Science Forums Data Analysis Report

Cathy Su

13/11/2019

Changes made since V1

- introduce the variables
- remove raw R
- formatting: remove R code and text running into margins
- add more explanations to captions of figures; also corrected figure numbers that didn't show.
- address outlier topics comment in Figure 3.
- expand executive summary to 1 page and expanded the introduction, adding citation

Executive Summary

Online forums have become ubiquitous in the internet age. They provide a place where people from many different backgrounds can interact at the same level by sharing their thoughts across topics of interest. The diversity of participation is a strength that can make these forums useful and interesting for all future readers, however, at the same time the anonymity and openness can allow strange or unpleasant discussions as well. Moderators at online forums are always interested in growing participation while keeping a high quality discussion. Therefore, at times it becomes necessary to delete or close topics that might become problematic where forum members get into really heated discussions or the posts become inappropriate. The purpose of this study was to provide insights into how to moderate online forums by assessing which variables influence which topics need to be closed.

Previous work on this topic has suggested moderators play a crucial role in growing online forums in a healthy way. For instance, in their study of six online health forums, Huh et al. found that moderators influenced the quality of the online forums by encouraging a respectful discussion, providing expertise, and reinforcing participation etiquette and forum rules (Huh et al. 2013). As part of their role, moderators need to monitor the forum to decide which topics should be closed. The type of post authors, category of topic, and number of deleted posts are important factors in which discussions will need to be shut down.

To explore the relationship between these covariates and the discussion status of a topic, we are using a processed dataset from Nifty Datasets repository taken from a highly active online discussion forum, ScienceForums.Net (SFN). Although it is a forum that focuses on science content there are also subforums relating to other diverse topics such as politics and religion. The forum is also open such that participants of all levels of expertise are represented. Due to high traffic and success, SFN employs a handful of staff who perform moderating functions such as shutting down discussions and banning authors.

Ten thousand discussion topics were randomly sampled from SFN for this study. The forum has operated since 2002 and contains more than 800 000 posts. 13 features were extracted from the data including the variables of interest such as author diversity and the status (closed versus open). Among the subforums, there are topics ranging across all of science including "Speculations" where especially controversial discussions will be moved. After removal of outliers, 9021 topics passed the inclusion criteria.

Originally 12 predictors were available and for our analysis we focused upon three qualitative and nine continuous covariates to study whether discussions will need to be closed. We built generalized linear models with a log link to better model the right skewed variables of interest. To account for the time since they were posted, we used a quasipoisson model of the view counts. We compared models with and without these predictors using the chi squared test and by examining the distribution of model residuals.

We found that the relationship between views and posts is positive and depends significantly upon the subforum by chi squared test ($p < 0.001$). There are outlier topics in subforums such as 'Medicine' and 'Speculations' which have disproportionately more views. Discussions with more authors tended to be closed

more often ($p < 0.05$). Discussions started from experienced authors also tend to be more successful ($p < 0.05$).

Additionally, we selected predictors using best subset regression and found that subforum category, number of unique authors, proportion of deleted posts and the seniority of the author who starts the topic are all statistically significant predictors of the status of a topic and/or its success in terms of views and number of posts. This suggests that these are the most important items for moderators to monitor when they consider whether a discussion can be allowed to continue.

Introduction

Online forums have grown in popularity over the past year as a place for all those with internet access to share thoughts. Participation in an online forum can be an enriching experience, but at the same time, the low barriers to entry can also allow harmful and unpleasant interactions. Therefore moderators at online forums are always interested in growing the participation while keeping a high quality discussion, but it's unclear what they should pay attention to when making this decision. Here we examine which topics on an online science forum, ScienceForums.Net, are at greatest risk of needing to be closed. This dataset represents a relevant scenario where moderators needed to make decisions on closing topics based upon our covariates of interest including author profile, author diversity, and subject matter of posts. Based on our literature review, we hypothesized that author diversity, author experience and subforum category are most important to determine number of posts, number of views and closed or open status of topics.

We processed the data from its original format as taken from teh Nifty Datasets database which contains 12 predictors and 9021 topics of interest spread across 13 subforums such as "Speculations" and "Medicine". Accordingly within this ScienceForums data, we have three major types of information: classification of the topic, statistics on the topics within the time period, and statistics about the participating authors. These are all informative characteristics for studying whether the topic will be closed.

Due to the nature of the count data and binary response variables we used generalized linear models to find the important explanatory variables behind the status of topics. We found important covariates by comparison of models with chi squared test and checked their predictive ability by using best subsets regression. We found the most important variables for building a model that is predictive of the number of posts include the date the topic was started, registration days and banned status of the author, category of the subforum, number of deleted posts and post rate.

Methods

Removal of outliers and missing data

Related to Figure 2, we removed the topics which are pinned since there seem to be very few cases so we may not be able to properly model what happened there. Looking at Figure 3, the only missing values in the data came from topics without categorization, which were also removed because we want to know the impact of the subforum type. Lastly we decided also on the basis of Figure 3 to remove topics with number of posts greater than 250 which seem to be extreme outliers.

Calculation of additional variables

For the substantive questions we calculated some new variables as follows:

- We converted the 'startdate' into POSIXct format which allows us to compare time elapsed precisely between dates. However to keep these numbers in a reasonable scale for the models, we subtracted 10^9 and divided the resulting number by the maximum startdate resulting in a (0,1] scale. In this scale, the entire approximattely 12 year period between the first and last topic is normalized to 1.
- the 'proportion_deleted' is the number of deleted divided by total posts in the topic which is necessary for Q1.
- The 'post_rate' is the number of posts divided by the 'duration' of the topic. If 'duration' is zero, the 'post_rate' was also assigned zero.

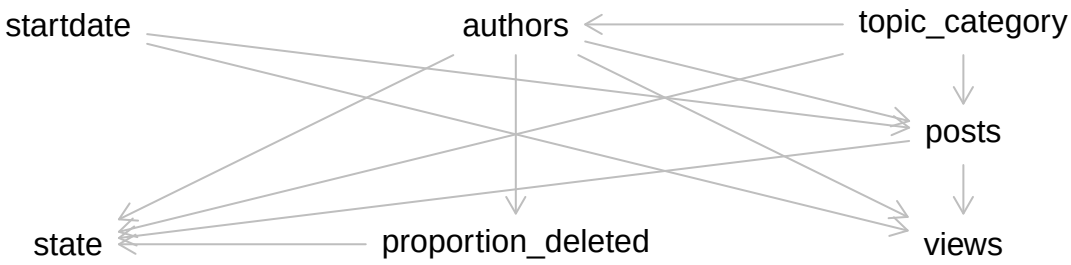

Figure 1: Causal diagram illustrates hypothesized relationships of experimental variables involved in relationship between proportion of deleted posts and topic status.

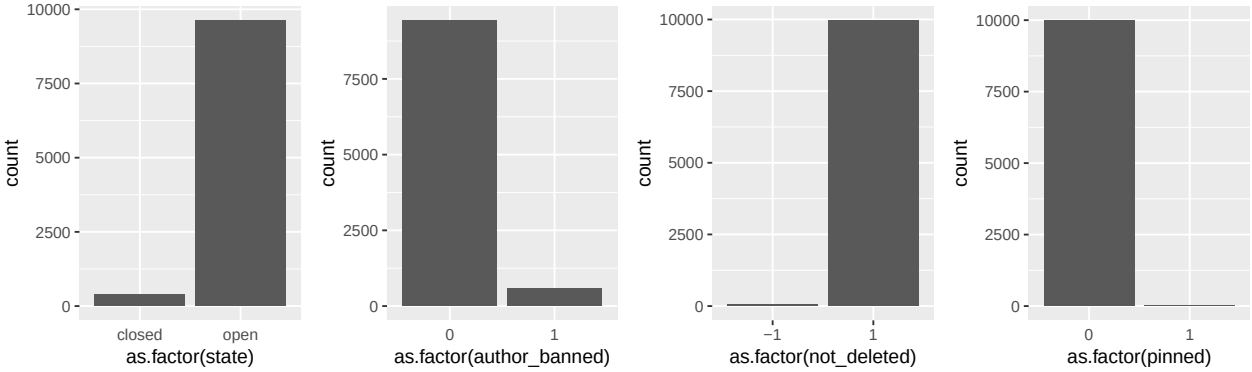

Figure 2: Distribution of binary categorical variables.

- 'posts_per_author' is number of posts divided by the number of distinct 'authors'.

Exploratory Data Analysis

We first performed EDA to study the relationship between dataset variables which fell within three major types of information: classification of the topic, statistics on the topics within the time period, and statistics about the participating authors. Each topic comes with an id (tid), subject category (e.g. "Miscellaneous sciences") and a state (open or closed). There are numeric variables describing the properties of each of the topics such as the number of posts, views, authors, and deleted posts. Lastly, some variables pertain to statistics about the authors who post on the topics such as the number of days the topic starter was registered (author_exp) and whether they were banned (author_banned). The categorical variables are visualized in Figure 2 and the quantitative variables are visualized in Figure 4.

Causal diagram

We may hypothesize that topics will need to be closed mainly due to offensive posts, or due to controversial discussion. Based on this, author diversity ('authors') could be important to affect the relationship between proportion of deleted posts, length of the discussion and topic status (open or closed) as in Figure 1.

Univariate variable distributions

Figure 2 shows the distribution of topics which fall into each category for the binary variables. This shows us that almost no topics are deleted from view or pinned. However a small portion of topics (<10%) are closed, or started by a banned author.

Additionally, although topic type is an important variable of interest, we found that there were many topic ids which were missing a categorization (about 10%, see the top boxplot of Figure 3) which we removed. Additionally the number of posts was very right skewed, and we trimmed the few topics with posts in excess

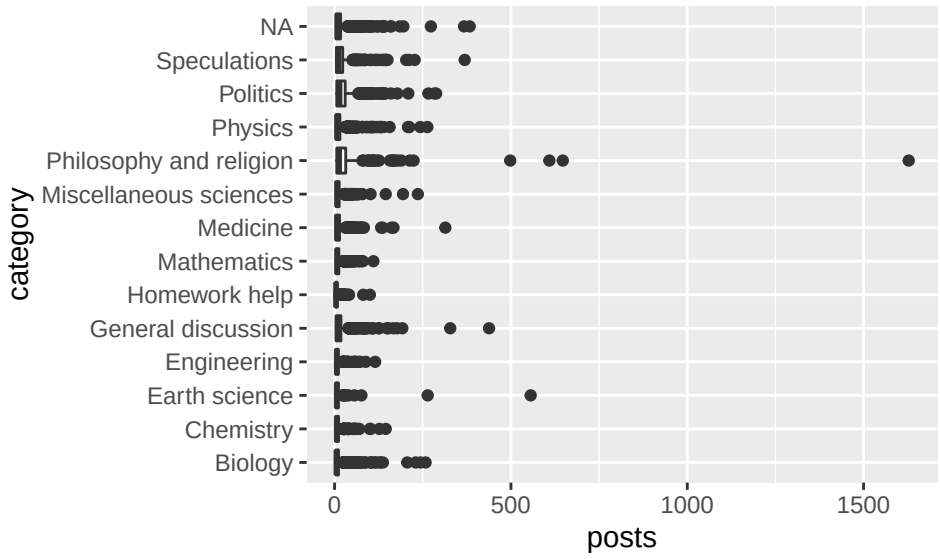

Figure 3: Boxplots giving breakdown of posts by subforum, showing many outlier topics which have an extreme number of posts. It seems Based on this raw data distribution, we trimmed the topics with an outlier number of posts.

of 500 since these seem were extreme outliers based on Figure 3). The distribution of views was similarly right skewed to the distribution of posts. This suggests that it may be better to use a quasipoisson model than a poisson model for these counts. Indeed, when we tested for overdispersion later we found that the views were both significantly overdispersed for a poisson model (Q1, Table 2 Model 1). To overcome this, we have use quasibinomial and quasipoisson models. However, as shown later these models are not perfect as there are still many outliers.

In Figure 4 we also see a relationship between duration, author_exp and startdate. If time increases by one unit in startdate, the number of views or posts may increase which suggests that we may want to use it as an offset. However the limitation of this startdate variable is that we picked an arbitrary offset for the time (see methods) so that we could apply a log transformation when we can use it as an offset for a poisson glm. As well, it's not clear that we need to know the time to very high precision as we do here.

Q1. Relationship between views and posts

Since views are a form of counts which are positive and do not have a defined maximum, to determine the relationship between views and posts, we start by fitting the glm with the poisson family, log link and log('starttime') as offset (Model 1 in Table 2). Based on our causal diagram, we chose to control for potential common cause 'authors' as well.We tested for overdispersion of this model with the package AER's dispersiontest we found that the views were both significantly overdispersed for a poisson model ($p < 0.001$, results not shown). Therefore we switched to using the quasipoisson which found the same significance levels for coeffcients with lower residual deviance (Model 2 in Table 2).

However the residuals of this basic model shown in Figure 5 suggest that the data have multiple outliers. The diagnostic plots are meant for linear models, which is why we see that for example the residuals are not highly normally distributed in the normal QQ plot. However we can still use the plot to see that there are a number of outliers very visible in the residuals vs. leverage plot. Rows 5483, 5592, 5997, and 6660 are some of the rows highlighted and these are topics belonging to "Miscellaneous sciences", "Speculations", and "Medicine".

We then added a term for subforum. We compared the models of these with and without the additional variable 'category' (which relates to their subforum) by chi squared test, and the results are in Table 5. It seems that the relationship with views and posts varies strongly by subforum since the chi squared test suggests the latter model has a significantly better fit ($p < 0.001$).

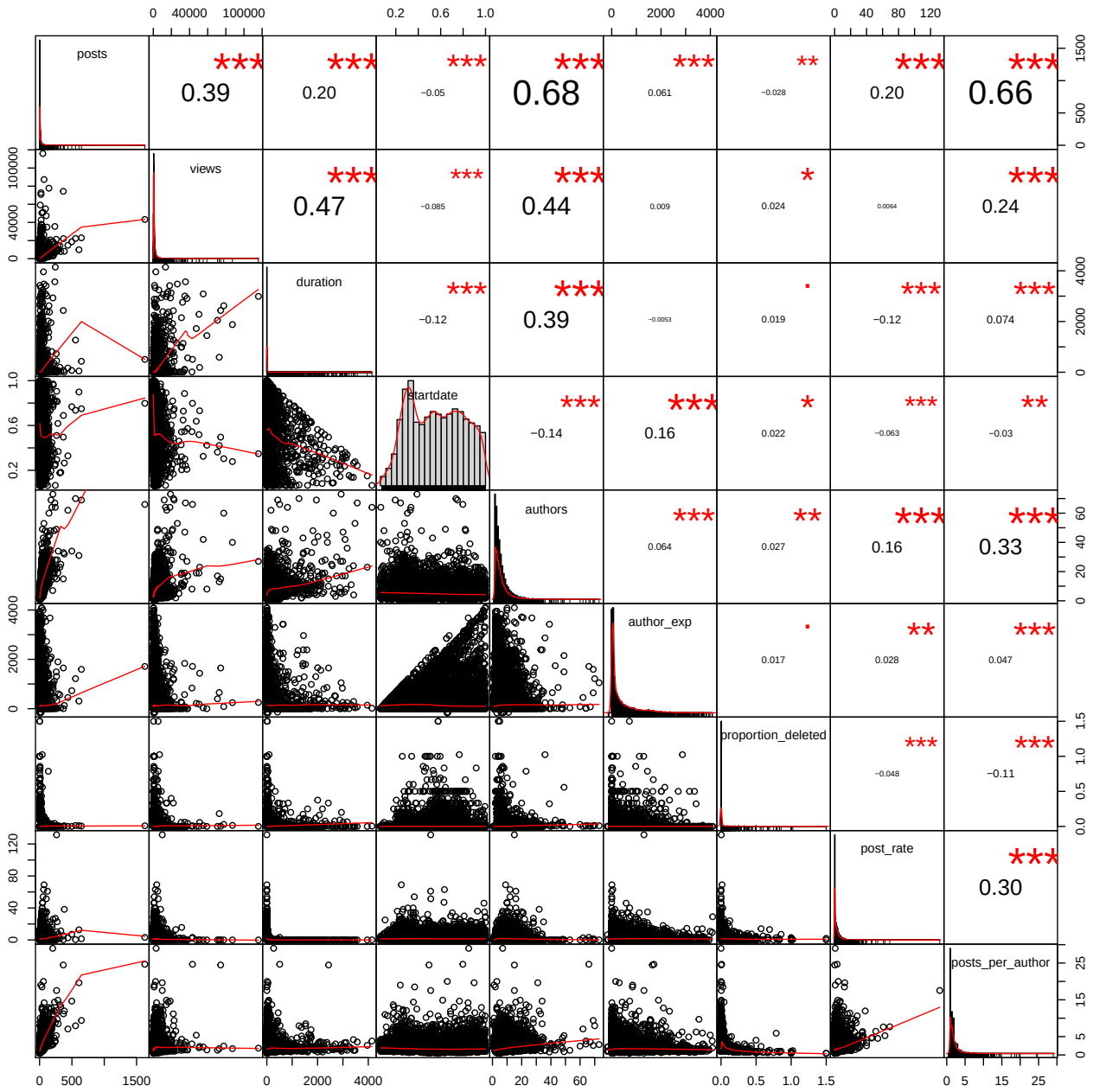

Figure 4: Pairwise correlations of important variables including their Pearson correlation coefficient. Significant correlations are marked by the corresponding number of red astericks. We can see from the univariate distributions (graphs on the diagonal) that with the exception of 'year_started', these variables are mostly very right skewed and positive count or rate data.

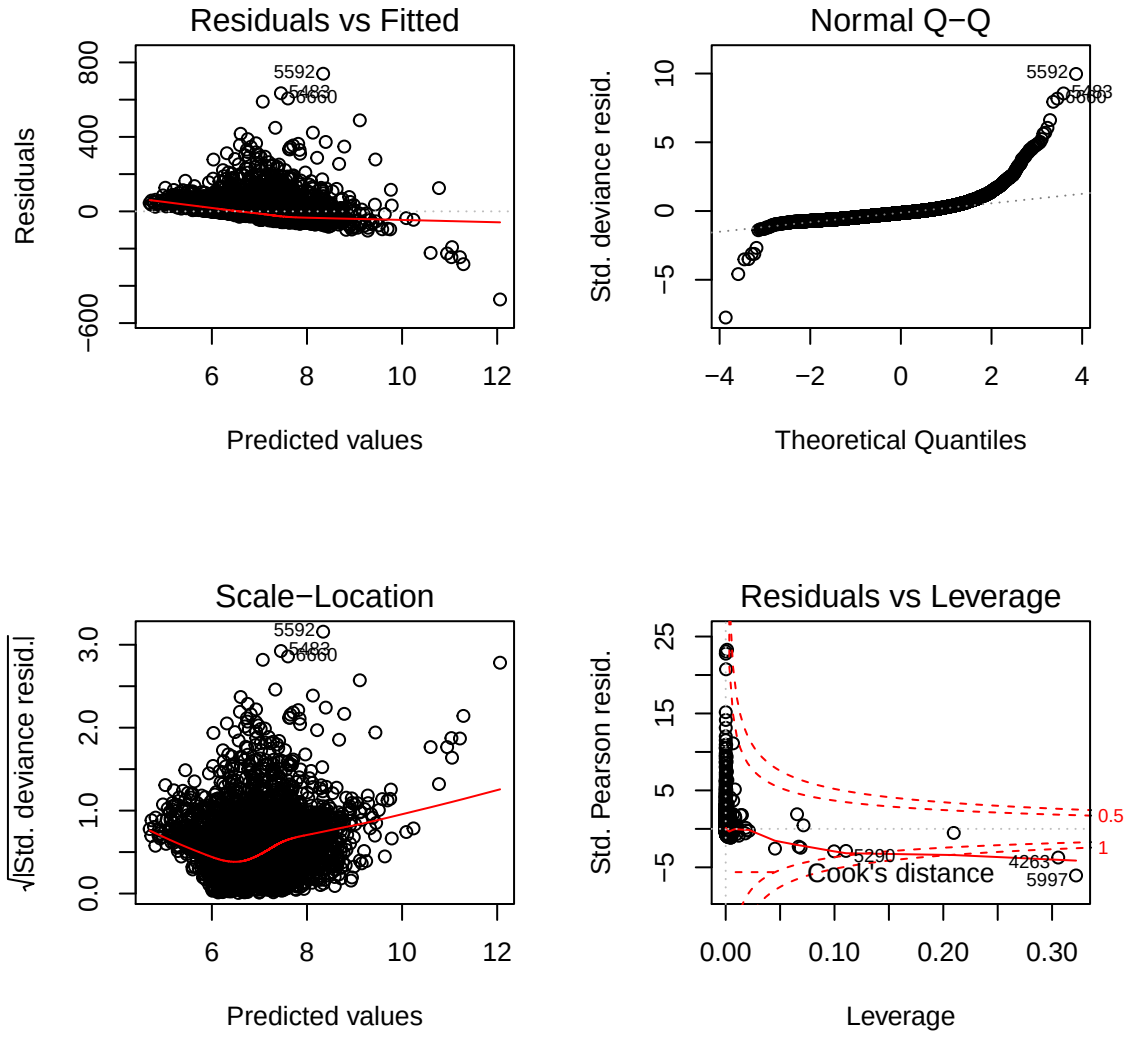

Figure 5: Diagnostic plots for the fit of model 'views ~ posts + authors + offset(log(startdate)'. Multiple outliers are apparent. These outliers occurred in topics belonging to Miscellaneous sciences, Speculations, and Medicine.

Table 1: Chi squared test models with and without controlling for subforum

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
9029	18189781	NA	NA	NA
9017	16742039	12	1447742	0

Table 2: Terms included in models of views vs. posts

	Dependent variable:		
	views		
	(1)	(2)	(3)
Constant	7.318^{***} (0.0004)	7.318^{***} (0.026)	7.317^{***} (0.060)
posts	0.001^{***} (0.00001)	0.001^{*} (0.001)	0.002^{***} (0.001)
categoryChemistry			0.582^{***} (0.076)
categoryEarth science			-0.436^{*} (0.226)
categoryEngineering			-0.066 (0.129)
categoryGeneral discussion			-0.602^{***} (0.091)
categoryHomework help			0.021 (0.092)
categoryMathematics			0.182^{*} (0.097)
categoryMedicine			0.485^{***} (0.081)
categoryMiscellaneous sciences			0.052 (0.091)
categoryPhilosophy and religion			-0.590^{***} (0.110)
categoryPhysics			-0.136^{*} (0.070)
categoryPolitics			-0.237^{**} (0.098)
categorySpeculations			-0.271^{***} (0.088)
authors	0.075^{***} (0.00004)	0.075^{***} (0.003)	0.076^{***} (0.002)
Observations	9,032	9,032	9,032
Log Likelihood	$-9,133,683.000$		
Akaike Inf. Crit.	18,267,373.000		

Q2. Diverse discussions closed or deleted

To check whether author diversity affects topic state, we modelled the responses 'state' and 'not_deleted' against the predictor 'authors'. Since the response is binary this time we used a quasibinomial family glm with logit link, with 'starttime' as offset. Based on our causal diagram we also controlled for subforum and number of posts. These were decent models of the topics that were open and not deleted according to the residual vs leverage plots but did not represent the closed and deleted topics very well (left hand side of Figure 6). This makes sense however as we have a binary variable where one outcome is significantly rarer as we saw in Figure 2.

However with authors added as covariate, the number of outliers is somewhat less (RHS of Figure 6). Furthermore the result of chi squared test is given in Table 3-4. This suggests that the model with authors better represents the data and discussions involving more authors are significantly more likely to be closed ($p < 0.05$) or deleted ($p < 0.001$). The coefficients on the covariate 'authors' are 0.042451 and 0.2245 in each case respectively. However only the coefficient of 0.042451 for the model of being closed is significant ($p < 0.01$)

The reference category of this model is Biology. The results can be interpreted to mean that the log odds of being closed increases by exp(0.042451) or about 4 percent per period of about 12 years (see methods for explanation of this unit) for every unique author, relative to a topic with no authors of the category Biology while controlling for the number of posts and using the start date as an offset.

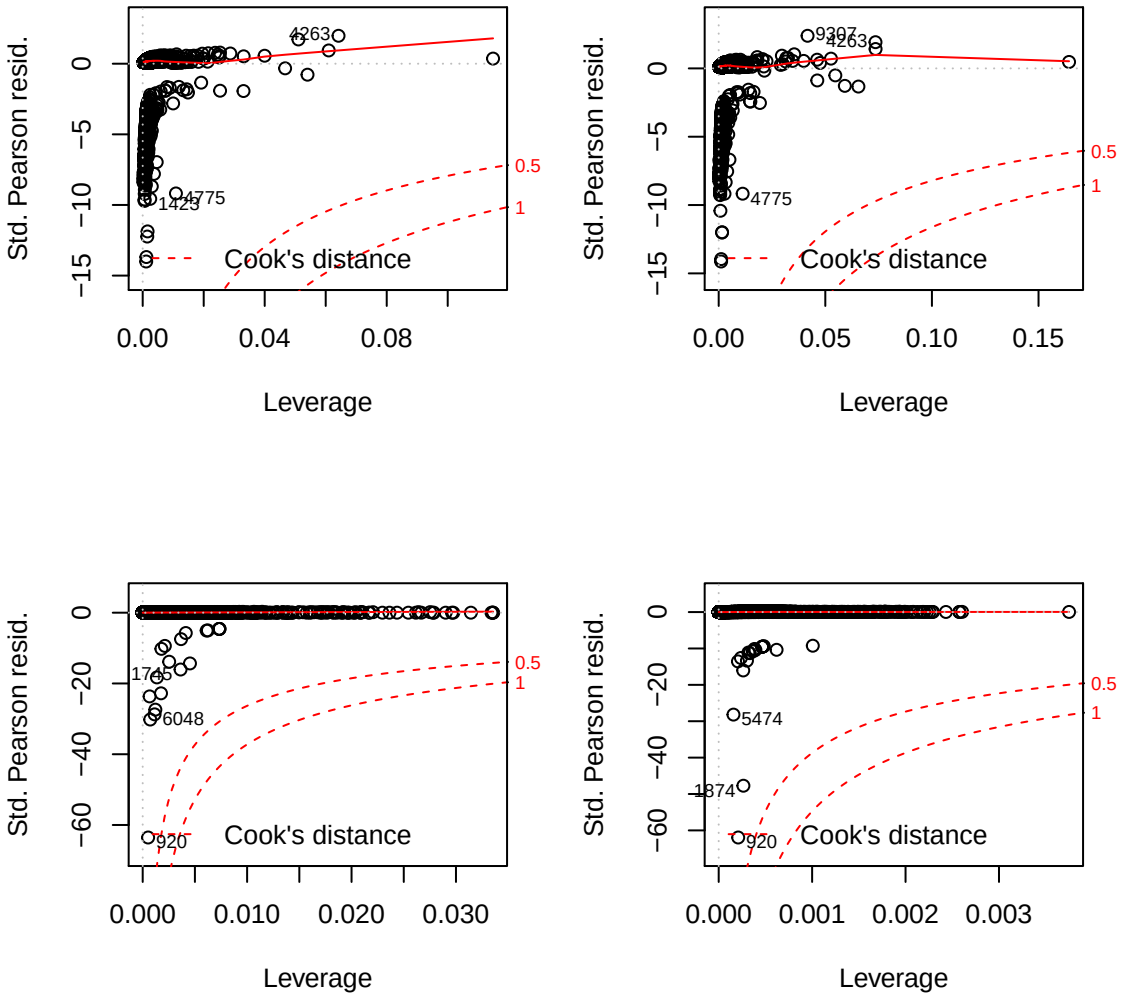

Figure 6: Residual plots of each of the models of 'state' (top row) and 'not_deleted' (bottom row). Each graph on teh right hand side includes the covariate 'authors' and those on teh left hand side do not.

Table 3: Comparison of model of deleted topics with and without controlling for authors

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
9018	240.594	NA	NA	NA
9029	273.393	-11	-32.799	0.002

Table 5: Coefficients of generalized linear model given in Q4

	x
(Intercept)	3.0491771
proportion_deleted	-0.8638655
author_exp	-0.0001327
categoryChemistry	-0.1446834
categoryEarth science	-0.1937571
categoryEngineering	-0.2149314
categoryGeneral discussion	0.5951648
categoryHomework help	-0.7321052
categoryMathematics	-0.0063990
categoryMedicine	0.2176237
categoryMiscellaneous sciences	0.1572621
categoryPhilosophy and religion	0.6706452
categoryPhysics	0.1222490
categoryPolitics	1.0395089
categorySpeculations	0.4325405

Q3 Do topics with deleted posts tend to get closed more often?

To check whether deleted posts tend to get closed more often, we checked if the full quasibinomial model of the response 'state' from Q2 **views ~ posts + category + authors + offset(log(startdate))**, could be improved by adding the predictor 'proportion_deleted' by chi squared test. The inclusion of proportion deleted does not significantly improve the model ($p > 0.05$) as shown in the Table 5. This suggests proportion of deleted posts is not a statistically significant predictor of topic state.

Table 4: Comparison of model of state from Q3 with and without controlling for proportion of deleted posts

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
9017	2494.462	NA	NA	NA
9016	2492.884	1	1.579	0.229

Q4. Are members who have been registered for longer before starting the topic more successful at starting active discussions?

To check whether long registered members may be more successful, we modelled the response 'posts' as proxy for activity of the topic against the predictor 'author_exp'. Since the response is a positive count we use quasipoisson with log link. From our causal diagram we decided to control for subforum and proportion_deleted with log(startdate) as offset. The coefficients of this model are given in Table 6 and the coefficient of 'author_exp' is -1.399e-04 which is significantly different from zero ($p < 0.01$).

The reference category of this model is also Biology. The results can be interpreted to mean that the count of posts increases by exp(0.042451) or about 1 post per period of about 12 years for every additional day that the topic starting author is active, relative to the post count in topics in Biology started by authors that are brand new.

Table 6: Coefficients of generalized linear model of binomial family picked by best subsets regression with AIC

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.8639712	0.3169385	15.3467369	0.0000000
startdate	-1.3445580	0.2805956	-4.7918001	0.0000017
author_exp	0.0007202	0.0001615	4.4586500	0.0000082
author_banned1	-1.0546636	0.1720972	-6.1283033	0.0000000
categoryChemistry	-0.6635955	0.3403092	-1.9499783	0.0511787
categoryEarth science	0.2027768	1.0438844	0.1942521	0.8459785
categoryEngineering	0.3600540	0.6392734	0.5632239	0.5732824
categoryGeneral discussion	-1.2669938	0.3112446	-4.0707339	0.0000469
categoryHomework help	1.7696428	0.7557996	2.3414178	0.0192107
categoryMathematics	-0.7574076	0.3624742	-2.0895489	0.0366583
categoryMedicine	-0.6173556	0.3710617	-1.6637547	0.0961615
categoryMiscellaneous sciences	-0.7756918	0.3409249	-2.2752571	0.0228905
categoryPhilosophy and religion	-1.4379328	0.3603807	-3.9900382	0.0000661
categoryPhysics	-0.1160288	0.3100434	-0.3742342	0.7082301
categoryPolitics	-0.5023689	0.4107707	-1.2229911	0.2213331
categorySpeculations	-1.3969316	0.3015193	-4.6329753	0.0000036
post_rate	-0.0447494	0.0091254	-4.9038431	0.0000009

Q5. Predicting whether a given topic will be closed

To build a classification model to predict whether a given topic will be closed, we chose 'state' as the response variable and then selected the potential predictors based upon whether they would be available while the topic is active. This means we could pick from only the following variables: the date the topic was started, registration days and banned status of the author, category of the subforum, number of deleted posts and post rate.

First we picked the best glm model using the training set, we performed best subsets regression using exhaustive search and AIC with the 'bestglm' package. Since the response is a binary variable, we chose models from the binomial family with logit link. We divided our data randomly into 10-fold and used 1 fold as the test set. Due to the categorical variables, we used AIC instead of cross validation to pick the best model since if we used cross validation some subsets will not contain all the categories.

We then performed model fitting of a glm binomial model to estimate the effect size and significance from our filtered list of covariates. The coefficients of the best model is shown in Table 6. We found that all of the covariates that we found were significant ($p < 0.05$). Specifically within the category of subforum, it significantly mattered relative to the baseline Biology topic ($p < 0.05$) if the forum topic fell within the areas of General Discussion, Homework Help, Math, Miscellaneous Sciences, Philosophy and religion, or Speculations.

Conclusion

We have examined teh relationship between many covariates of interest for the ScienceForums.Net dataset in predicting which topics will need to be closed. The data consists of counts and binary factors. Based upon our analysis we found that the relationship between views and posts is positive and depends significantly upon the subforum. There are outlier topics in categories such as 'Medicine' which have disproportionately more views. Discussions with more authors also significantly tend to be closed, but those with a higher proportion of deleted posts may not be. Lastly, discussions started from experienced authors tend to be more successful.

However, there are some caveats to this analysis. We found in our EDA that the counts are overdispersed. To overcome this, we have use quasibinomial and quasipoisson models. However these models were still a bit difficult to assess using traditional residual plots and we found outliers indicating they are not a perfect

fit. We also chose to discard some variables based upon our EDA and our reasoning about the relationship between variables collected in the study but we do not have expert knowledge in this area. Depending on the outliers removed, the conclusions of the models we have built could change.

Bibliography

Huh, Jina, David W. McDonald, Andrea Hartzler, and Wanda Pratt. 2013. "Patient moderator interaction in online health communities." *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium* 2013:627–36.