Generally Curious: Thematically Distinct Datasets of General Threads on 4chan/pol/

Emilija Jokubauskaitė
University of Amsterdam
e.jokubauskaite@uva.nl

Stijn Peeters
University of Amsterdam
stijn.peeters@uva.nl

Abstract

Over the second half of the 2010s, the /pol/ ('politically incorrect') forum on the 4chan image board has emerged as a space within which various extreme political ideologies are discussed and cultivated, occasionally informing off-site acts of political extremism. While previous research has often studied this space as a unified whole, it is relevant to more specifically demarcate different publics within 4chan's /pol/ board, apart from studying it as an 'amorphous blob'. This paper focuses specifically on 'generals' - recurring threads with a specific thematic focus identified by a particular vernacular phrase or tag. By identifying them it is possible to partition the board's archive into multiple distinct datasets comprising discussions about a particular topic, such as Donald Trump, the Syria war, or British politics. We provide a dataset containing 58,841 opening posts and 13,697,738 replies to those, divided over 329 thematically distinct general thread collections. In this paper we outline our data collection and query protocol, the structure of the data and its rationale, as well as a number of suggested research uses for this new data.

Introduction

The 4chan imageboard is a low-tech forum that was initially set up in 2003 for the discussion of Japanese anime (Phillips 2015; Beran 2019). More recently the forum has been popularized as a breeding ground for internet memes (Zannettou et al. 2018) and, in the last few years, as a space where extreme political ideologies are freely discussed and promoted (Tuters and Hagen 2019; Nagle 2017). The latter phenomenon is largely particular to one of the various thematically distinct discussion boards hosted on the site, the 'Politically Incorrect' board or /pol/ as it is known on 4chan itself.

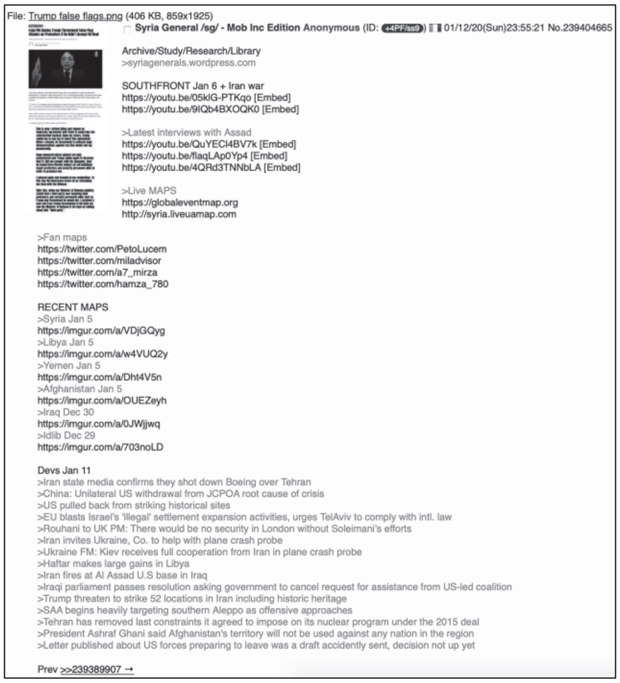
The board particularly rose to prominence in the context of the 2016 election of Donald Trump as the president of the United States of America and its aftermath. It has been argued that /pol/'s *anons* (users of the board, short for 'Anonymous') 'memed the president into office' (Beran 2019). The election further seems to have served as a catalyst for more extreme ideologies from the right side of the spectrum, an undercurrent of antisemitism and xenophobia to blossom into what can now be characterized as a far-right discussion and even recruitment space (Phillips 2018). Such a profile of the online space together with the fact of recently becoming the most popular of the 70 boards, can be regarded an impetus to study this particular image board.

However, in academic and journalistic coverage 4chan and, especially, /pol/ have often been presented as an 'amorphous entity' (Coleman 2015; Phillips, Coleman and Beyer 2017). This can be attributed to a lack of empirical approaches to dissect the large data body of short, anonymous posts; but this view is also at least partially shared by users of the site, who tend to think of themselves as part of a blob or mass. Recall, for example, the slogan 'we are legion' popularized by the early-2000s 'Anonymous' movement originating on 4chan (Beraldo 2017). Furthermore, this view of 4chan as a subculturally coherent mass is arguably shaped by the site's technical affordances. As discussed in previous research (e.g. Bernstein et al. 2011; Hagen 2018), 4chan's two most significant affordances are anonymity and ephemerality. Anonymity is an important feature of the board as on 4chan, to participate in the discussion one does not need a username, nor do they have to present their actual name — by default, someone posting on the board is identified as 'Anonymous'. While one can choose another screen name when posting, this name cannot be 'claimed' and may also be used by others, always leaving the identity of the poster in doubt.

In terms of ephemerality, 4chan's interface sorts threads by how recently they have been posted or replied to. However, after reaching a specific threshold (usually 300 replies on /pol/), threads are no longer sorted to the top of the board regardless of how recent the last activity in them was. Ultimately, they are deleted from the site altogether, usually within a couple of hours or even quicker (Bernstein et al. 2011). While external archives such as 4plebs[1] and research tools such as 4CAT (Peeters and Hagen 2018) provide the possibility to access and collect data from 4chan/pol, there have been few attempts to disassemble the dataset into separate topics of importance. Other authors have proposed various ways to go beyond the ‘amorphous entity’; for example, Hine et al. (2017), in their analysis of a then-current 4chan /pol/ dataset, provide various breakdowns of the data based on the users’ locations, type of links contained within a post, and popular shared images. A similar analysis of an updated dataset was done by Papasavva et al. (2020), who additionally utilized Natural Language Processing techniques to extract popular topics of discussion from the data. Nevertheless, such analyses do not demarcate clear publics or distinct discussion spaces *within* /pol/, and indeed few scholars have attempted to do so. We propose that one vernacular practice of 4chan/pol/ - that of posting recurring, thematically coherent and semi-consistently labeled general threads - provides an opportunity to address this. This paper looks at the cultural practice of creating general threads by 4chan users and how they may be repurposed to provide a more granular reading of the board.

In essence, general threads can be understood as a practice that aims to counter the aforementioned ephemerality of the imageboard’s software. On imageboards like 4chan and 8chan, creating and posting in general threads forms a way to maintain long-lasting thematic conversations in spite of the software’s automatic deletion of data. In practice, this concerns groups of users manually creating new discussion threads with repeated opening texts on an overarching topic, for example the war in Syria (see Figure 1). For some general threads, this is done even before the previous post is deleted. For example, ‘/ptg/ President Trump general’ has been shown to be posted on /pol/ as often as every half an hour (Bach et al. 2018), allowing for specific discussions to continue and live updates to be included. Previous research has looked at the ‘Pizzagate general’ threads in relation to the creation of the Pizzagate conspiracy theory (Tuters, Jokubauskaitė and Bach 2018) and the organization of responsibilities in two most frequent general threads, namely ‘/sg/ Syria general’ and ‘/ptg/ President Trump general’ (Bach et al. 2018). However, an approach that provides a more comprehensive overview of the board from the perspective of the general threads is missing from academic inquiry.

The dataset provided within this paper covers the temporal mapping of the most prominent themes on 4chan/pol/ through the general threads found between December 2013 and end of May 2019. In this paper, we first present the general approach to query design and data collection, followed by a description of the contents of the dataset. Finally, we conclude with a brief reflection on the purposes this data may be used for. The data itself is available on the Zenodo data repository[2].



```
File: Trump false flags.png (406 KB, 859x1925)
Syria General /sg/ - Mob Inc Edition Anonymous (ID: +4PF/ss9) 01/12/20(Sun)23:55:21 No.239404665

Archive/Study/Research/Library
>syriagenerals.wordpress.com

SOUTHFRONT Jan 6 + Iran war
https://youtu.be/05klG-PTKqo [Embed]
https://youtu.be/9IQb4BXOQK0 [Embed]

>Latest interviews with Assad
https://youtu.be/QuYECl4BV7k [Embed]
https://youtu.be/flaqLAp0Yp4 [Embed]
https://youtu.be/4QRd3TNNbLA [Embed]

>Live MAPS
https://globaleventmap.org
http://syria.liveuamap.com

>Fan maps
https://twitter.com/PetoLucem
https://twitter.com/miladvisor
https://twitter.com/a7_mirza
https://twitter.com/hamza_780

RECENT MAPS
>Syria Jan 5
https://imgur.com/a/VDjGQyg
>Libya Jan 5
https://imgur.com/a/w4VUQ2y
>Yemen Jan 5
https://imgur.com/a/Dht4V5n
>Afghanistan Jan 5
https://imgur.com/a/OUEZeyh
>Iraq Dec 30
https://imgur.com/a/0JWjjwq
>Idlib Dec 29
https://imgur.com/a/703noLD

Devs Jan 11
>Iran state media confirms they shot down Boeing over Tehran
>China: Unilateral US withdrawal from JCPOA root cause of crisis
>US pulled back from striking historical sites
>EU blasts Israel’s 'illegal' settlement expansion activities, urges TelAviv to comply with intl. law
>Rouhani to UK PM: There would be no security in London without Soleimani’s efforts
>Iran invites Ukraine, Co. to help with plane crash probe
>Ukraine FM: Kiev receives full cooperation from Iran in plane crash probe
>Haftar makes large gains in Libya
>Iran fires at Al Assad U.S base in Iraq
>Iraqi parliament passes resolution asking government to cancel request for assistance from US-led coalition
>Trump threaten to strike 52 locations in Iran including historic heritage
>SAA begins heavily targeting southern Aleppo as offensive approaches
>Tehran has removed last constraints it agreed to impose on its nuclear program under the 2015 deal
>President Ashraf Ghani said Afghanistan's territory will not be used against any nation in the region
>Letter published about US forces preparing to leave was a draft accidently sent, decision not up yet

Prev >>239389907 →
```

Figure 1. An opening post of Syria General (12th January, 2020)

Query Design

Detecting manually constructed posts that include typos, variations[3] and gradual changes[4] in the subject line poses a number of difficulties when working with a large dataset. Nonetheless, from an initial overview of a data sample and previous observations (Bach et al. 2018; Tuters, Jokubauskaitė and Bach 2018) it is known that the general threads can be recognizable through the subject of the opening post (OP), i.e. the title of a thread, usually containing the word ‘general’, like ‘president trump general’, and/or an abbreviation between two slashes, like ‘/ptg/’. Therefore, to identify tentative general threads within the broader 4chan corpus, we first extracted all between-slashes abbreviations or phrases that preceded the word ‘general’ which occurred more than 10 times across all post subjects.

The output was then cleaned by analysing a sample of opening posts retrieved with each abbreviation or phrase and subsequently removing false positives. Following that, each of the resulting datasets was manually analysed and a number of them were combined where appropriate. For

[1] archive.4plebs.org

[2] Data is available at doi.org/10.5281/zenodo.3603292

[3] For example, ‘christian general’ and ‘christianity general’

[4] The ‘national socialism general’ had iterated through a number of abbreviations and signifying symbols over time, for example, ‘natsoc’, ‘/natsoc/’, ‘/nsg/’, a swastika symbol etc., used in various combinations

instance, '/mlg/ marxism leninism' and '/mlg/ marxism-leninism' were merged as were other general threads with similar minor differences in titles or those that indicated a change in naming conventions over time. Final per-general queries were then designed based on the organized phrases and abbreviations. This resulted in 329 separate queries for as many coherent, recurring general threads, with anywhere between 8 and 14,853 opening posts in each.

It is worth noting that the approach presented produces a 'snapshot of query design' rather than a number of specific queries that one could apply on a different dataset (e.g. with another date range), since new general threads may emerge over time and their titles evolve. New vernacular may surface that necessitates other querying strategies. The contribution of this paper then is twofold; the general query design strategy may be adapted to differently demarcated data source as researchers see fit, while we provide a dataset created through a set of queries for a specific, significant period of time that is in itself a useful dataset for further research.

Data Collection

The generals' datasets presented here comprise all posts within the distinct general threads demarcated through the querying strategy described in the previous section. Posts within these general threads were collected through 4CAT, a research tool developed by the Open Intelligence Lab (Peeters and Hagen 2018). The tool allows one to query the full archive of 4chan's /pol/ board, thus providing a complete dataset from which general threads may be partitioned. This initial full dataset consisted of posts retrieved through 4chan's API. It covers the full history of the board from its beginning in late 2013 up to the moment the datasets were created, on 25 May 2019.

While 4chan's public API[5] does not explicitly allow retrieving data for research purposes, its terms of use can generally be characterized as permissive, only disallowing excessive use and misuse of the 4chan brand. Furthermore, 4chan by its nature is an anonymous platform, and usernames are not unique, minimizing the risk of the data containing personally identifying information. As such, our data collection does not violate either 4chan's own terms of use or wider laws regarding data collection, such as the EU's General Data Protection Regulation.

Queries were made for posts between 30 November 2013 (shortly after /pol/ was created) and 25 May 2019. While due to the large number of distinct general threads found on /pol/ it would be impractical to list all queries here, we provide those for three of the largest general threads here as an example:

- **President Trump General**: 'president trump general' | 'president drumpf general' | '/ptg/' (14,853 opening posts, 4,530,199 replies)
- **Syria General**: 'syria general' '/sg/' (6,213 opening posts, 1,825,142 replies)
- **Communism General**: 'communism general' -'anti-communism' -'anti communism' (2,276 opening posts, 123,588 replies)

Other queries as well as statistics per general such as the number of opening posts and replies per general may be found in the Zenodo dataset accompanying this paper[6].

These queries follow the binary search query syntax supported by 4CAT's search engine, Sphinx[7], and more generally by well-known search engines such as Google. Datasets for specific generals were created by selecting all posts that either have a title matching the query, or are a reply to such a post. Combined, these queries matched 58,841 opening posts across all distinct general threads as well as 13,697,738 replies within those threads. This is considered to cover all posts within general threads on /pol/.

While a further, in-depth analysis of these generals would be an interesting avenue for future work, we can provide an overview of the key themes these datasets (and 4chan's issue space) cover, retrieved via a cursory, inductive break-down of the topics into broad categories. The aforementioned Donald Trump-related general threads take up a category of their own with almost half (46.4%) of all opening posts in the datasets presented in this paper. In comparison, threads dedicated to other politicians all combined take up only 1.7% of all generals. Two other significant categories are those that cover discussions of left (3.5%) and right (3.4%) wing political ideologies. This, however, does not necessarily signal that some 4chan/pol users subscribe to leftist political thought, but rather that there is a small issue space where this topic is discussed, criticized and ridiculed. Finally, two other major categories include country or state politics, in which 21.7% of general threads falls, as well as the topic of war or conflict with 7.8% of the threads. The remaining 15.5% are divided among smaller categories, including religion, disasters, conspiracies, self-improvement, identity politics, et cetera.

Data Format

For each general thread, a csv (comma-separated values) file is provided containing all posts ordered by post time (from old to new). For each post, a subset of the data fields provided by the 4chan API is included:

[5] Documented at github.com/4chan/4chan-API.

[6] These may be found in the '_index.csv' file.

[7] sphinxsearch.com

- **thread_id**: The identifier of the thread the posts belong to, equivalent to the identifier of the first post in the thread.
- **id**: The identifier of the post. For opening posts, this is equivalent to thread_id.
- **timestamp and unix_timestamp**: The moment the post was created on 4chan, in both human-readable and UNIX epoch-based format.
- **body**: The full post text, stripped of HTML tags.
- **subject**: The post subject (title). This will be empty for all posts but the very first one in a thread.
- **author**: The name of the author. This will be 'Anonymous' for the vast majority of posts, as this is the default username on 4chan and not many change it.
- **image_file**: The original file name of the image attached to the post, if any.
- **image_md5**: An MD5 hash digest of the image file attached to the post, if any. This may be used to retrieve the original image via third-party 4chan archive sites.
- **country_code**: One of the distinguishing features of the /pol/ board on 4chan is that it allows people to add a country identifier to their post. By default, this is set to the country the poster's internet connection is located (via their IP address) at. This field mostly follows the ISO 3166 alpha-1 country code standard, though 4chan also allows people to identify themselves with fictional 'meme countries' for which this field may be empty or contain a non-standard code.
- **country_name**: The full name of the country the poster is identified with, if available.

While more data fields are available in the original 4CAT database, which mostly follows the 4chan API's data structure, 4chan provides a large number of such fields of which many are somewhat trivial (e.g. image thumbnail height), can be derived from the other data if needed (the amount of images in a thread) or are mostly internally relevant to 4chan (the numeric ID of a posted image). The data fields included in the dataset, then, are those most immediately relevant to researchers who want to study 4chan activity from a cultural or linguistic perspective, retaining the full information necessary to reconstruct posting activity and content for the generals while removing information that is less useful and would increase the file size of the data files. As such, the dataset comprises a complete view of the generals - within the queries' date range - for research purposes.

Conclusions and Future Work

This paper has presented 329 distinct datasets, together forming a larger data body of general threads on the 'Politically Incorrect' board on 4chan. This data positions 4chan as a collection of distinct - and crucially, *distinguishable* - conversations, as an alternative to studying it as a monolithic discussion space. As such, the datasets on offer here may contribute to investigations of 4chan's demographics, which have been particularly difficult given its anonymous nature; or to track the interests of the platform over time, providing greater understanding of the intellectual preoccupations of the 'underbelly of the internet'. As a corpus of over 13 million distinct and categorized posts, it can also serve as a basis for (socio-)linguistic analysis of 4chan's subcultures, which have been noted for their vernacular innovation (De Zeeuw and Tuters 2020).

The paper also serves as a showcase for the method implemented in retrieving the datasets. Besides using the provided data for future research work, the principal query design steps discussed here can be reproduced when dealing with a similar 4chan dataset (different timeframe or board) as well as data from other imageboards with a similar structure and utilizing a comparable cultural practice of general threads. Therefore, our hope is that the use of both the data provided and its collection methods can serve as a foundation for a more nuanced understanding of anonymous image boards as a discussion and issue space.

Acknowledgements

We would like to thank Daniel Bach and Sal Hagen for their assistance in collecting and organizing the data.

Funding

Stijn Peeters received funding from the Horizon 2020 program, ERC grant agreement number 732942.

References

Bach, D.; Tsapatsaris, M.R.; Szpirt, M.; and Custodis, L. 2018. *The Baker's Guild: The Secret Order Countering 4chan's Affordances*. Open Intelligence Initiative.

Beraldo, D. 2017. Contentious Branding: Reassembling Social Movements Through Digital Mediators. PhD Dissertation. Faculty of Social and Behavioural Sciences, Amsterdam Institute for Social Science Research, Amsterdam.

Beran, D. 2019. *It Came from Something Awful: How a Toxic Troll Army Accidentally Memed Donald Trump into Office*. New York: St. Martin's Publishing Group.

Bernstein, M.S.; Monroy-Hernández, A.; Harry, D.; André, P.; Panovich, K.; and Vargas, G. 2011. 4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

Coleman, G. 2015. *Hacker, Hoaxer, Whistleblower, Spy: The Many Faces of Anonymous*. New York: Verso.

De Zeeuw, D.; and Tuters, M. 2020. Teh Internet Is Serious Business: On the Deep Vernacular Web Imaginary. *Cultural Politics* 16(2).

Hagen, S. 2018. *Rendering legible the ephemerality of 4chan/pol/*. Open Intelligence Initiative.

Hine, G.E.; Onaolapo, J.; De Cristofaro, E.; Kourtellis, N.; and Leontiadis, I. 2017. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and its Effects on the Web. *arXiv*. arXiv:1610.03452

Nagle, A. 2017. *Kill All Normies: Online Culture Wars from 4Chan and Tumblr to Trump and The Alt-Right*. Winchester: Zero Books.

Papasavva, A.; Zanettou, S.; De Cristofaro, E.; Stringhini, G.; and Blackburn, J. 2020. Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board. *arXiv*. arXiv:2001.07487

Peeters, S., and Hagen, S. 2018. *4CAT: Capture and Analysis Toolkit*.

Phillips, W. 2018. *The Oxygen of Amplification: Better Practices for Reporting on Extremists, Antagonists, and Manipulators Online*. New York: Data & Society Research Institute.

Phillips, W. 2015. *This Is Why We Can't Have Nice Things: Mapping the Relationship Between Online Trolling and Mainstream Culture*. Cambridge: MIT Press.

Phillips, W.; Coleman, G.; and Beyer, J., 2017. *Trolling Scholars Debunk the Idea That the Alt-Right's Shitposters Have Magic Powers*. Vice.

Tuters, M., and Hagen, S. 2019. (((They))) rule: Memetic antagonism and nebulous othering on 4chan. *New Media & Society*. doi.org/10.1177/1461444819888746.

Tuters, M.; Jokubauskaitė, E.; and Bach, D. 2018. Post-Truth Protest: How 4chan Cooked Up the Pizzagate Bullshit. *M/C Journal* 21(3).

Zannettou, S.; Caulfield, T.; Blackburn, J.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Suarez-Tangil, G. 2018. On the Origins of Memes by Means of Fringe Web Communities. In *Proceedings of the Internet Measurement Conference 2018,* 188–202. Boston: Association for Computing Machinery. doi.org/10.1145/3278532.3278550