to engage in a more direct investigation of the specific narratives behind some of the edits we observe in this study.

Our analysis is also limited by the small sample of Tor edits that get through the ban relative to those that might occur if the ban was not in place. As shown in Table 1, most Tor nodes are blocked before they are used for a single edit. Because of the language barrier, our analysis of Tor users’ behavior across languages other than English is limited and must rely heavily on ORES as a result.

CONCLUSIONS AND IMPLICATIONS FOR DESIGN

Wikipedia’s imperfect blocking of Tor provides a unique opportunity to gain insight into the quality of contributions made by Tor users. Our findings suggest that Tor users are not substantially different from other groups of Wikipedia editors with less experience and established reputation in terms

---

28https://en.wikipedia.org/wiki/Wikipedia_talk:Blocking_policy/Tor_nodes
of quality. Research has shown that overzealous policing of newcomers may discourage potentially devoted editors from becoming a regular contributor Halfaker et al., 2013. Some of these discouraged editors are very likely Tor users.

We found that Tor editors in our dataset were quite similar to other unregistered IP editors. Although they edited similar topics, they focused more on topics related to religion, technology, and politics and less on topics related to sports and music. Using ORES, we found that the proportion of Tor and other edits that are damaging and in good faith to be similar. The one area that Tor editors differed substantially was in the proportion of their edits that are reverted. However, closer examination showed that this difference disappears when we account for the fact that Tor users’ reversion actions are more frequently reverted. When reversions of substantive edits are considered, Tor users are again statistically similar to IP editors. We also found interesting patterns across non-English languages over time, suggesting that the Tor community differs by language.

Our analysis provides evidence for the value and potential of Tor-based contributors. The Tor network is steadily growing, with approximately 2 million active users at the time of writing. Many communities around the world face Internet censorship and authoritarian surveillance. In order to be Wikipedia contributors, these communities need privacy solutions. We suggest that the potential value to be gained by creating a pathway for Tor-based contributors may exceed the potential harm. Likewise, given the advances of the privacy research community (including anonymous blacklisting tools such as Nymble Tsang, Kapadia, Cornelius, and Smith, 2011), and the improvements in automated damage-detecting tools in Wikipedia, alternatives to an outright ban on Tor-based contributions may be feasible without substantially increasing the burden already borne by the vandal-fighting efforts of the Wikipedia community.

REFERENCES


Figure 5: A raster diagram of topic proportions
Figure 6: Monthly edit trends by Tor users to different language editions: English (enwiki), German (dewiki), French(frwiki), Spain (eswiki), and Russian (ruwiki).

Figure 7: ORES prediction of Damaging and Good Faith edits, drawn as a non-parametric LOESS curve.