Bias in Published Academic Papers from Gaps in a Public Reddit Archive

Devin Gaffney and Nate Matias

17 November 2016

Abstract

On July 2, 2015, Jason Baumgartner, known as /u/Stuck_in_the_Matrix on Reddit, published a dataset claimed to comprise "every publicly available Reddit comment", which was quickly shared on Bittorrent and the Internet Archive. Within a short period, the data became the basis of a series of academic papers on topics including machine learning, social behavior, politics, breaking news, and hate speech. When exploring this data in our own research, we discovered substantial gaps and limitations in the dataset which may contribute to bias in the findings of research that relies on it. This report documents our work to identify those gaps and consider the risks to research validity that they represent. In summary, we identify strong validity risks to research that considers user histories or network analysis, moderate risks to research that relies on sums of participation, and minimal risk to machine learning research that trains models from content analysis rather than making representative claims about behavior and participation on Reddit.

The Baumgartner Dataset

Trace data sourced from online platforms has become an essential component for many forms of research ranging from sentiment analysis (Pak & Paroubek, 2010) to epidemiological modeling (Abdullah & Wu, 2011) and economics (Bollen, Mao, & Zeng, 2011). 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 (Bond et al., 2012). The reason for this, as Tufekci (2014) argues, is that large platforms (specifically Twitter, in this analogy) serve as a *model organism* for the social sciences, one that allows 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 (Matias, 2016; Newell et al., 2016; Baumgartner, 2016) and the Gamergate controversy (Massanari, 2015). Subsequently, many researchers have adopted the dataset, and have used it's unique affordances to study the evolution of social networks (Fire & Guestrin, 2016b, 2016a) and user migration through online platforms (Tan & Lee, 2015; Newell et al., 2016). Aside from the technical differences in terms of the architecture of this platform as compared to other platforms dominant in research, the distinguishing factor of the Baumgartner Reddit dataset is it's completeness, which allows for a higher degree of validity in findings which need not have any ambiguities about how unsampled data may differ from samplings typically present in research on other platforms (Lotan et al., 2011). This dataset, however, is not actually complete.

Sequential ID Analysis

The reason that the entire contents of the platform could be systematically collected, and the reason that it has been shown that the dataset is in fact not complete. 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 real data created by users with ID 1,000,001. Barring this, however, if an object contains an ID of n, then it is plausible to assume that there are at least n objects within the database. By personal correspondence, Baumgartner has explained that this intuition led to the development of the systematic collection of all data on Reddit – the 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 – found objects will be returned by the request. This 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 many platforms, some error may be returned if the data has been deleted – with Reddit, no error is returned – instead, a truncated object reflecting this deletion has occurred is returned. Therefore, barring technical issues, there should be a complete accounting for every ID within the range 1-n for all comments and submissions within the dataset, which allows for a validity check of the completeness claim.

Diagnosing Missing Data

The completeness problem was found relatively early in working with the data – a random sample of subreddits was selected, and a timeline was generated for the daily counts of comments and submissions on the subreddits. Plots showed impossible results given the architecture of Reddit – some comment timelines started earlier than their corresponding submission timelines. On Reddit, comments can only refer to other comments or submissions, therefore, a submission would have to exist in order for a comment to subsequently refer to

Data Type	Known Unknowns	Unknown Unknowns
Comments	101,257	943,755
Submissions	405,911	1,539,583

Table 1: Totals for missing data in the Baumgartner dataset

it, and the order of these events is unidirectional with respect to time. Digging further, many instances of references to missing data begun to appear throughout the data. An exhaustive search through the data was therefore necessary.

Two sets of data within the dataset, comments, and submissions, ultimately have been shown to be missing data. Current work on this front has uncovered two issues with the data currently referred to as the "known unknowns" and the "unknown unknowns". Known unknowns are comments which refer to other comments or parent submissions, but the referred-to comment or parent submission is not contained within the Baumgartner dataset. "Unknown unknowns" are cases where there are gaps within the ID space of the Baumgartner dataset. The earliest comment in the Baumgartner dataset is comment #2, which allows for us to know that, between the oldest and the newest within the dataset, there are 943,755 total "gaps", or number of missing comments (which could be a known unknown, a truly unknown unknown comment that can be mined, or content not available for technical reasons (deleted data on Reddit is validly returned by the Reddit API, and included in the data as a deleted record)). Submissions are much trickier. While there are 943,755 "gaps" in the space of IDs for submissions, the first submission in the Baumgartner dataset starts at #9,970,002 – many small checks between #1 and #9,970,001 have been conducted, and submissions found, which suggests potentially significant problems with the collection process particularly in the early development of the forum.

Rough Estimates

A random sample of 7,400 accounts were selected from the Baumgartner dataset, and summary statistics were generated for the users – the average user posts 6.8 times and comments

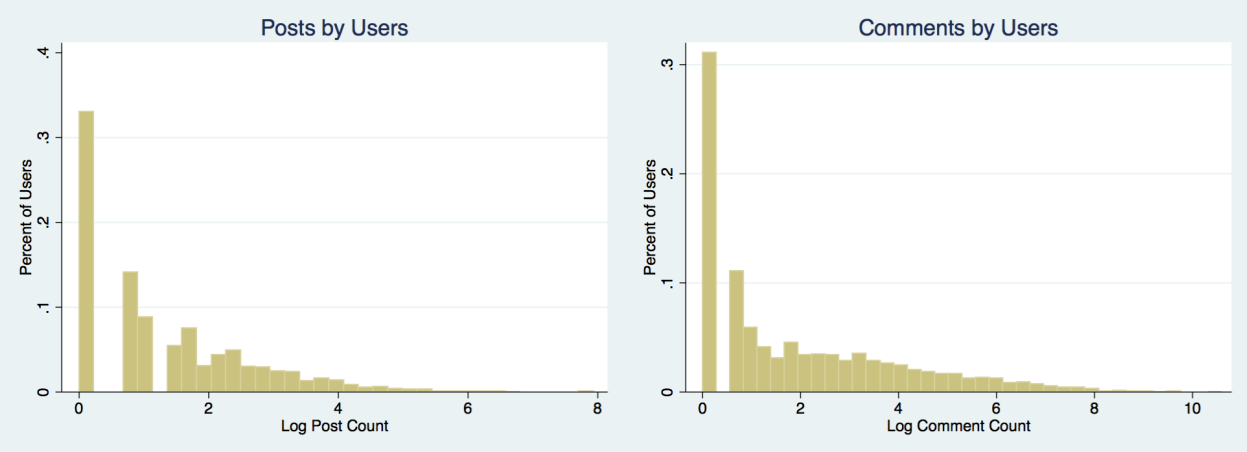

Figure 1: Histograms of sampled user Submission and Comment counts

96.6 times, though this is a highly skewed distribution, as the log-histograms in 1. Based on 1, the known maximum amount of missing comments and submissions is 943,755 and 1,539,583, respectively – "known unknowns" are a subset of "unknown unknowns". Across the platform, this is, admittedly, a very small amount of missing data – Across the entire Baumgartner dataset, only 0.043% and 0.65% of comments and submissions, respectively, are missing. The issue has a compounding effect, however: the skewedness of the dataset shows that a small number of users create a large amount of the content on the platform – the more posts and comments they generate, ceteris paribus, the more likely their histories will be affected by the missing data issue.

Conservatively, we can provide a rough accounting for the qualitative degree to which this could affect a wide range of research by considering rough probabilities of data loss for individual Redditor histories. In reality, the missing data is far from uniformally distributed through the corpus, but relaxing this fact, we can simply compound 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 are only very rough approximations to help get a qualitative sense of how this missing data issue may create an appreciable problem for some forms of research – a more considered typology of errors is considered below.

Error Distribution

Far from being uniformally distributed throughout the dataset, the instances of missing data are very bursty. This creates spaces within the data where the issue of missing data may be less of an issue or considerably more of an issue. Importantly, significant gaps were found for comments created around the time of the SOPA/PIPA protests (Benkler, Roberts, Faris, Solow-Niederman, & Etling, 2015), while significant gaps in submissions were found in the months leading up to the Reddit blackout (Matias, 2016), though this work falls short of drawing any direct causal inference in these two cases.

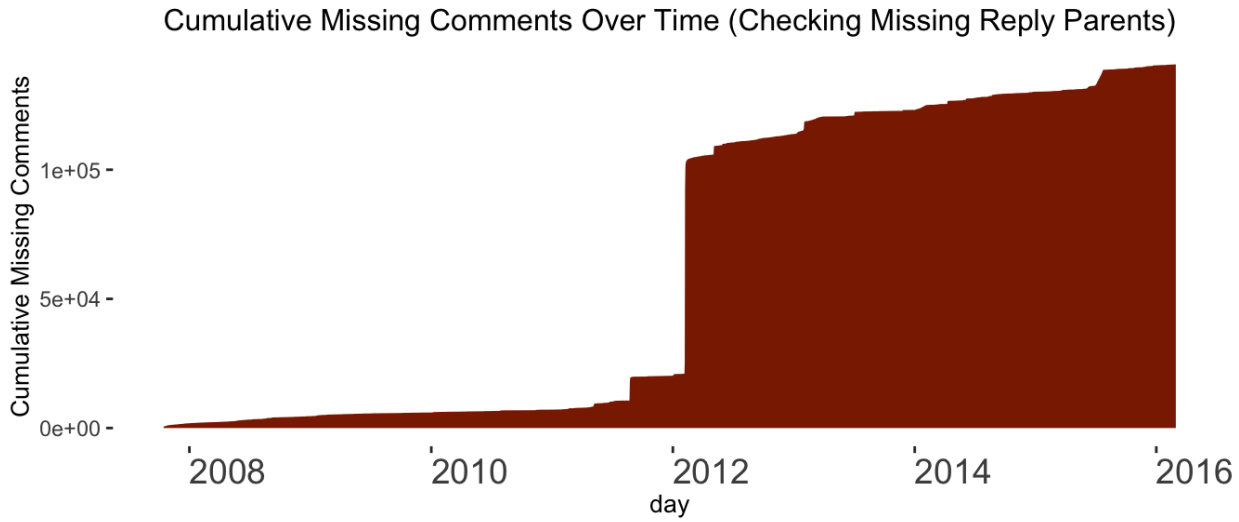

Figure 2: Cumulative missing comments by dates of "known unknowns" referencing missing comments. Note the steep rise in 2012.

Overall, figures 5, 2, 3, and 4 illustrate an erratic distribution of errors throughout the dataset – and they appear to occur directly within time spaces of inherent research interest,

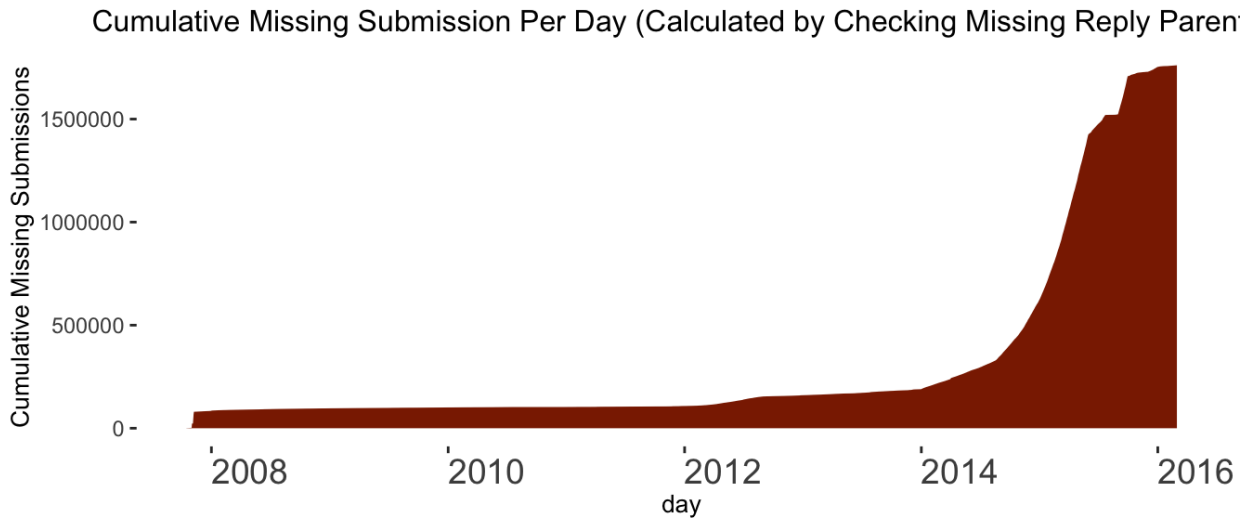

Figure 3: Cumulative missing submissions by dates of "known unknowns" referencing missing submissions. Note the steep rise over months leading to February 2016.

which may affect results found by several works already (Matias, 2016; Newell et al., 2016). As alluded to earlier, some methodological designs may be affected more significantly by these issues than others that range from strong risks to minimal risks. Reviewing current literature that has used this dataset, it is possible to convey these issues as a typology of methodological concerns.

Typology of Errors

User history analysis papers also face the **highest risks**, since a missing comment or submission could hide an important part of that users history. A network analysis may fail to include a users participation in a particular community or interaction with a key user. Furthermore, survival analyses might mis-estimate the moment of departure or participation levels due to gaps in the dataset. **Network analysis** papers 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

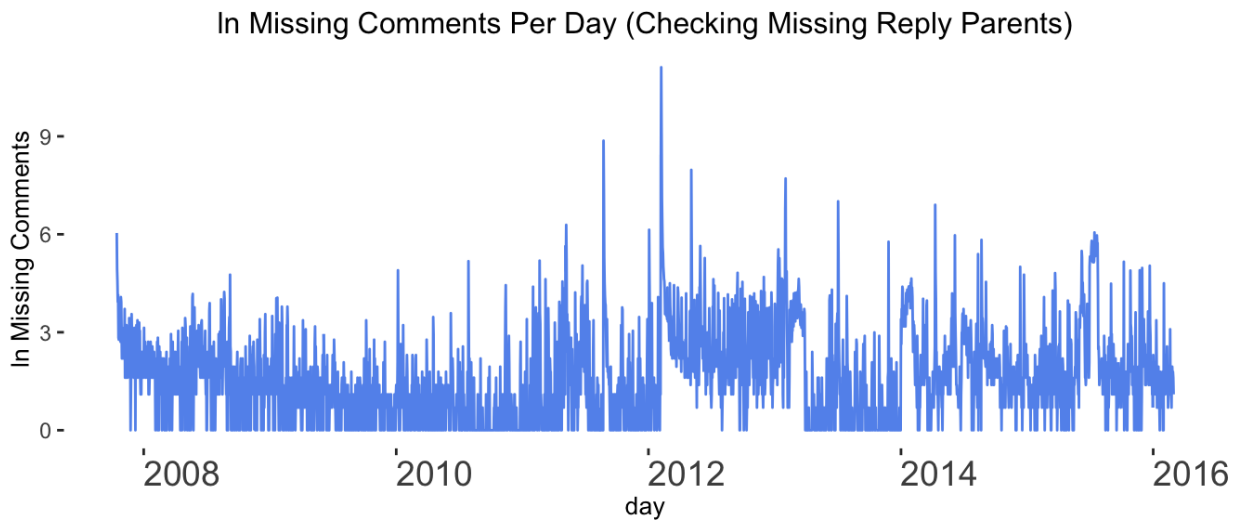

Figure 4: Log plot of missing comments over time - notice extremely large gaps by peaks that reach several orders of magnitude larger for brief moments.

face **minimal risk** because their systems rarely make claims about the population of Reddit users.

User History Analysis

Papers that test hypotheses based on user histories that may have substantial gaps in them. Analyses that are especially sensitive to high-volume users are more likely, on average, to consider users whose histories have gaps. Hessel, Tan, and Lee (2016), for example, observes and compares sums of comment participation between subreddits, and observes the full chain of user history – Hessel, Schofield, Lee, and Mimno (2015) continues with a similar approach. Barbosa, Cosley, Sharma, and Cesar Jr (2016) compares year cohorts of individual-level behavior across all of Reddit, and as has been shown, some years are more affected than others. Additionally, the large amount of potential missing submissions from Reddit's inception may also affect these findings. Depending on how the missing data is missing beyond what has already been shown, some findings may be substantially affected when using this methodology.

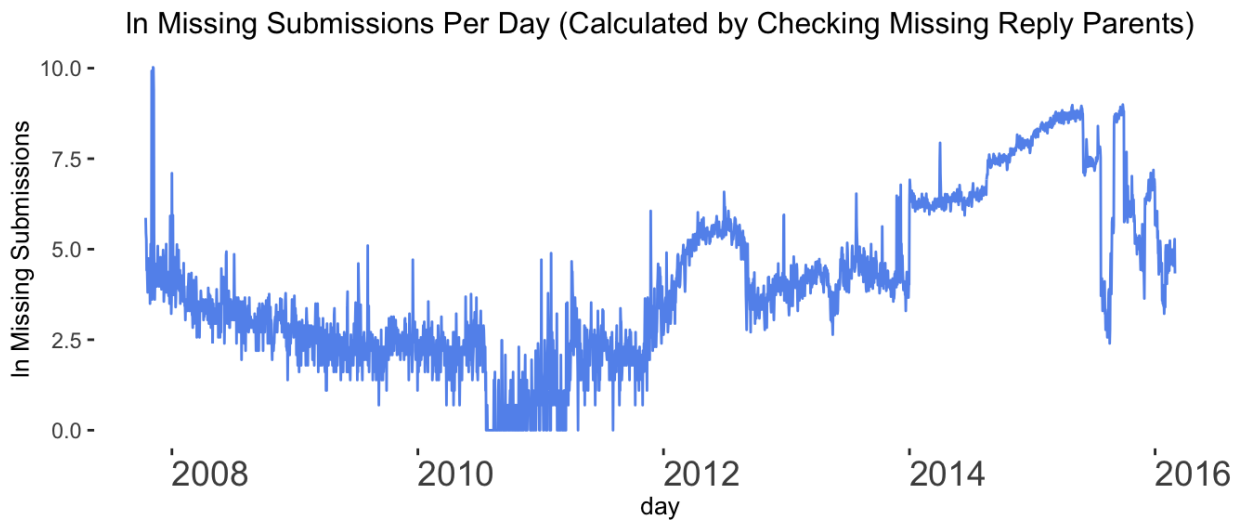

Figure 5: Log plot of missing submissions over time - notice extremely large gaps by peaks that reach several orders of magnitude larger for brief moments.

Network Analysis

Some papers test network hypotheses by constructing interaction networks between users or communities, sometimes over time. Gaps represent a high risk to these papers, since missing submissions may result in unobserved ties in the network. Tan and Lee (2015) observes histories of user accounts participating in different communities, while Fire and Guestrin (2016a) and Fire and Guestrin (2016b) observe network ties over time modeled on user histories. Again, significant blocks of missing data, along with the potentially large amount of missing submissions from Reddit's nascency present potential issues to validity.

Sums of Participation

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 (2016) 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. (2016) observes comment volumes within subreddits. Barthel (n.d.) observes comments about political candidates across Reddit during a

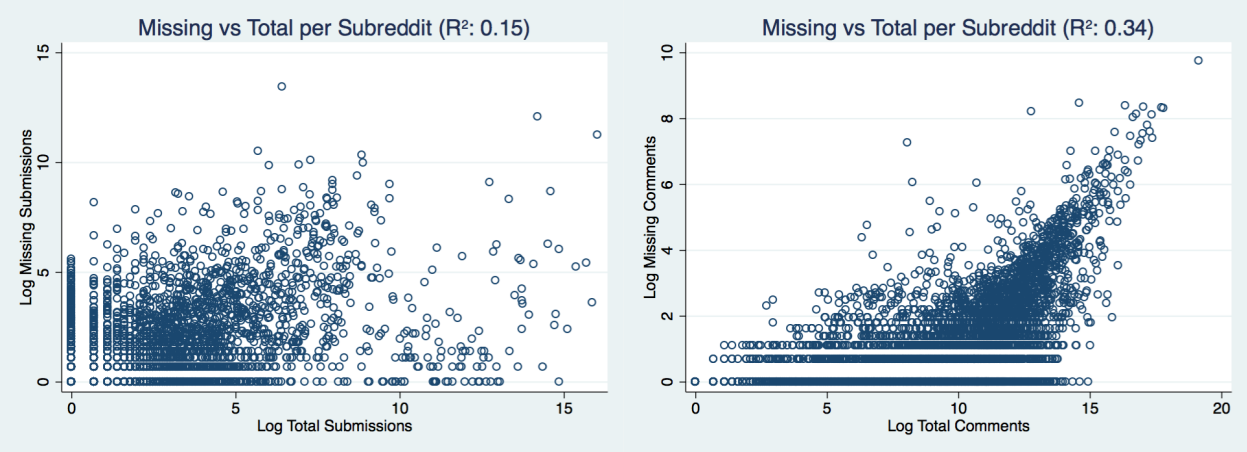

Figure 6: Correlations between the size of subreddits according to total historical counts of comments and submissions versus observed known unknowns by subreddit - comments are mildly correlated, while submissions are not very correlated.

period where many submissions are within the dataset. Barbaresi (2015) analyzes German language text to identify relative commenting rates about places in Germany. While in this instance, where /r/de shows only small amounts of known unknown comments (12) and submissions (450).

Figure 6 shows a deeper introspection – a null hypothesis would state that the number of missing comments and submissions would be tightly correlated to the size of the subreddit. 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 certainly conclude that studies on some subreddits could lead towards very biased results due to higher than random amounts of missing data - in practice, there are 78 subreddits where at least 20% of the comments are missing, and 1,755 subreddits where at least 20% of the submissions are missing – of subreddits that have any missing data according to known unknowns, on average they are missing at least 35% of their submissions.

Content Analysis

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. Saleem, Dillon, Benesch, and Ruths (2016) trains machine learning models on comments from particular subreddits. Though the missing data is non-uniformally distributed over time, and mildly non-uniformally distributed across subreddits, it is unlikely that it favors any form of content over another. Saleem et al. (2016)'s work is specifically focused on communities that have since been quarantined or banned, however, and from a qualitative review of 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 actually has 696,642 comments that reference missing submissions, and only 606 known submissions (on that particular subreddit, however, only 1,100 of 1,585,014 comments were known to be missing). Again, as stated in the previous section, particular subreddits may have larger issues than others. Still, with a large enough corpus of training, it is likely that there is no bias about *which* specific pieces of content are missing within most selection frames, so there is a minimal risk for work employing content analysis given a sufficient scale.

Going Forward

These issues have all been raised in direct with Baumgartner, whom has graciously and quickly made large strides in addressing these issues, and re-processing missing data. By publication time of this paper, it is likely that the issue will have been fully addressed, and the data re-published. In terms of what steps must be taken forward in increasing the integrity of this dataset, little is left and all steps forward are known. The other larger discussion, however, is about the role researchers play in vetting datasets sourced from non-academic providers. All datasets, just by nature of operationalizing a concept for the purpose of research, will have biases present – in the process of designing research, these biases are sought after and addressed. This case highlights a failure of checking for biases – namely that Baumgartner's "completeness" statement should have not been taken as truth.

The stunning scale of the dataset provides for an easy cognitive bias towards assuming completeness, and to be completely clear, Baumgartner's collected dataset is impressive, a generous gift for computational social scientists, and in no way is the result of this case a call to demand for complete rigor placed on non-academic authors – that responsibility lies firmly on academics. In this particular case, the effects of the missing data will likely be marginal, and may not even change results found by all published works using the dataset before the issue is addresssed and the dataset re-published. Even with complete data coverage, there is still a problem with deleted content – a user who deletes even one comment along their posting history induces the exact same problem, and research should analyze the degree to which the magnitude of deleted content on the platform affects research in a similar typological approach. The problem of deleted content, however, is practically speaking impossible to resolve, barring some un-deletion process by Reddit for the sake of academics, which in and of itself raises ethical questions on behalf of the users being studied. The problem of missing data, for what appears to be problems associated with the data collection process itself, is something preventable, and it is the responsibility of academics to check for this problem, and address it by either acknowledging it as yet another potential source of bias, or in the best case, work to address it.

Concluding Remarks

This work began by discussing the impressive dataset collected by Baumgartner, and the impressive array of research already conducted with the dataset. Computational social scientists have been given a gift by a generous benefactor, and it is likely that more useful research will continue as a result of this gift. The dataset, however impressive, contains issues with missing data. While the raw amount of missing data is minimal, depending on methodological approaches, it may actually have significant potential to needlessly negatively affect results of previous research. The missing data affects research differently according

to the typological attributes of the methodological approach, but may also affect research regardless of methodology if the study is, unluckily, conducted on particular subreddits or particular time frames where the missing data issue disproportionately damages the integrity of the dataset. All of this can be resolved by careful introspection of the dataset, and this work has conducted that introspection. The dataset is expected to be resolved of this issue by the original author, and the research community should be grateful that the author has graciously donated more effort to address these concerns. Soon, with a more complete dataset, increasingly useful research will doubtlessly continue.

References

Abdullah, S., & Wu, X. (2011). An epidemic model for news spreading on twitter. In *2011 ieee 23rd international conference on tools with artificial intelligence* (pp. 163–169).

Barbaresi, A. (2015). Collection, description, and visualization of the german reddit corpus. In *2nd workshop on natural language processing for computer-mediated communication* (pp. 7–11).

Barbosa, S., Cosley, D., Sharma, A., & Cesar Jr, R. M. (2016). Averaging gone wrong: Using time-aware analyses to better understand behavior. In *Proceedings of the 25th international conference on world wide web* (pp. 829–841).

Barthel, M. (n.d.). *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/`. (Accessed: 2011-11-07)

Baumgartner, J. (2016). *I have every publicly available reddit comment for research. 1.7 billion comments at 250 gb compressed. any interest in this? : datasets.* `https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/`. (Accessed: 2011-11-07)

Benkler, Y., Roberts, H., Faris, R., Solow-Niederman, A., & Etling, B. (2015). Social mobilization and the networked public sphere: Mapping the sopa-pipa debate. *Political Communication*, *32*(4), 594–624.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295–298.

Fire, M., & Guestrin, C. (2016a). Analyzing complex network user arrival patterns and their effect on network topologies. *arXiv preprint arXiv:1603.07445*.

Fire, M., & Guestrin, C. (2016b). Time is of the essence: Analyzing the effect of vertex-joining time on complex network evolution. *arXiv preprint arXiv:1603.07445*.

Hessel, J., Schofield, A., Lee, L., & Mimno, D. (2015). What do democrats do in their spare time? latent interest detection in multi-community networks. *arXiv preprint arXiv:1511.03371*.

Hessel, J., Tan, C., & Lee, L. (2016). Science, askscience, and badscience: On the coexistence of highly related communities. In *Tenth international aaai conference on web and social media.*

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., et al. (2011). The arab spring—the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International journal of communication*, *5*, 31.

Massanari, A. (2015). # gamergate and the fappening: How reddits algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 1461444815608807.

Matias, J. N. (2016). 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* (pp. 1138–1151).

Newell, E., Jurgens, D., Saleem, H. M., Vala, H., Sassine, J., Armstrong, C., & Ruths, D.

(2016). 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.*

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Lrec* (Vol. 10, pp. 1320–1326).

Saleem, H., Dillon, K., Benesch, S., & Ruths, D. (2016). 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).*

Tan, C., & Lee, L. (2015). All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of the 24th international conference on world wide web* (pp. 1056–1066).

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *arXiv preprint arXiv:1403.7400*.