Votes and Comments in Recommender Systems: The Case of Digg

Tasos Spiliotopoulos
Madeira Interactive Technologies Institute
University of Madeira

ABSTRACT

In this paper we describe Digg, a successful social news aggregator web site. Digg allows users to submit links to news stories, as well as vote and comment on other submitted stories. It also enables users to designate other users as friends and makes it easy to track their activities, thereby creating a social network within Digg.

We perform a statistical analysis of a sample of 1000 popular stories on Digg. We explore relationships and correlations between the two main ways of interacting with submissions on the website and we explain that votes (diggs) and comments constitute qualitatively different mechanisms for providing recommendations. Furthermore, we investigate the voting and commenting behavior for different content categories and discover significant differences among them.

Author Keywords

Digg.com, recommender systems, content categories, collaborative rating, news aggregation, statistical analysis

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

The social news aggregator website Digg is a successful example of a recommender system. In a typical recommender system users provide recommendations, which the system then aggregates and directs to the appropriate recipients. In some cases the primary transformation is in the aggregation, whereas in others the system's value lies in its ability to make good matches between the recommenders and those seeking recommendations [14]. The argument that the "wisdom of the crowd" can provide better recommendations seems very convincing to the internet user, while the increasing need to overcome overload and bias in the information provided electronically justifies tapping into this "pool of wisdom" even further. This accounts for the recent phenomenal rise in the popularity of social media sites, such as Digg, Slashdot, Reddit, StumbleUpon, Fark, del.icio.us and their like, as well as the countless blogs that emerge every day.

Digg is a social media site that exhibits the typical characteristics of recommender systems and attempts to combine them. On Digg, users can submit links to news stories, vote up (digg), vote down (bury) and comment on other users' links. In addition, the website allows its users to digg or bury other users' comments. Like other social media sites, Digg also allows users to create a profile and designate other users as friends and track their activities, such as what stories they submitted, and what they voted or commented on.

The main idea behind Digg is that the user voting system allows for the best stories on the web to surface to the front page and become available to all users. The promotion mechanism, therefore, does not depend on the opinion of a few editors, but emerges from the activities of its users. Digg is not edited by its staff, so that the people can collectively determine the value of the content submitted. However, occasionally the Digg staff might intervene to remove stories that violate the terms of use.

In this study we undertake an analysis of a sample of 1000 popular stories on Digg. We explore relationships and correlations between the two main ways of interacting with submissions on the Digg website and providing recommendations, namely diggs and comments. We also investigate the use of the website with regards to different categories of content.

RELATED WORK

An analysis of aggregate voting and rating behavior in the context of recommender systems [14] explains that such systems exhibit economies of scale; the bigger the number of users using this system, the more likely it is that the recommendations will be unbiased and the more likely one is to find someone with the same preferences and interests. Kostakos [9] takes a quantitative approach at analyzing users' voting and rating behavior and concludes that there is considerable bias in this behavior, which can be framed in terms of the voting mechanisms of the website. Studies on reviewer bias seek to provide a method that attempts to detect associations between people as an indicator of potential bias in online reviews [15].

Research focusing on the social aspect of the Digg website, shows that users with larger social networks within Digg are more successful in getting their stories promoted to the front page [13]. In addition, other studies suggest that the pattern of the spread of interest in a story on the Digg network is indicative of how successful the story will become. In particular, stories that spread mainly outside of the submitter's social neighborhood go on to be very popular, while stories that spread primarily through the submitter's neighborhood prove not to be very popular [10].

A study carried out in 2007 analyzes the role of popularity and novelty in attracting the attention of users, and finds that interest in a Digg story peaks when the story first hits the front page and then decays with time, with a half-life of about a day [21]. In a similar effort, an attempt was made to predict the long-time popularity of online content from early measures of user's access, such as views and votes [16].

Researchers argue that as social media move into the mainstream, the threat of vote spam becomes bigger. Towards combating this problem, they propose a machine-learning based ranking framework for social media that integrates user interactions and content relevance [1].

Social media websites typically exhibit long-tailed distributions in many aspects of their behaviour. Research attributes these long-tails to large differences among user activity rates, the time users devote to the site, and qualities of the rated content [7].

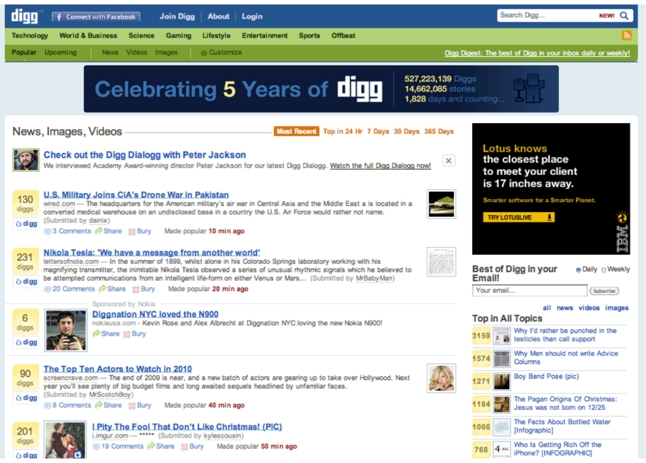
Addressing the problem of information overload can be viewed as an attempt to "turn down the noise" in the blogosphere. Towards this end, an an automated algorithm can be provided for learning user preferences from limited feedback [5].

Other researchers explain that the number of user supplied comments on a news article may be indicative of its importance or impact and try to predict the comment volume prior to an article's publication [20]. Another study focusing on blog comments, found that blogs act essentially as echo chambers since agreement outnumbers disagreement in comments by more than 3 to 1, with this ratio depending heavily on the blog's genre [6].

THE DIGG WEBSITE

History and Features

The Digg website (Figure 1) was originally launched on December 2004 with limited functionality. Version 2.0 was released on July 2005 and included a new interface, primitive social network characteristics (a friends list) and made use of GoogleAds. The third version of Digg was released on June 2006 and was updated to include specific categories for Technology, Science, World and Business, Videos, Entertainment and Gaming. The Digg website is constantly updated and, as of the time of this report, it offers the following categories of content: Technology, World and Business, Science, Gaming, Lifestyle, Entertainment, Sports, and Offbeat. Each of these categories comprises from 3 to 11 subcategories.



Check out the Digg Dialogg with Peter Jackson
We interviewed Academy Award-winning director Peter Jackson for our latest Digg Dialogg. Watch the full Digg Dialogg now!

130 diggs digg
U.S. Military Joins CIA's Drone War in Pakistan
wired.com — The headquarters for the American military's air war in Central Asia and the Middle East a is located in a converted medical warehouse on an undisclosed base in a country the U.S. Air Force would rather not name. (Submitted by dainix)
3 Comments Share Bury Made popular 10 min ago

231 diggs digg
Nikola Tesla: 'We have a message from another world'
lettersofnote.com — In the summer of 1899, whilst alone in his Colorado Springs laboratory working with his magnifying transmitter, the inimitable Nikola Tesla observed a series of unusual rhythmic signals which he believed to be attempted communications from an intelligent life-form on either Venus or Mars… (Submitted by MrBabyMan)
20 Comments Share Bury Made popular 20 min ago

Sponsored by Nokia

6 diggs digg
Diggnation NYC loved the N900
nokiausa.com - Kevin Rose and Alex Albrecht at Diggnation NYC loving the new Nokia N900!
Share Bury

90 diggs digg
The Top Ten Actors to Watch in 2010
screencrave.com — The end of 2009 is near, and a new batch of actors are gearing up to take over Hollywood. Next year you'll see plenty of big budget films and long awaited sequels headlined by unfamiliar faces. (Submitted by MrScotchBoy)
8 Comments Share Bury Made popular 40 min ago

201 diggs digg
I Pity The Fool That Don't Like Christmas! (PIC)
i.imgur.com — ***** (Submitted by kylescousin)
19 Comments Share Bury Made popular 50 min ago

3159	Why I'd rather be punched in the testicles than call support
1574	Why Men should not write Advice Columns
1271	Boy Band Pose (pic)
1184	The Pagan Origins Of Christmas: Jesus was not born on 12/25
1066	The Facts About Bottled Water [Infographic]
768	Who Is Getting Rich Off the iPhone? [INFOGRAPHIC]

Figure 1. A screenshot of the Digg.com front page

The website is very flexible it terms of the ways that it allows a user to access the stories. The user may browse the stories by category or subcategory, view the currently popular stories (i.e. stories that made it to the front page), the upcoming stories (i.e. stories that were very recently submitted but have not been made popular yet), the most recent stories, and the most popular stories in the last 24 hours, 7 days, 30 days, or 365 days. In addition, Digg features a new recommendation engine that suggests stories to users based on their recent digging activity.

Digg enables and encourages its users to create a profile and take advantage of the social network features it provides. Users have a significant degree of control over their profile. They can customize it by providing details and photographs of themselves, declare their favorite stories, and filter the topics and the media types that they wish to view stories about. Users have access to all their history, and specifically all their diggs, submissions, comments, favorites, and profile activities. They also have access to a privacy control panel, which enables them to supervise and manage the extent of their personal information and usage behaviour that is displayed to their friends and other users of the website. The "friends' activity" panel is a running list of all of a user's friends' actions within Digg, such as diggs, comments and submissions.

Controversy

Since Digg is a very popular news web site, an appearance of a story or link on its front page can have a significant impact and can even lead to the "digg effect", a sudden and massive increase in traffic to the web site in question. However, there has been a lot of controversy regarding the

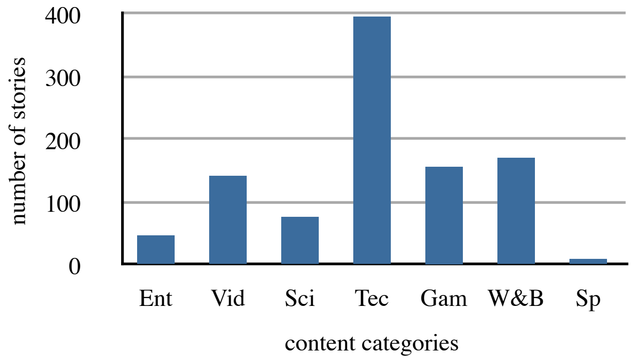


Figure 2. Distribution of stories across the categories.

way that stories get promoted to the front page of the web site. Although the particular algorithm that is used for this is kept secret and is updated by the Digg staff whenever there is a suspicion of unfairness, there appear to have been successful attempts in the past to take advantage of shortcomings in the algorithm and manipulate the promotion system [2]. This has lead to a gradual ban of a large number of Digg members, even users that were regular and active members of the community. In fact, there are reports of an actual industry based on getting stories promoted to the Digg front page [3]. Interestingly enough, one of the all-time most dugg stories [4] is dedicated to complaints about the unfairness of the algorithm and calls the web site undemocratic.

In June 2009 Digg introduced a new kind of advertisements, the Digg Ads [17]. These are advertisements that are presented like user-submitted stories and people can vote on them, just as they would with submitted stories. The number of diggs influences how often an ad gets displayed, and ultimately how much the advertiser pays per click. This is expected to encourage advertisers to create content as compelling as organic Digg stories, and to indirectly give users more control over which ads they see on the web site. However, these ads can easily be confused with actual stories by users not accustomed to this presentation of advertising material.

Another point of controversy has been the DiggBar, a feature introduced in April 2009. This is a toolbar above the top of a site which allows the user to produce shortened URLs and access digg comments and analytics without leaving the page [19]. When a user follows a link from a Digg story, the DiggBar does not redirect the user to the original URL, but instead frames the linked page and displays it within a Digg shortened URL. After several complaints from the webmasters of other sites that the shortened URLs fail to give link credit to the original sites and thus lower their page rank, several content management systems released plugins that block the DiggBar.

There was an internet-wide discussion and massive concern when a story appeared on the Digg front page that contained the AACS encryption key, something that was considered to be illegal. This story was later removed by the Digg staff, along with all other submissions that contained this number, having also the users that submitted it banned from the web site. After public outcry, Digg stepped back and decided not to delete any new stories mentioning this number [18]. This raises the important question of the extent of accountability that a community site like Digg should have for any inappropriate/illegal content posted by its users.

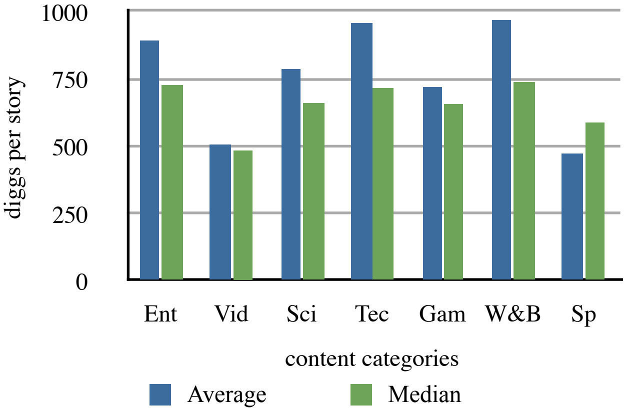


Figure 3. Mean and median of the number of diggs per story for each content category.

METHOD

For this study we used a dataset of the most recent 1000 popular Digg stories, as of January 30th 2007. By popular, it is assumed that we mean the stories that made it to the front page. The information collected for each story comprise the number of diggs, the number of comments, the container (i.e. primary category of the content), the sub-category, and the title of the story. In our analysis, we did not take into account the sub-category of the content and we used the title of a story as its identifier. We used this dataset because it was freely available and therefore convenient to acquire. It was downloaded from the IBM *Many Eyes beta* web site [8]. Although this tool can provide a large number of visualizations for such data, we decided to perform further analysis of our own in order to identify trends on the data.

As of January 2007, when the data were collected, stories were classified in 7 main categories, namely Entertainment, Videos, Science, Technology, Gaming, World and Business, and Sports. Figure 2 shows the distribution of articles in these categories. As evident from Figure 2, almost 40% of the number of stories belonged to the Technology category. This segment represents more than twice the percentage of the next category, something than can be attributed to the fact that Digg started as a technology oriented website. Categories that were added later, such as Entertainment and Sports account for a significantly smaller proportion of the submitted stories. Today that the website has reached a more mainstream audience, the distribution of popular

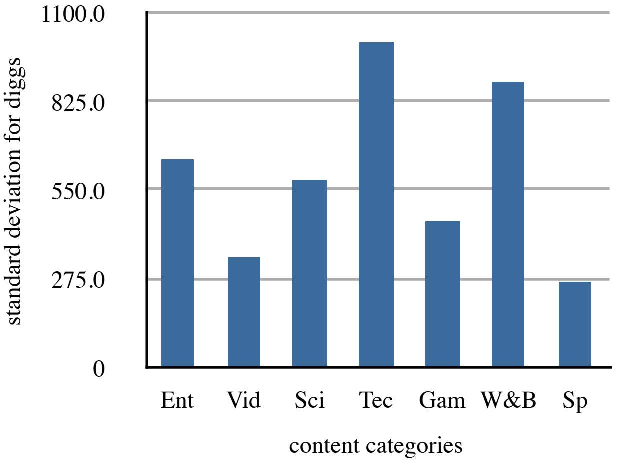


Figure 4. Standard deviation STDEVP of the number of diggs per story for each content category.

stories across categories would be expected to be more uniform.

For each content category in the dataset, we calculated the arithmetic mean (average), median and standard deviation (STDEVP) of the number of diggs for each story, and the mean, median and standard deviation of the number of comments for each story. We expected the comments-to-diggs ratio of each story to provide useful insights on hoe these two features were used, so we calculated the mean, median and standard deviation of this metric, too.

RESULTS

Figure 3 shows the mean of the number of diggs per story for each category, as well as the respective median. The mean of the number of diggs per story ranges from 472 for the Sports category to roughly the double of that number for the World and Business category. The median is always slightly lower, except for the sports category where it is slightly higher.

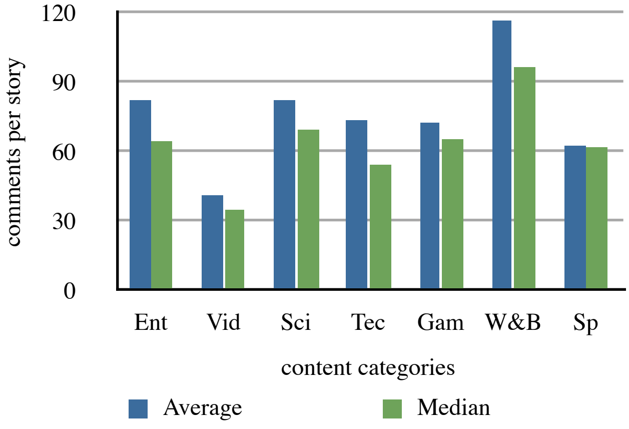


Figure 5. Mean and median of the number of comments per story for each content category.

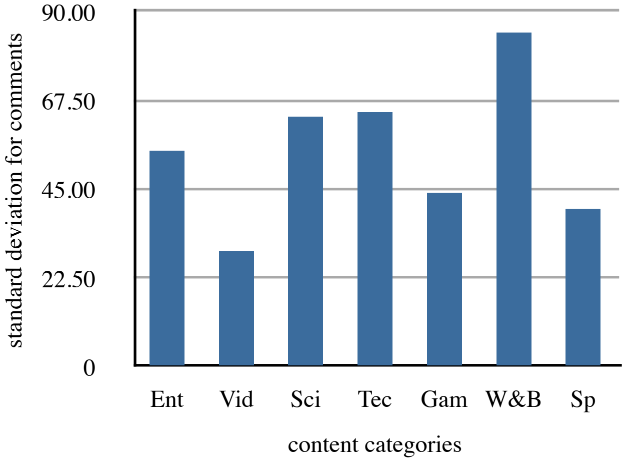


Figure 6. Standard deviation STDEVP of the number of comments per story for each content category.

The standard deviation STDEVP of the number of diggs per story for each category is shown in Figure 4. The standard deviation ranges from 265 for the Sports category to almost the quadruple of that (1008.1) for the Technology category. Interestingly, there appears to be a correlation between the average number of diggs per story and the corresponding standard deviation; the categories with a higher average number of diggs per story exhibit a higher standard deviation. However, this can be considered normal since we calculated the standard deviation in absolute numbers.

Figure 5 shows the mean of the number of comments per story for each category, as well as the corresponding median. The mean of the number of comments per story ranges from 40.6 for the Videos category to almost the triple of that number (116.3) for the World and Business category. The median is again slightly lower for all categories.

The standard deviation STDEVP for the number of comments per story for each category is depicted in Figure 6. The standard deviation ranges from 29.36 for the Videos category to 84.61 for the World and Business category. Again, the correlation between the average number of comments per story and the corresponding standard deviation for each category can be explained by the fact that we are measuring standard deviation in absolute numbers.

We measured the comments-to-diggs ratio for each story and then calculated the mean, median and standard deviation for each category. Interestingly, the mean and median values were not found to be uniform across categories. The World and Business category and the Sports category were found to have a comments-to-diggs ratio of 0.184 and 0.167 respectively, with the rest of the 5 categories being quite uniform with a ratio between 0.119 and 0.134. Once more, the median values were consistently lower as shown in Figure 7.

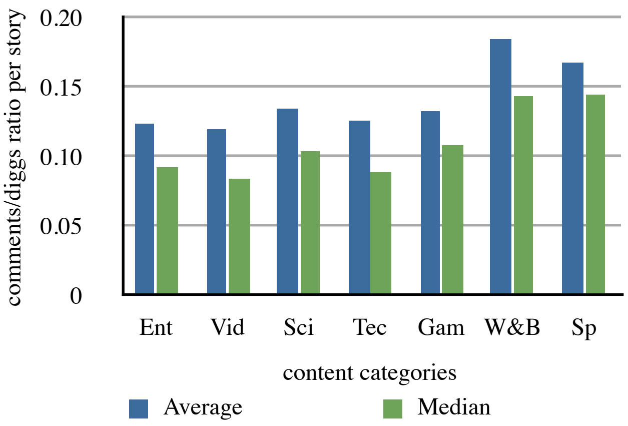


Figure 7. Mean and median of the comments-to-diggs ratio per story for each content category.

The standard deviation for the comments/diggs metric (Figure 8) shows a high value for the World and Business (0.159) and the Entertainment (0.151) categories, with the other 5 categories exhibiting consistently a significantly lower value ranging from 0.091 to 0.106.

DISCUSSION

We observed that the mean of the number of diggs and comments for each story is almost always higher than the median. We attribute this to the fact that there are a few number of stories with a very high number of diggs and comments which raise the average. For instance, there was a story with 8293 diggs (more than 11 times the median value) and a story with 547 comments (more than 10 times the median value for its category), both in the Technology category. We attribute the only exception to this rule - the median number of diggs is higher than the mean for the Sports category - to our small sample, since there were only 10 stories in the Sports category.

We find that the mean and median of the number of diggs and comments per story is significantly diversified across categories. This merely means that the readers generally find stories in some categories more interesting than stories in other categories and, thus, digg and comment on them more.

The high standard deviation of the comments-to-diggs ratio implies that votes and comments are qualitatively different mechanisms for providing recommendations. Comments provide more information than simple diggs and some stories are more suitable for comments than diggs. So, some stories in the Entertainment and World and Business categories elicited a relatively high number of comments, thus raising the average comment-to-digg ratio for these categories. It seems that certain stories with regards to entertainment (e.g. celebrities) and world and business (e.g. politics or economy) evoked considerably more discussion in the form of comments than stories in other categories.

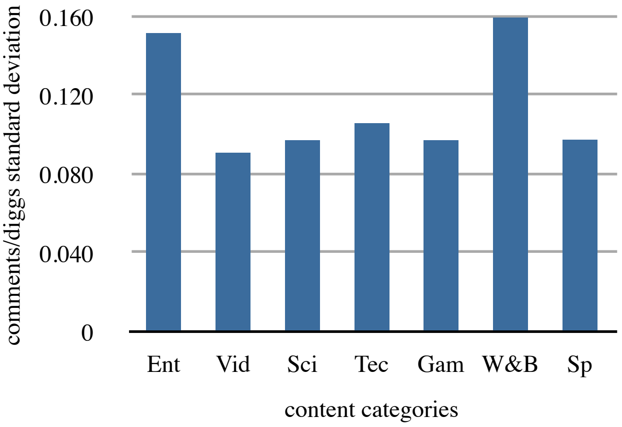


Figure 8. Standard deviation STDEVP of the comments-to-diggs ratio per story for each content category.

One would expect that the number of diggs a story receives would be very highly correlated to the number of comments, resulting in a rather uniform distribution of the comments-to-diggs ratio across categories. However, that is not the case. If we dismiss the value for the Sports category, due to the small sample, it is evident that the World and Business category exhibits a significantly higher comments-to-diggs ratio than the other 5 categories, which are fairly consistent. This indicates that different content categories exhibit different recommendation patters and more controversial topics present a higher comments-to-diggs ratio. This observation is inline with one of the findings from analyses on blogs [6], which indicates that blog genre has a significant impact on agreement proportions. In that study, technology and entertainment blogs were found to inspire less polarization and have a much lower agreement to disagreement ratio than the other three genres.

CONCLUSIONS AND FUTURE WORK

We performed a statistical analysis of a sample of 1000 popular stories on the social news website Digg and explored relationships and correlations between the two main ways of interacting with submissions on the website and providing recommendations, namely votes and comments. The high standard deviation of the comments-to-diggs ratio that we observed implies that votes and comments are qualitatively different mechanisms for providing recommendations. Furthermore, contrary to our expectations, the distribution of the comments-to-diggs ratio across categories was not found to be uniform, indicating that different content categories exhibit different recommendation patterns, with certain topics presenting a higher comments-to-diggs ratio than others.

It would be of great interest to investigate the effectiveness of this collaborative determination of the value of content that Digg attempts, by measuring the effectiveness and the user satisfaction both of the site's suggestions (the popular stories) and the Digg recommendation engine. Furthermore, the differences between a moderated news site (such as Slashdot) and an unmoderated one (such as Digg) for the same content can be a very interesting topic to explore. Finally, another stimulating question would be to determine the extent to which a moderated story or a comment by a site editor can influence the outcome of the users' discussion in the comments section of a moderated news site.

REFERENCES

1. Bian, J., Liu, Y., Agichtein, E., and Zha, H. A few bad votes too many? Proceedings of the 4th international workshop on Adversarial information retrieval on the web - AIRWeb '08, ACM Press (2008), 53.

2. BusinessWeek.com. Digg Argues It Has Ways to Prevent Manipulation. http://www.businessweek.com/the_thread/blogspotting/archives/2006/03/digg_argues_it.html.

3. CNET News. Digg continues to battle phony stories. http://news.cnet.com/Digg-continues-to-battle-phony-stories/2100-1025_3-6144652.html.

4. Digg.com. Digg this if your sick of power users stealing stories. http://digg.com/tech_news/Digg_this_if_your_sick_of_power_users_stealing_stories.

5. El-Arini, K., Veda, G., Shahaf, D., and Guestrin, C. Turning down the noise in the blogosphere. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09, ACM Press (2009), 289.

6. Gilbert, E., Bergstrom, T., and Karahalios, K. Blogs Are Echo Chambers: Blogs Are Echo Chambers. HICSS 2009.

7. Hogg, T. and Szabo, G. Diversity of User Activity and Content Quality in Online Communities. Proceedings of the Third International ICWSM Conference, (2009), 58-65.

8. IBM. Many Eyes website. http://manyeyes.alphaworks.ibm.com/manyeyes/.

9. Kostakos, V. Is the crowd's wisdom biased? A quantitative analysis of three online communities. Adjunct proceedings of IEEE SocialComm, International Symposium on Social Intelligence and Networking (SIN09), (2009).

10. Lerman, K. and Galstyan, A. Analysis of social voting patterns on digg. Proceedings of the first workshop on Online social networks - WOSP '08, ACM Press (2008), 7.

11. Lerman, K. Social networks and social information filtering on digg. Arxiv preprint cs/0612046, (2006).

12. Lerman, K. Dynamics of collaborative document rating systems. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis - WebKDD/SNA-KDD '07, ACM Press (2007), 46-55.

13. Lerman, K. Social Information Processing in Social News Aggregation. Information Sciences, 1 (2007), 1-20.

14. Resnick, P. and Varian, H.R. Recommender systems. Communications of the ACM 40, 3 (1997), 56-58.

15. Staddon, J. and Chow, R. Detecting reviewer bias through web-based association mining. Proceeding of the 2nd ACM workshop on Information credibility on the web - WICOW '08, ACM Press (2008), 5.

16. Szabo, G. and Huberman, B.A. Predicting the popularity of online content. Computing, (2008).

17. The Digg Blog. Ads You Can Digg…or Bury. http://about.digg.com/blog/ads-you-can-digg?or-bury.

18. The Digg Blog. Digg This: 09-f9-11-02-9d-74-e3-5b-d8-41-56-c5-63-56-88-c0. http://about.digg.com/blog/digg-09-f9-11-02-9d-74-e3-5b-d8-41-56-c5-63-56-88-c0.

19. The Digg Blog. DiggBar Launches Today! http://about.digg.com/blog/diggbar-launches-today.

20. Tsagkias, M., Weerkamp, W., and de Rijke, M. Predicting the volume of comments on online news stories. Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09, ACM Press (2009), 1765.

21. Wu, F. and Huberman, B.A. Popularity, novelty and attention. Proceedings of the 9th ACM conference on Electronic commerce - EC '08, ACM Press (2008), 240.

The columns on the last page should be of approximately equal length.