PLOS ONE

RESEARCH ARTICLE

Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus

Devin Gaffney[1]*, **J. Nathan Matias**[2]

1 Network Science Institute, Northeastern University, Boston, Massachusetts, United States of America, **2** Princeton University, Princeton, New Jersey, United States of America

* gaffney.d@husky.neu.edu

Abstract

As researchers use computational methods to study complex social behaviors at scale, the validity of this computational social science depends on the integrity of the data. On July 2, 2015, Jason Baumgartner published a dataset advertised to include "every publicly available Reddit comment" which was quickly shared on Bittorrent and the Internet Archive. This data quickly became the basis of many academic papers on topics including machine learning, social behavior, politics, breaking news, and hate speech. We have discovered substantial gaps and limitations in this dataset which may contribute to bias in the findings of that research. In this paper, we document the dataset, substantial missing observations in the dataset, and the risks to research validity from those gaps. In summary, we identify strong risks to research that considers user histories or network analysis, moderate risks to research that compares counts of participation, and lesser risk to machine learning research that avoids making representative claims about behavior and participation on Reddit.

Citation: Gaffney D, Matias JN (2018) Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. PLoS ONE 13(7): e0200162. https://doi.org/10.1371/journal.pone.0200162

Editor: Christopher M. Danforth, University of Vermont, UNITED STATES

Published: July 6, 2018

Data Availability Statement: Amended data surfaced in our work is publicly available online at http://files.pushshift.io/reddit/requests/1-10m_submissions.zip as well as https://github.com/DGaffney/baumgartner_missing_data_paper.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

1 The Baumgartner Reddit Corpus

Trace data sourced from online platforms has become an essential component for many forms of research ranging from sentiment analysis [1] to epidemiological modeling [2] and economics [3]. Dominant social platforms such as Twitter and Facebook have provided researchers with opportunities to directly study complex phenomena that, at their root, rely strongly on the nature of social interaction [4]. The reason for this, as Tufekci [5] argues, is that large platforms (specifically Twitter, in this analogy) serve as a *model organisms* for the social sciences, ones that allow for ideal conditions for measurement of many phenomena in a relatively accessible form.

On July 2, 2015, a new model organism was provided to researchers by Jason Baumgartner —a "complete" copy of one of the largest forums, Reddit, which has gained high visibility in the past several years due to events such as the Reddit blackout [6–8] and the Gamergate controversy [9]. Subsequently, many researchers have adopted the dataset, and have used it to study a wide range of questions, including the evolution of social networks [10], user

migration through online platforms [7, 11], hate speech [12], and online behavior research methodology [13], among others.

As a social news platform, Reddit hosts discussions about text posts and web links across hundreds of communities called "subreddits" [14, 15]. Discussions from public subreddits are aggregated by a variety of news aggregators to create the "front page of the web" that Reddit was founded to provide to its readers [16]. While the site also provides chatrooms and features for live discussions of breaking news [17], the most common Reddit experience is centered around top-level *submissions* and the *comments* that people post when discussing those submissions within their subreddit communities. The Baumgartner dataset follows this common experience and includes submissions and comments.

Researchers are drawn to the Baumgartner Reddit dataset for its completeness. In principle, a complete dataset improves research validity by avoiding the ambiguities of samples provided by platform application programming interfaces (APIs) and third-party data resellers [18, 19]. In this paper, we show that this dataset, as distributed and used by researchers, is not as complete as reported. We report on gaps in this data, categorize the risks to research validity from these gaps, and share collaborative re-analyses of peer-reviewed papers that have used this dataset. Finally, we conclude with reflections on the sensitivity of online behavioral research to the kinds of gaps we found in the Baumgartner Reddit Dataset.

1.1 Sequential ID analysis

The Baumgartner Reddit dataset came about through a convergence of factors: a mostly-public conversation platform, engineering details specific to the design of the Reddit system, and a creative data scientist who capitalized on these characteristics to contribute a unique dataset to public knowledge.

Many databases include the concept of an Identity column, or a column that generates an internal ID to serve as a unique reference to the row, or object, within the database. In many cases, this value auto-increments—the first value in the database assumes a value of 1, the next, a value of 2, and so forth. This number can be artificially shifted within the space—for instance engineers may partition early IDs of 1-1,000,000 for experimenting with data, for some reason, and start all production-system data created by users with ID 1,000,001. Aside from this possibility, if an object contains an ID of n, then it is plausible to assume that there are at least n objects within the database.

In personal correspondence, Baumgartner explained that this intuition led him to develop systems designed to systematically-collect all data on Reddit. Baumgartner's algorithm batches up 100 integers, converts them to the Base 36 representation that Reddit uses to represent their objects, and then queries for those objects. Reddit then returns the request with a set of all public, found objects. Baumgartner's algorithm can be run in a highly parallel environment—many batches of 100 IDs can be concurrently requested, with no need to interact with one another. On other platforms, some error may be returned if data has been deleted. With Reddit, no error is returned—instead, a truncated object reflecting that this deletion has occurred is returned. Therefore, barring technical issues, this method should provide a complete accounting for every ID within the range 1-n for all public comments and submissions within the dataset. Using this method, Baumgartner archived the public record of Reddit comments and submissions from the platform's creation through July 2015. Baumgartner has continued to provide this data as a freely-available resource.

In this analysis, we consider the full dataset as released by Baumgartner in July 2015, supplemented with updates published by Baumgartner through the end of February 2016. We also include a followup analysis extended to June 2017.

2 Diagnosing missing data

Because Reddit comments and submissions have unique, sequential IDs, we can analyze gaps in the sequence to evaluate the completeness of the dataset. We observed two kinds of missing information: dangling references (known unknowns) and gaps indicated by the absence of information that we would expect to exist given the use of sequential integers to index comments and submissions (unknown unknowns).

We discovered the completeness problem when working with this dataset for our own research. Taking a random sample of subreddits and generating a timeline of daily comments and submissions, plots showed impossible results given the architecture of Reddit: some comment timelines started earlier than their corresponding submission timelines.

The first kind of gap we discovered were dangling references. On Reddit, comments can only occur within a discussion of a submission and can only refer to other comments or submissions. In all cases, a submission would have to exist for a comment to refer to it, a relationship that is unidirectional in time. By traversing these relationships, we observed many references to missing comments and submissions. These can be thought of as "known unknowns:" comments which refer to other comments or to a parent submission, where the referred-to comment or parent submission is not contained within the Baumgartner dataset.

We also observed a second kind of gap: objects that are never referenced in the dataset but are likely missing. If all comments and submissions are given sequential integer IDs, we would expect an unbroken sequence of integers to be associated with information in the dataset. This is not the case. Consider the comments dataset: the earliest comment in the Baumgartner dataset is comment #2 and the highest is #29,484,960,643. In October 2007, the Reddit Company incremented the comment IDs by several billion IDs. When accounting for this difference, we assume that any other gaps in the sequence of comment IDs can be attributed to gaps in the dataset: we count 943,755 total potentially-missing comments up to February 2016.

Missing comment IDs could be attributed to many possible causes. These IDs could be dangling references, public information that for some reason were not returned by Reddit's systems to Baumgartner's software at that moment, unobservable technical errors within Reddit's architecture, or information that was part of a community that had set its discussions to be private. A major point of evidence in this direction is that during the Reddit blackout [6], the number of missing comments and submissions spiked significantly. In this case, due to the constraints of the platform, it is not possible to disambiguate private content from truly missing content—lookups on private IDs appear to yield no data (assuming that the API request is not done by an authenticated user who has access to the private content).

In some cases, it is possible that some of the missing IDs were never associated with any content. In correpondence, Baumgartner reported successfully retrieving many of the IDs not present in the original corpus, confirming that many of these missing IDS are genuinely associated with content. Of the initial set of 666,542 distinct comment ids and 864,598 distinct submission ids from the beginning of Reddit to February 2016 (when we initially contacted Baumgartner with missing ID lists to check), we found 101,257 existing comments and 405,911 existing submissions within those requested sets, which is substantial enough that not all "missing" data claims are spurious. Furthermore, these missing IDs are not associated with deleted content, since the Reddit platform returns information about deleted data, which is included in the Baumgartner dataset—in any event where a deleted comment is requested via the Reddit API (and subsequently, the Baumgartner corpus), a stubbed object is returned, clear of most metadata, simply stating that the object was deleted. In that event, we know the status of the object, and can confirm that it is not falsely "missing" from the dataset, but is instead truly missing in that it has been intentionally erased.

For submissions, we are less confident about the magnitude of missing unknown unknowns. While we have observed 1,539,583 "gaps" in the space of IDs for submissions through February 2016, the first submission in the Baumgartner dataset starts at 9,970,002. When searching for submissions between #1 and #9,970,001, we have successfully found some submissions, leading us to believe that millions of submissions from the early history of Reddit may be absent from this dataset, though that figure only represents an upper bound.

Deleted content, which is included in this dataset, represents a risk to validity that we do not consider here. A user who deletes even one comment in their posting history introduces many of the problems we describe in this paper, even if the fact of the comment is recorded in the Baumgartner dataset.

2.1 The per-user risk of missing data

How likely is a researcher to encounter these gaps? To address this question, we estimate the per-user risk of missing data, using a random sample of 7,400 accounts from the Baumgartner dataset.

The average user in this sample commented 6.8 times and commented 96.6 times from late January 2006 through February 2016. These averages occur on a highly skewed distribution, as illustrated by the log-histograms in Fig 1. Based on Table 1, the known maximum amount of missing comments and submissions is 943,755 and 1,539,583, respectively—dangling references are a subset of "unknown unknowns." Across the entire Baumgartner dataset, only 0.043% and 0.65% of comments and submissions, respectively, are missing. The issue has a compounding effect, since a small number of users create a large amount of the content on the

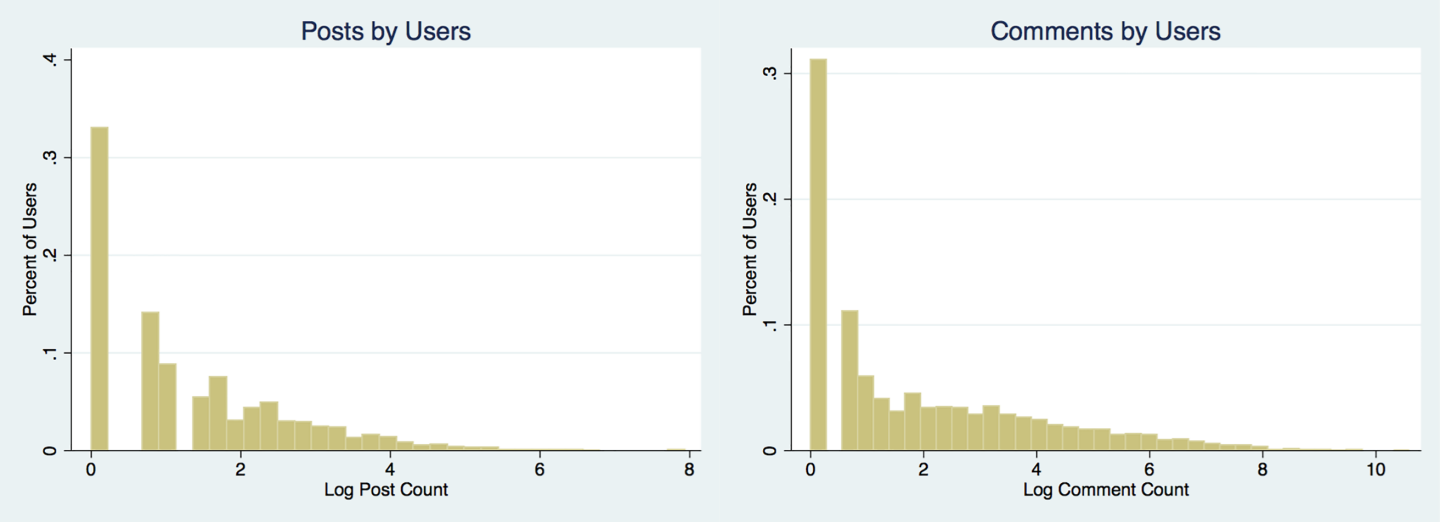

Fig 1. Log-histograms of sampled user submission and comment counts.

https://doi.org/10.1371/journal.pone.0200162.g001

Table 1. Totals for missing data in the Baumgartner dataset.

Data Type	Comments	Submissions
Dangling References (to Feb 2016)	101,257	405,911
Unknown Unknowns (to Feb 2016)	943,755	1,539,583
Unknown Unknowns (to Jun 2017)	35,801,325	27,795,423

https://doi.org/10.1371/journal.pone.0200162.t001

platform. The more posts and comments someone produces, all else being equal, the more likely their histories will be affected by the missing data issue. As we have also shown, unknown unknowns expanded dramatically in the 16 months following February 2016 and now include 36 million missing comments and 28 million missing submissions.

What is the probability of data loss for an individual Redditor history? While in reality the missing data is not uniformally distributed throughout the corpus, we can estimate the effect by compounding probabilities to assess the degree to which a user could be affected by only a small amount of missing data. Using the averages from earlier, we can calculate the risk of any individual submission r_s or comment r_c being missing simply by $\sum_c^n r_c$ and $\sum_s^n r_s$, respectively. In this case, the "average" Redditor may be exposed to a total maximum risk level of $\propto$ 4.18% likelihood for missing at least one comment and $\propto$ 4.46% for missing at least one submission. In the 7,400 individual set, approximately 2% of the sampled users had a 50% or greater chance of having a missing comment, and 2.6% of the sampled users had a 50% or greater chance of having a missing submission. These estimates were based on the census of dangling references and unknown unknowns from the beginning of the corpus to February 2016; we expect relatively similar rates in later data, since the rate of missed content has been consistent for the past several years. We offer these rough approximations to communicate a qualitative sense of how this missing data issue may create an appreciable problem for some forms of research. We include a more detailed typology of possible errors below.

2.2 Distribution of gaps across time

Far from being uniformally distributed throughout the dataset, the instances of missing data appear to be "bursty"—clustered at certain moments of time. Consequently, certain spaces in the Reddit network or certain time periods may be at greater risk of missing data than others. Importantly, we found significant gaps for comments at key moments in Reddit history that have been subjects of research, including the SOPA/PIPA protests [20] and the months leading up to the Reddit blackout [6]. Leaning on Jo et al [21], we employ a measure of "burstiness", defined as $B = \frac{\sigma_t - \mu_t}{\sigma_t + \mu_t}$, where σ_t and μ_t are the standard deviation and mean of the size of missing id gaps for each month of data from the Baumgartner corpus. This measure considers the relative dispersion of errors throughout the ID space per each month of gathered data. This measure is bounded from [-1, 1], where a score of -1 indicates completely evenly dispersed errors, and a score approaching 1 indicates that errors are located in a more concentrated set of missing blocks. Fig 2 shows many high positive burstiness scores, indicating that missing blocks are often distributed unevenly within months throughout the dataset.

Overall, Figs 3 and 4 illustrate an initially erratic distribution of errors throughout the dataset. For researchers concerned about the dispersion of missing objects, consider the dark blue line which shows the cumulative percent of missing objects or the simple percent of missing content per month of data in the medium blue. For researchers concerned with the percent of missing content to date, consider the light blue line which charts how much content appears to be missing from the beginning of the corpus until the end of our current analysis. These errors appear to occur directly within periods of substantial research interest and may affect several published results [6, 7]. While the rate of error was particularly erratic in early years, and the distribution of errors per ID gap continues to be erratic (Fig 2), the error rate per month has evened out to around 1% missing data per month.

2.3 Distribution of gaps across communities

We also considered the degree to which missing content differentially affects individual subreddits. If data from some communities were more affected by gaps than others, the gaps

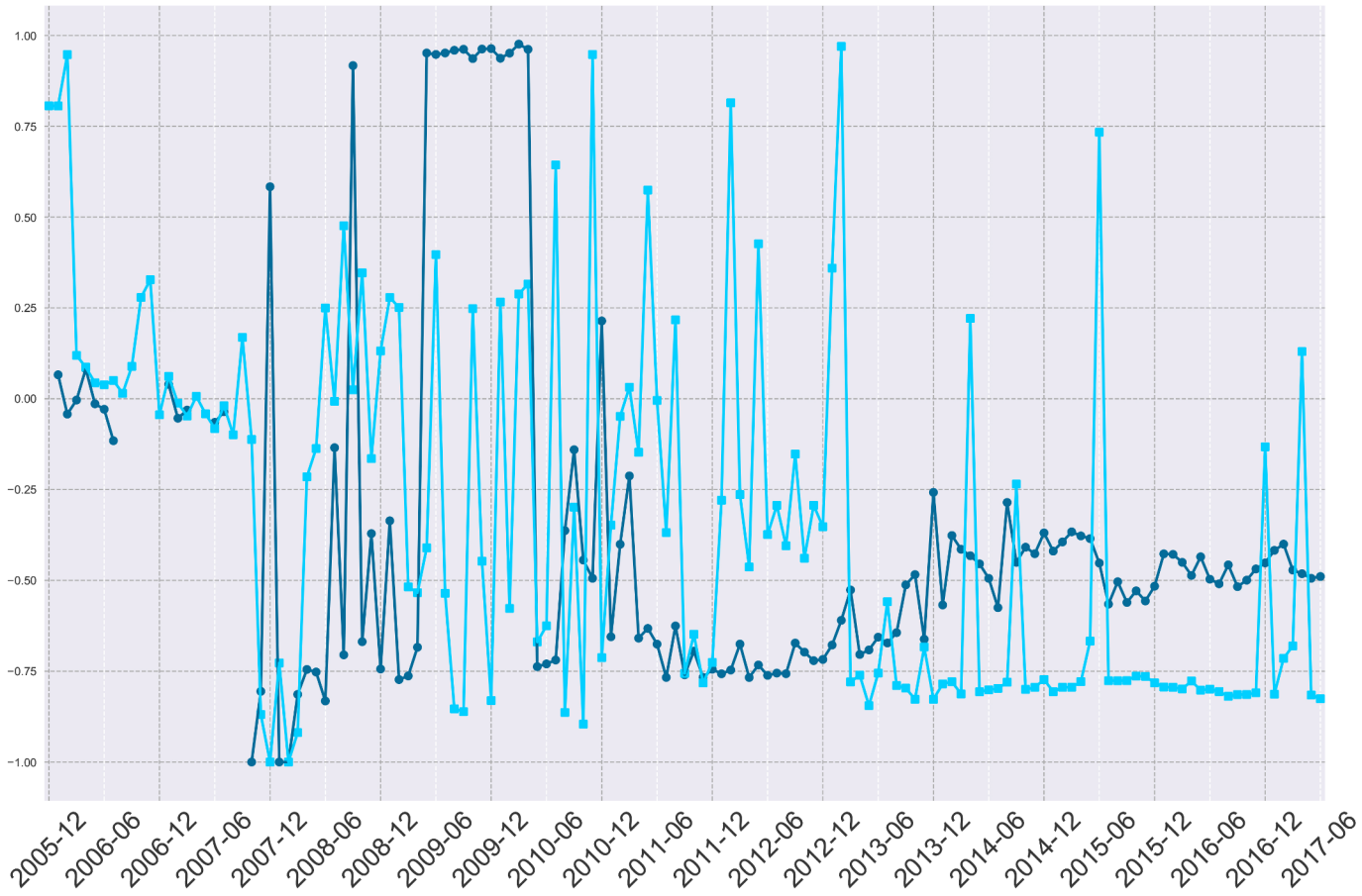

Fig 2. Burstiness of missing submissions and comments per month, 2005-June 2017.

https://doi.org/10.1371/journal.pone.0200162.g002

could influence the results of comparative research about populations communities [22]. If gaps affected communities equally, we would expect that the number of missing pieces of content monotonically rises with the number of overall pieces of content posted to a subreddit. As Fig 5 shows, we find only marginal evidence for such a supposition. While more missing content is positively and significantly associated with larger subreddits, we do not find a direct relationship. One confounding factor may be the temporal "center of gravity" of a subreddit—older subreddits are positioned at a time when more content was missing, on average, which may differentially affect older subreddits. We attempted to control for subreddit age in a multiple linear regression which accounted for the size of subreddits as well as the time at which those subreddits were created; we did not find any meaningful increase in explanatory power in the adjusted model. Table 2 provides the output from two regressions, one on missing submissions and one on missing comments, where observations are individual subreddits, and we hold the total number of known missing objects as dependent against the total number of found objects as well as the date the subreddit was created. The time at which a subreddit was created, however, is a poor proxy for the true "center of gravity" of content (i.e. the time at which a subreddit was most active), a characteristic that these models do not account for.

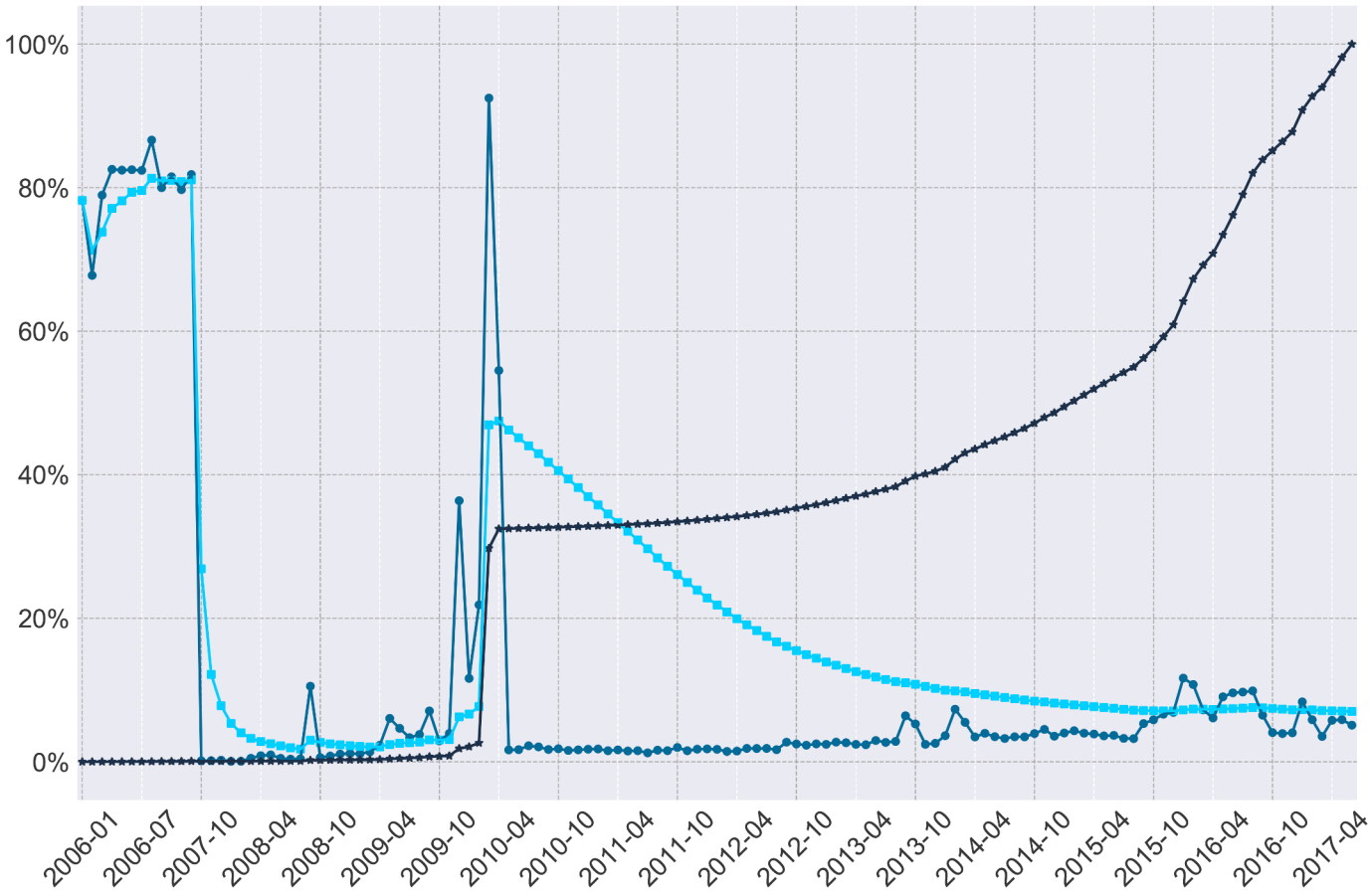

Fig 3. Varied measures of missing submissions per month. Medium blue circles denote the percent of submissions missing for each month of data, bright blue squares denote the average percent of missing submissions to date, and dark blue stars denote the cumulative total percent of missing submissions to date.

https://doi.org/10.1371/journal.pone.0200162.g003

In the above sections, we have considered the influence of potentially-missing content on analyses of users, behavior over time, and groups. We observed numerous sources of potential bias in research: a substantial percentage of users could be affected by these gaps, the gaps are not evenly distributed across time, and gaps are not evenly distributed across communities.

3 How missing data affects common research methods in computational social science

How might these gaps influence research in practice? We expect that researchers asking different kinds of questions will face different kinds of risks from missing data. In the following sections, we categorize published literature that uses this dataset and offer a typology of the risks that these gaps represent to common research methods in computational social science.

User history analysis papers face the *highest risks* from missing data, since a missing comment or submission could hide an important part of that user's history. A network analysis may fail to include a user's participation in a particular community or interaction with a key user. Furthermore, survival analyses might mis-estimate the moment of a person's departure

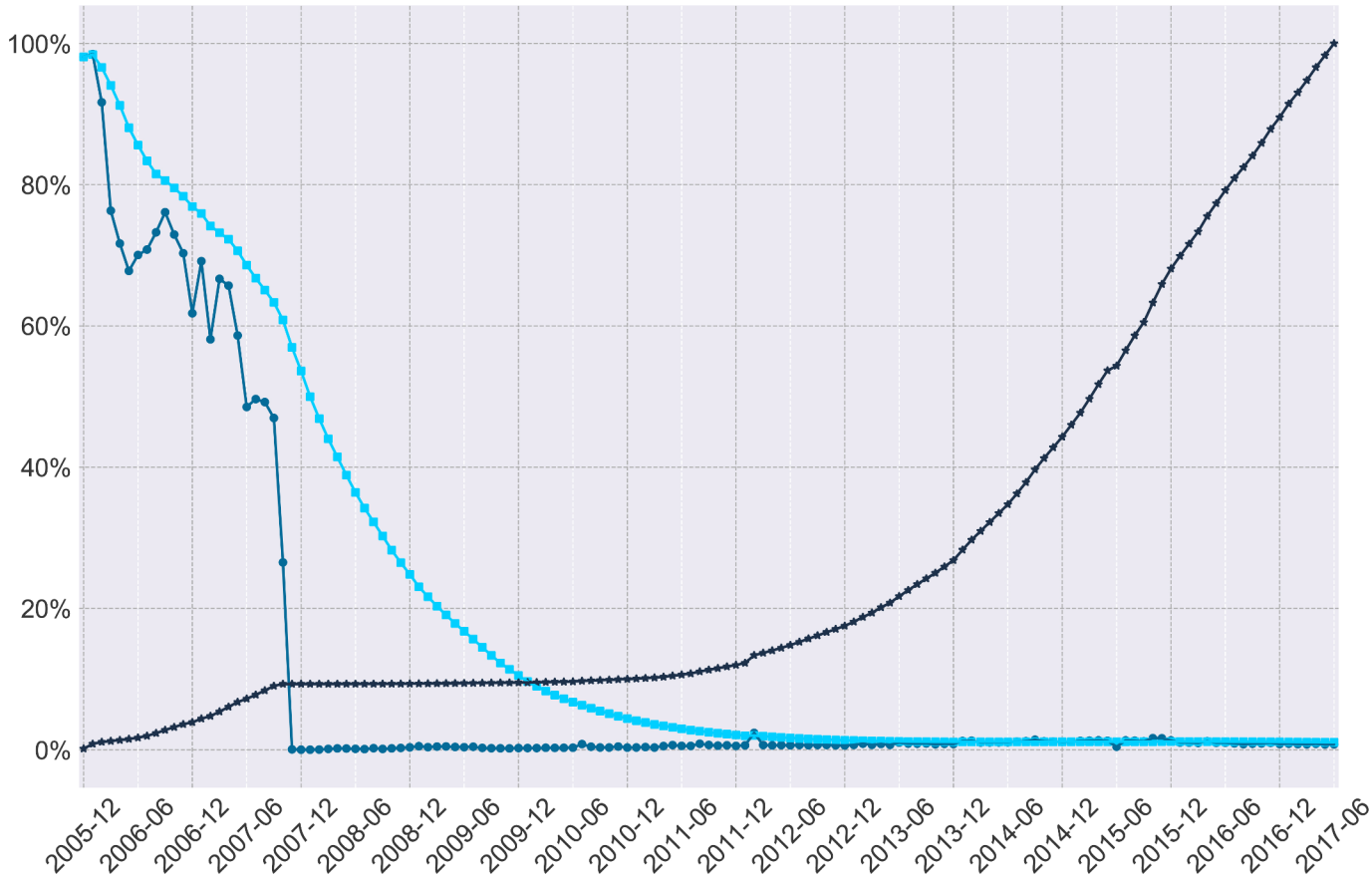

Fig 4. Varied measures of missing comments per month. Medium blue circles denote the percent of comments missing for each month of data, bright blue squares denote the average percent of missing comments to date, and dark blue stars denote the cumulative total percent of missing comments to date.

https://doi.org/10.1371/journal.pone.0200162.g004

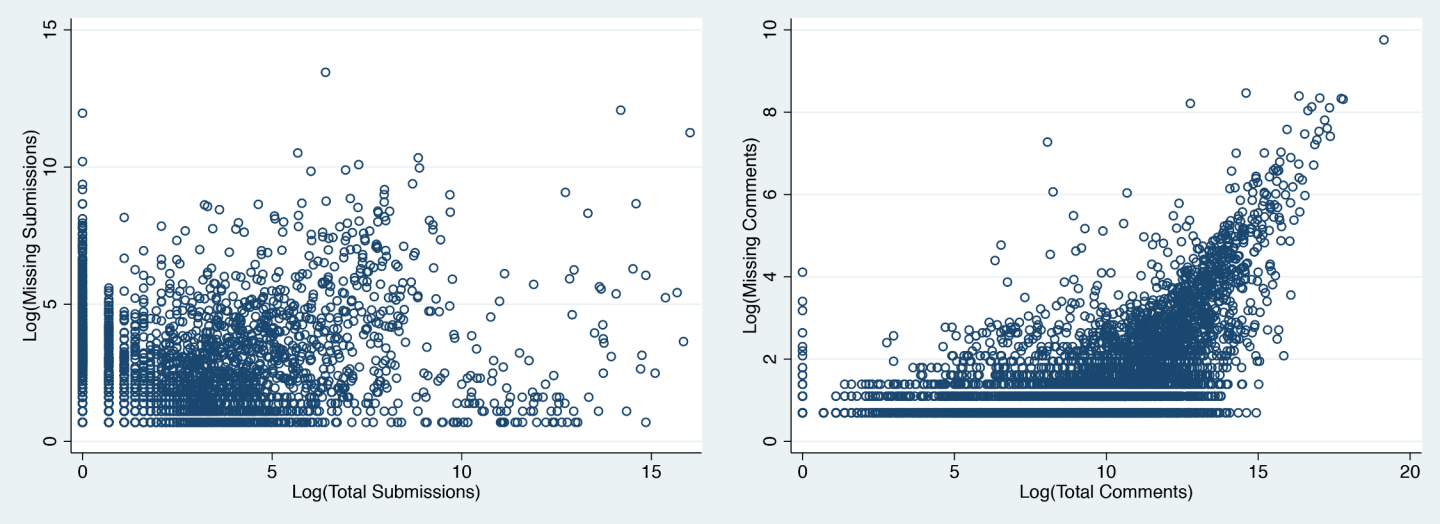

Fig 5. Gaps are not evenly distributed across communities. The total historical counts of comments per community comments are mildly correlated with the number of dangling references, while submissions are not very correlated with the number of dangling references.

https://doi.org/10.1371/journal.pone.0200162.g005

Table 2. Regression exploring the relationship between amount of missing content per subreddit and total amount of known content per subreddit, and month in which the subreddit was created. We expect that these two variables would have meaningful explanatory power for where missing content is—we find that this appears to be the case for missing comments but not for missing submissions, as evidenced by the relative R^2 values.

Variable	Submissions	Comments
Total Content	0.212*** (0.008)	0.217*** (0.006)
Month Subreddit Created	0.005*** (0.001)	0.002** (0.001)
Constant	1.198*** (0.094)	-0.518*** (0.095)
Observations	8,176	4,341
R^2	0.086	0.306
Adjusted R^2	0.086	0.305

* $p<0.1$

** $p<0.05$

*** $p<0.01$

https://doi.org/10.1371/journal.pone.0200162.t002

or their participation level. *Network analysis* papers also face *high risks*, since the presence or absence of a tie could be dependent on the missing data. *Sum analyses* that count the size or incidence rate of participation in subreddits or the use of certain kinds of language face *moderate risk*, especially when analyzing small communities and rare events. *Content analysis* that involves training machine learning systems on Reddit comments face *minimal risk* because their research rarely includes claims about the population of Reddit users.

3.1 Risk to user history analyses

Papers that test hypotheses based on user histories on Reddit may have substantial gaps in the histories that they seek to test. Analyses on user histories that consider the history in full are, in general, exposed to the highest risk—analyses that are especially sensitive to high-volume users are very likely, on average, to consider users whose histories have gaps. Hessel et al [23], for example, observes and compares sums of comment participation between subreddits, and observes the full chain of user history—Hessel et al [24] adopts a similar approach. Barbosa et al [13] compares year cohorts of individual-level behavior across all of Reddit, and as has been shown, some years are more affected by gaps than others. Additionally, the large number of potential missing submissions from Reddit's earliest years may also affect these findings. If a user history analysis requires the complete posting history between subreddits for a given user, gaps in such transmissions may constitute meaningful gaps in explaining a wide array of hypotheses.

3.2 Risks to network analyses

Some papers test network hypotheses by constructing interaction networks between users or communities, sometimes over time. Data gaps also represent a high risk to these papers, since missing submissions may result in unobserved ties in the network. Tan and Lee [11] observe histories of user accounts participating in different communities, while Fire and Guestrin [10] observe network ties over time modeled on user histories. Substantial blocks of missing data, including the potentially large amount of missing submissions from Reddit's nascency could redraw the map of community ties on the platform. Tree structures reconstructing threads are

also similarly affected, such as work by Hessel et al [25] and Fire and Guestrin [26], which through linkages of comments and submissions similarly face issues due to missing submissions (i.e. parents of threads) or comments.

3.3 Risks to research that counts and compares participation between communities

Other papers test hypotheses based on participation sums within communities. Gaps that are biased toward particular communities will represent a risk to the validity of these studies. Matias [6] observes levels of subreddit participation by moderators, observes relative participation levels of subreddit commenters in other subreddits, and observes moderator participation in "metareddits". Newell et al [7] observes comment volumes within subreddits. Barthel [27] observes comments about political candidates across Reddit during a period where many submissions are within the dataset. Barbaresi [28] analyzes German language text to identify relative commenting rates about places in Germany. Horne and Adali [29] consider posts within /r/worldnews to determine linguistic characteristics of why some news frames are more visible than others. Dosono et al [30] considers a specific set of communities associated with self-expression of Asian-American Pacific Islander (AAPI) identity on the platform.

As we showed in Fig 5, gaps do not appear to be evenly distributed across communities, since the number of missing comments and submissions per community is not strongly correlated to the number of observed comments and submissions in that community. While a simple statistical regression between the total counts of missing data and known data shows the relationship to be significant, the R^2 is low enough in both cases to lead us to conclude that studies on some subreddits could lead towards very biased results due to higher than random amounts of missing data.

In practice, we observe 78 subreddits where at least 20% of the comments are missing, and 1,755 subreddits where at least 20% of the submissions are missing. Among subreddits that have any dangling references, on average they are missing at least 35% of their submissions. The R^2 score in a model predicting the volume of a community's missing observations from the volume of observed comments and submissions only explains 30% of the variance of missing comments and 10% of the variance of missing submissions (Fig 5). The risk to any specific study will depend on the distribution of gaps across the specific communities being compared.

3.4 Risks to machine learning models

Finally, some studies train machine learning models and conduct linguistic analysis of the Baumgartner dataset. Insofar as these studies do not make claims about populations, gaps represent a minimal risk to the validity of this research. For example, Saleem et al [12] trains machine learning models on comments from particular subreddits that have since been quarantined or banned by Reddit for harmful behavior.

In our observations of communities where the mass of missing data is pooled, it seems to trend towards such communities—across the three subreddits considered in their work, one of those subreddits has a large number of dangling references: observed comments refer to 696,642 unique missing submissions in the dataset for this one community alone. Among comments, 1,100 of 1,585,014 total comments were known to be missing. Saleem, Dillon, Benesch, and Ruths have re-analyzed their data after filling some gaps and fail to find any substantial differences in the performance of their machine learning models (citation forthcoming). Furthermore, since the purpose of this kind of machine learning research is to make inferences about out-of-sample observations rather than to test hypotheses about a population, such research may be less sensitive to variation due to missing data.

4 Discussion

All datasets have biases, no matter how complete we wish them to be. In the process of designing research, conscientious researchers will study those biases, document them, and account for them as best as possible. In this paper, we have shown ways in which an influential public dataset does not represent the "complete" record that its publisher and users aspired to. We have documented per-user risks of missing data, risks from the uneven distribution of missing data over time, and risks in the uneven distribution of missing data across communities. We have outlined the risks to research validity represented by these data gaps, including some of our own work.

We have raised these issues in direct conversation with Baumgartner, who has quickly and graciously re-processed ID blocks with missing data and filled in any gaps that are able to be filled. By publication time of this paper, we believe that any missing data that can be filled will have been done so for datasets provided directly by Baumgartner up to February 2016. As of now, Baumgartner has acknowledged future steps to be taken in terms of ensuring the integrity of future data by double-checking for missing content—while his commitment to increasing the integrity of the data is not required, it is highly appreciated [31]. Data shared from any other source may still include these missing observations. Since any missing data that Reddit does not provide will still be missing from the corrected datasets, we encourage researchers to check the integrity of your data when publishing results from this dataset. Additionally, Hessel and his colleagues have provided a response to the issues raised in this work, which we have included as supporting information.

More widely, the case of this so-called complete dataset draws attention to the risks to validity from research cultures that move fast to produce new results when new data is released. While many researchers have utilized Baumgartner's generous work on this Reddit dataset to investigate important questions, too few of us questioned a "completeness" statement that shouldn't have been accepted as truth. This dataset has numerous omissions, and those issues affect different research agendas with varying levels of severity.

As researchers, we need to protect ourselves from the dazzling scale of large datasets. We encourage more people in Baumgartner's position to collect data, share it in an ethical manner, and contribute to knowledge through the research that it enables. It will not always be possible or reasonable to place strict methodological expectations upon such citizen scientists—that responsibility lies firmly on academics. We hope this paper will encourage other researchers to test their assumptions and document data quality when conducting social scientific research with large datasets that they did not collect.

Supporting information

S1 File. Response letter from Hessel, Lee, Mimno and Tan. Authors of all cited works were solicited for opportunities to respond to our findings—Jack Hessel, Lillian Lee, David Mimno, and Chenhao Tan jointly have provided a response.
(PDF)

Author Contributions

Conceptualization: Devin Gaffney, J. Nathan Matias.

Data curation: Devin Gaffney.

Formal analysis: Devin Gaffney, J. Nathan Matias.

Investigation: Devin Gaffney.

Methodology: Devin Gaffney, J. Nathan Matias.

Project administration: Devin Gaffney, J. Nathan Matias.

Resources: Devin Gaffney.

Software: Devin Gaffney.

Supervision: J. Nathan Matias.

Validation: Devin Gaffney, J. Nathan Matias.

Visualization: Devin Gaffney.

Writing – original draft: Devin Gaffney, J. Nathan Matias.

Writing – review & editing: Devin Gaffney, J. Nathan Matias.

References

1. Pak A, Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: LREc. vol. 10; 2010. p. 1320–1326.
2. Abdullah S, Wu X. An epidemic model for news spreading on twitter. In: 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence. IEEE; 2011. p. 163–169.
3. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. Journal of Computational Science. 2011; 2(1):1–8. https://doi.org/10.1016/j.jocs.2010.12.007
4. Bond RM, Fariss CJ, Jones JJ, Kramer AD, Marlow C, Settle JE, et al. A 61-million-person experiment in social influence and political mobilization. Nature. 2012; 489(7415):295–298. https://doi.org/10.1038/nature11421 PMID: 22972300
5. Tufekci Z. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. arXiv preprint arXiv:14037400. 2014;.
6. Matias JN. Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM; 2016. p. 1138–1151.
7. Newell E, Jurgens D, Saleem HM, Vala H, Sassine J, Armstrong C, et al. User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest. In: Tenth International AAAI Conference on Web and Social Media; 2016.
8. Baumgartner J. I have every publicly available Reddit comment for research. 1.7 billion comments at 250 GB compressed. Any interest in this?: datasets; 2016. https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/.
9. Massanari A. # Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. New Media & Society. 2015; p. 1461444815608807.
10. Fire M, Guestrin C. Analyzing Complex Network User Arrival Patterns and Their Effect on Network Topologies. arXiv preprint arXiv:160307445. 2016;.
11. Tan C, Lee L. All who wander: On the prevalence and characteristics of multi-community engagement. In: Proceedings of the 24th International Conference on World Wide Web. ACM; 2015. p. 1056–1066.
12. Saleem HM, Dillon K, Benesch S, Ruths D. A Web of Hate: Tackling Hateful Speech in Online Social Spaces. In: First Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS 2016); 2016.
13. Barbosa S, Cosley D, Sharma A, Cesar Jr RM. Averaging Gone Wrong: Using Time-Aware Analyses to Better Understand Behavior. In: Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee; 2016. p. 829–841.
14. Leavitt A, Clark JA. Upvoting hurricane Sandy: event-based news production processes on a social news site. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM; 2014. p. 1495–1504. Available from: http://dl.acm.org/citation.cfm?id=2557140
15. Massanari A. Participatory Culture, Community, and Play: Learning from Reddit. 2nd ed. New York: Peter Lang Publishing Inc.; 2015.
16. Massanari A. # Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. New Media & Society. 2015; p. 1461444815608807.

17. Leavitt A, Robinson JJ. The Role of Information Visibility in Network Gatekeeping: Information Aggregation on Reddit during Crisis Events. In: CSCW; 2017. p. 1246–1261.
18. Lotan G, Graeff E, Ananny M, Gaffney D, Pearce I, et al. The Arab Spring—the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. International journal of communication. 2011; 5:31.
19. Diaz F, Gamon M, Hofman JM, Kiciman E, Rothschild D. Online and Social Media Data As an Imperfect Continuous Panel Survey. PLOS ONE. 2016; 11(1):e0145406. https://doi.org/10.1371/journal.pone.0145406 PMID: 26730933
20. Benkler Y, Roberts H, Faris R, Solow-Niederman A, Etling B. Social mobilization and the networked public sphere: Mapping the SOPA-PIPA debate. Political Communication. 2015; 32(4):594–624. https://doi.org/10.1080/10584609.2014.986349
21. Jo HH, Karsai M, Kertész J, Kaski K. Circadian pattern and burstiness in mobile phone communication. New Journal of Physics. 2012; 14(1):013055. https://doi.org/10.1088/1367-2630/14/1/013055
22. Hill BM, Shaw A. Studying Populations of Online Communities. The Handbook of Networked Communication Oxford University Press, New York, NY. 2017;.
23. Hessel J, Tan C, Lee L. Science, AskScience, and BadScience: On the Coexistence of Highly Related Communities. In: Tenth International AAAI Conference on Web and Social Media; 2016.
24. Hessel J, Schofield A, Lee L, Mimno D. What do Democrats do in their Spare Time? Latent Interest Detection in Multi-Community Networks. arXiv preprint arXiv:151103371. 2015;.
25. Hessel J, Lee L, Mimno D. Cats and captions vs. creators and the clock: Comparing multimodal content to context in predicting relative popularity. In: Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee; 2017. p. 927–936.
26. Fire M, Guestrin C. The Rise and Fall of Network Stars. arXiv preprint arXiv:170606690. 2017;.
27. Barthel M. How the 2016 presidential campaign is being discussed on Reddit—Pew Research Center;. http://www.pewresearch.org/fact-tank/2016/05/26/how-the-2016-presidential-campaign-is-being-discussed-on-reddit/.
28. Barbaresi A. Collection, Description, and Visualization of the German Reddit Corpus. In: 2nd Workshop on Natural Language Processing for Computer-Mediated Communication; 2015. p. 7–11.
29. Horne BD, Adali S. The impact of crowds on news engagement: A reddit case study. arXiv preprint arXiv:170310570. 2017;.
30. Dosono B, Semaan B, Hemsley J. Exploring AAPI identity online: Political ideology as a factor affecting identity work on Reddit. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems. ACM; 2017. p. 2528–2535.
31. Baumgartner J. My response to the paper highlighting issues with data incompleteness concerning my Reddit Corpus: datasets; 2018. https://www.reddit.com/r/datasets/comments/884vkh/my_response_to_the_paper_highlighting_issues_with/.