CMPUT 660 Assignment 1

Logan Gilmour

October 5, 2013

1 Schema

I used the MSR Mining Challenge 2013 StackOverflow Posgresql dump [1].

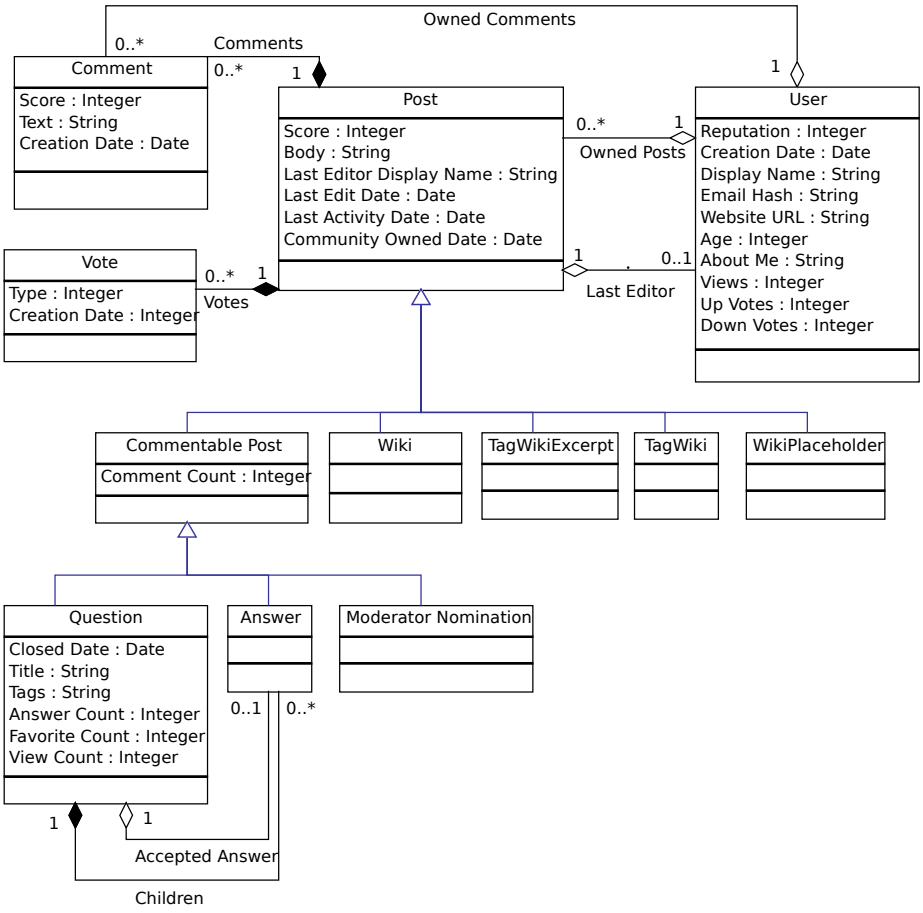
Though there are only four tables in the schema given, both the Votes and Posts tables contain a 'Type Id' field.

By counting the non-null values for each column for each Post type, I was able to infer which fields were only present in specific types of posts. Specifically, the 'Question' type has several fields not found in any other type. There are also only three types of posts with comments; the 'Commentable Posts' class is a synthetic construct to indicate that.

Votes also have distinct types (see table), however, they all have the same fields, so I have left them as one class with the type field intact. I have listed the separate vote types and their individual counts in the next section.

In order to make the schema more presentable in the UML diagram given here, I replaced all underscores with spaces. All tables also contain Id fields, which are not shown in the UML diagram.

Field that refer to Ids from different tables are turned into connecting arrows, and inverted to make more sense as a composite structure. For example, the 'post id' field in the 'votes' table became a 'votes' composition on the 'posts' table.


Owned Comments
0..*
Comments
1
1
Comment
Score : Integer
Text : String
Creation Date : Date
Post
Score : Integer
Body : String
Last Editor Display Name : String
Last Edit Date : Date
Last Activity Date : Date
Community Owned Date : Date
0..*
1
Owned Posts
User
Reputation : Integer
Creation Date : Date
Display Name : String
Email Hash : String
Website URL : String
Age : Integer
About Me : String
Views : Integer
Up Votes : Integer
Down Votes : Integer
Vote
Type : Integer
Creation Date : Integer
0..*
1
Votes
1
0..1
Last Editor
Commentable Post
Comment Count : Integer
Wiki
TagWikiExcerpt
TagWiki
WikiPlaceholder
Question
Closed Date : Date
Title : String
Tags : String
Answer Count : Integer
Favorite Count : Integer
View Count : Integer
Answer
Moderator Nomination
0..1
0..*
1
1
Accepted Answer
Children

2 Size Metrics

2.1 Counts

Posts	
Type	**Count**
Question	3453742
Answer	6858133
Wiki	167
TagWikiExcerpt	13095
TagWiki	13095
ModeratorNomination	138
WikiPlaceholder	1
Total	10338371

Users
1295620

Comments
13252467

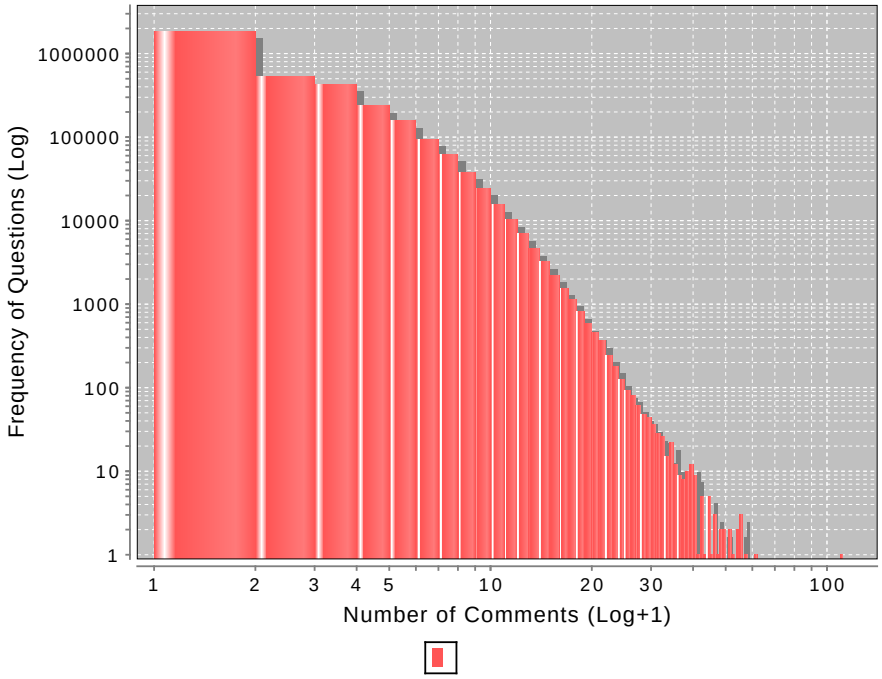
Votes	
Type	**Count**
AcceptedByOriginator	2152420
UpMod	19948316
DownMod	1508097
Offensive	513
Favorite	1877602
Close	475359
Reopen	3327
BountyStart	38358
BountyClose	37959
Deletion	994180
Undeletion	65192
Spam	1946
ModeratorReview	194570
ApproveEditSuggestion	273861
Total	27571700

2.2 Summary Statistics

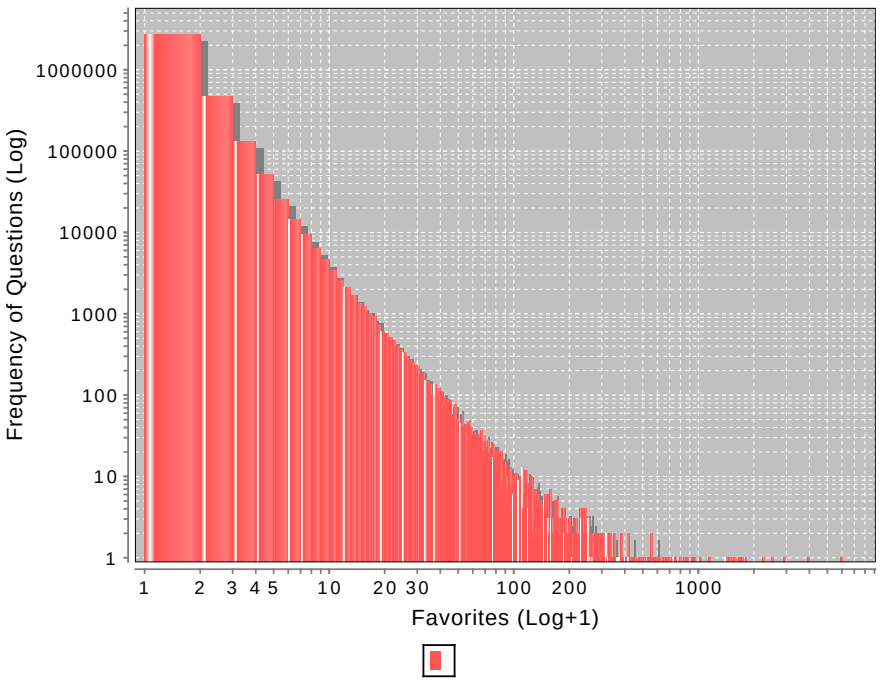
I have focused on the Question and Answer types of posts for this section, as they comprise actual question and answer data of StackOverflow. The moderator election process seems unrelated to the kinds of questions I would ask. Also, my charts are likely terribly labeled.

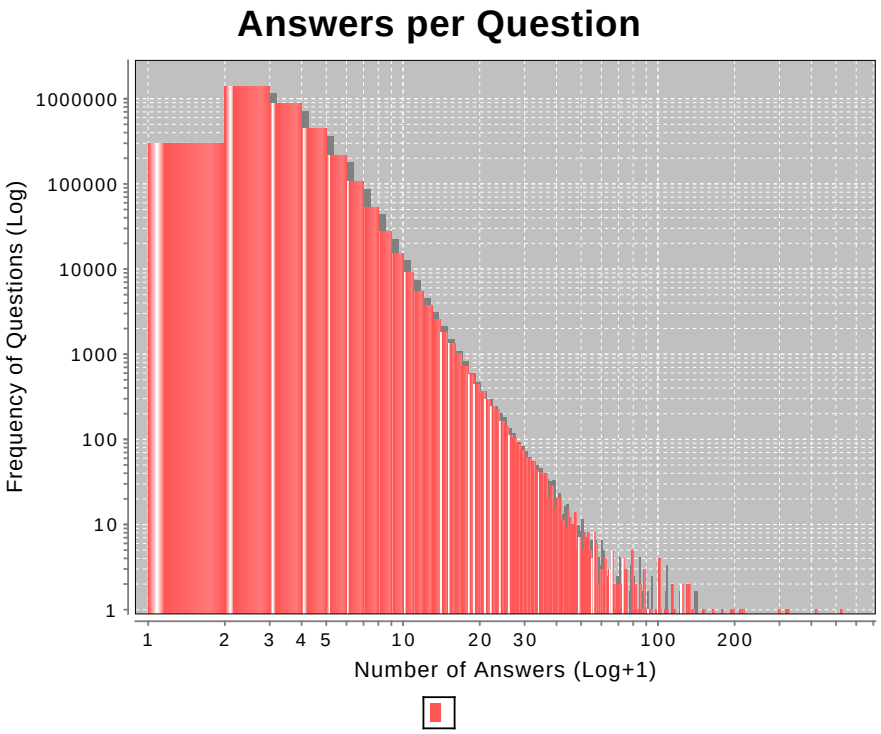
Questions					
	Score	**Comment Count**	**Favorite Count**	**Answer Count**	**View Count**
min	-132	0	0	0	1
max	2499	109	5894	519	1051784
median	1	0	0	2	205
mean	1.4952	1.3276	0.5178	1.986	730.1
var	45.1488	4.5475	41.1802	3.550	10825646.4
std.dev	6.7193	2.1325	6.4172	1.884	3290.2
skewness	81.4	2.9	401.1	20.6	43.6
kurtosis	15366	21	279068	3299	5516

Comments per Question

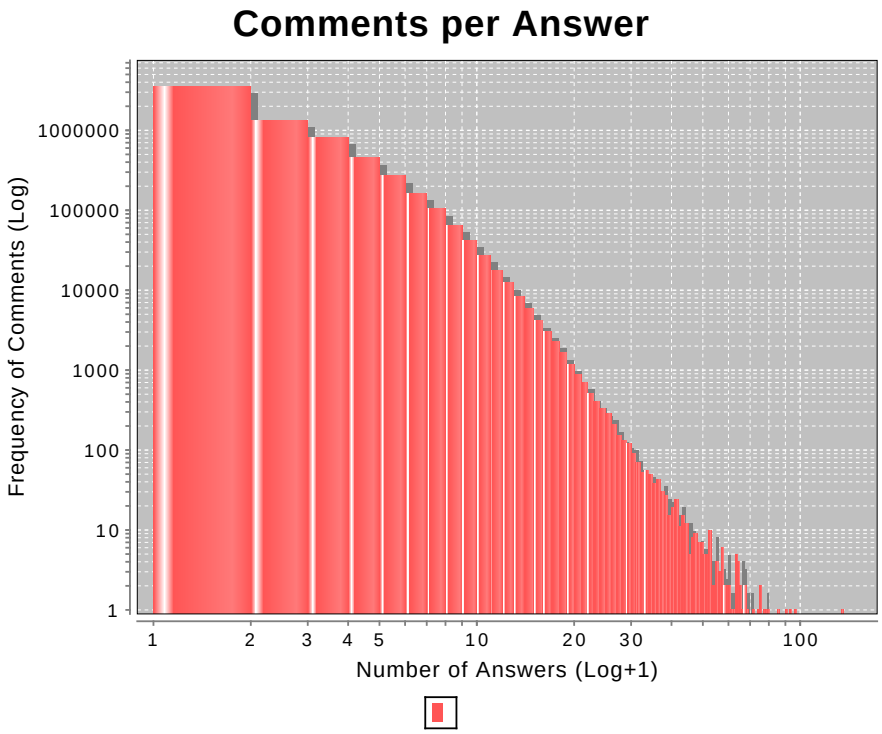


Favorites per Question



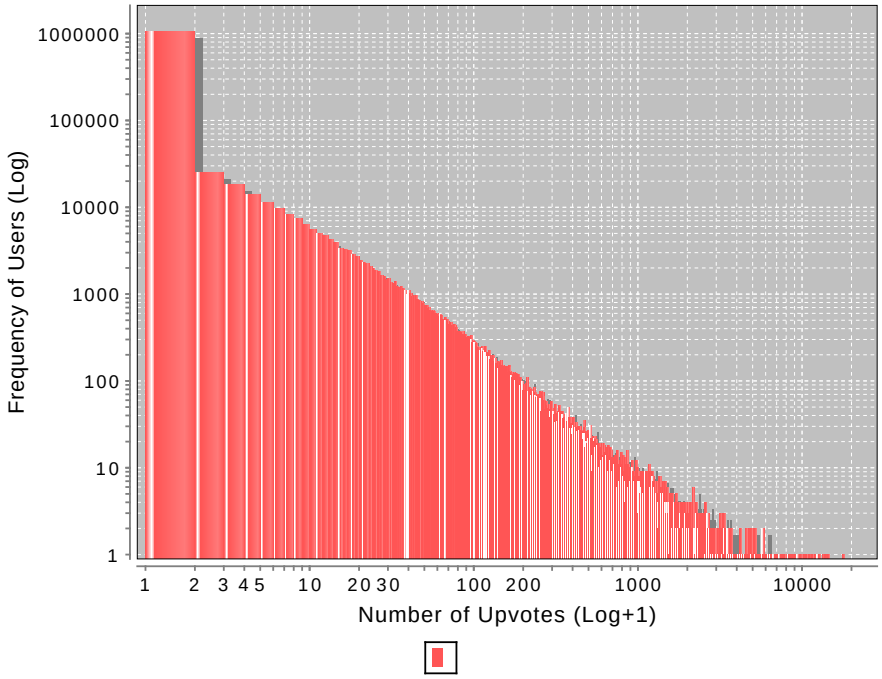


Answers		
	Score	**Comment Count**
min	-59	0
max	4432	133
median	1	0
mean	1.8672	1.26376
var	53.1593	4.25170
std.dev	7.2910	2.06196
skewness	117.8	3.5
kurtosis	37770	34

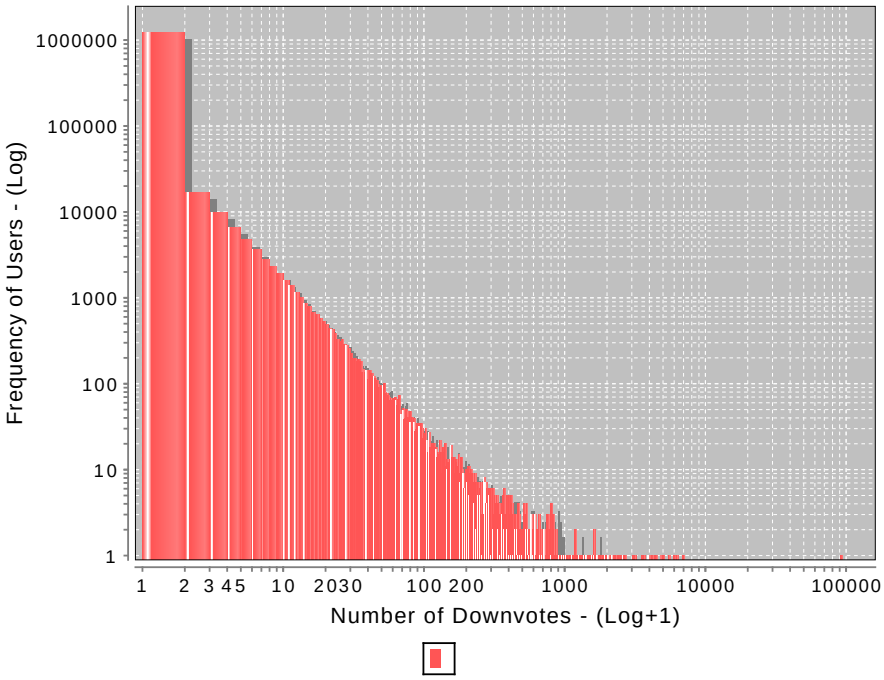


Users			
	Views	**Up Votes**	**Down Votes**
min	0	0	0
max	195496	17587	91092
median	0	0	0
mean	6.92	15.40	1.164
var	44714.20	19512.37	7141.739
std.dev	211.46	139.69	84.509
skewness	643	31	971
kurtosis	569104	1741	1042095

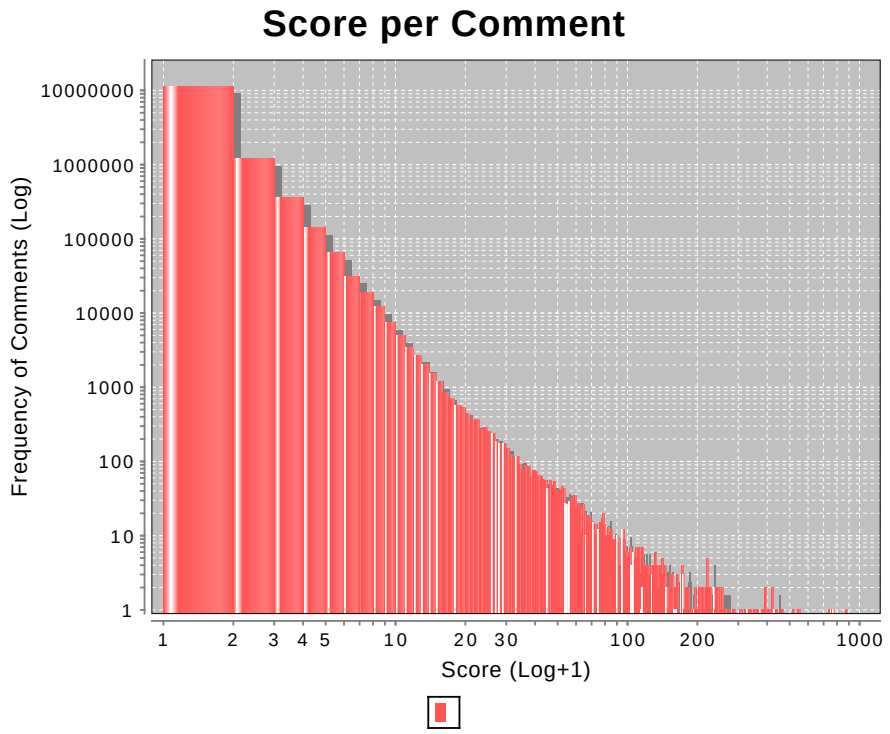
Upvotes per User



Downvotes per User



Comments	
	Score
min	0
max	872
median	0
mean	0.26117
var	2.31396
std.dev	1.52117
skewness	124
kurtosis	40884



3 Traceability

To find bug/issue IDs, I adapted the bug ID regex given by Fischer et al. [2]

I used regexes for email addresses and hyperlinks that I found here: http://net.tutsplus.com/tutorials/other/regular-expressions-you-should-know/. Plain text was contained primarily in the 'text' field of comments, and the 'body' field of posts.

Post Body			
Type	**Count**		
	Email Addresses	Issue IDs	Hyperlinks
Question	49748	608	12130779
Answer	30847	2321	12252559
Wiki	0	0	92
TagWikiExcerpt	0	0	728
TagWiki	19	1	21194
ModeratorNomination	1	0	272
WikiPlaceholder	0	0	9
Total	80615	2930	24405633

Comment Text			
Parent Type	**Count**		
	Email Addresses	Issue IDs	Hyperlinks
Question	2675	285	1241586
Answer	5680	539	2128905
ModeratorNomination	0	0	282
Total	8355	824	3370774

References

[1] Alberto Bacchelli. Mining challenge 2013: Stack overflow. In *The 10th Working Conference on Mining Software Repositories*, page to appear, 2013.

[2] M. Fischer, M. Pinzger, and H. Gall. Populating a release history database from version control and bug tracking systems. In *Software Maintenance, 2003. ICSM 2003. Proceedings. International Conference on*, pages 23–32, 2003.