2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

An Exploration of Discussion Threads in Social News Sites: A Case Study of the Reddit Community

Tim Weninger Xihao Avi Zhu Jiawei Han
Department of Computer Science
University of Illinois Urbana-Champaign
Urbana, Illinois 61801
Email: {weninge1, zhu40, hanj}@illinois.edu

***Abstract*—Social news and content aggregation Web sites have become massive repositories of valuable knowledge on a diverse range of topics. Millions of Web-users are able to leverage these platforms to submit, view and discuss nearly anything. The users themselves exclusively curate the content with an intricate system of submissions, voting and discussion. Furthermore, the data on social news Web sites is extremely well organized by its user-base, which opens the door for opportunities to leverage this data for other purposes just like Wikipedia data has been used for many other purposes. In this paper we study a popular social news Web site called Reddit. Our investigation looks at the dynamics of its discussion threads, and asks two main questions: (1) to what extent do discussion threads resemble a topical hierarchy? and (2) Can discussion threads be used to enhance Web search? We show interesting results for these questions on a very large snapshot several sub-communities of the Reddit Web site. Finally, we discuss the implications of these results and suggest ways by which social news Web site's can be used to perform other tasks.**

I. Introduction

Social news Web sites are platforms in which (1) users generate or submit links to content, (2) submissions are voted on and ranked according to their vote totals, (3) users comment on the submitted content, and (4) comments are voted on and ranked according to their vote totals. These platforms provide a type of *Web-democracy* that is open to all comers. These social news Web sites have become exponentially more popular during the past few years. Some popular social sews sites include Digg, Reddit, Slashdot, StumbleUpon and Technorati among hundreds of others.

These social frameworks represent a stark departure from traditional media platforms in which a news organization, *i.e.*, a handful of television, radio or newspaper producers, sets the topics and directs the narrative. Social news sites increasingly set the news agenda, cultural trends, and popular narrative of the day. The notion that Web blogs, in particular, drive the media narrative was initially presented in research literature by the MemeTracker project [1]. As this trend continues and grows, the number of blogs, news outlets, and other sources of user generated content has outpaced the rate at which Web users can consume information. Social news sites and their many subtopic pages are able to automatically curate, rank and provide commentary on the top content of the day by harnessing the power of the masses.

One of the most interesting and important features of social news sites is the ability for users to comment on a submission. These comment threads provide a user-generated and user-curated commentary on the topic at hand. Unlike message board or Facebook-style comments that list comments in a flat, chronological order, or Twitter discussions that are person-to-person and oftentimes difficult to discern, comment threads in the social news paradigm are permanent (although editable), well-formed and hierarchical. The hierarchical nature of comment threads, where the discussion structure resembles a tree, is especially important because this allows divergent sub-topics resulting in a more robust overall discussion.

In this paper we explore the social news site Reddit in order to gain a deeper understanding on the social, temporal, and topical methods that allow these types of user-powered Web sites to operate. This paper presents first-of-a-kind, large-scale study of posts and comments on a social news site. Specifically, we explore the typical structure of a comment thread, and how does it evolve over time?

II. Dataset Description

User-powered social news sites such as Reddit, Slashdot and others have similar setups and user interaction schemes. Web users may access these sites anonymously (without an account) in read-only mode where they can browse postings and comments, but not contribute, vote or comment. Account creation typically only requires a username, password, and the passage of a challenge-response test (*e.g.*, Captcha-test); thus users typically remain anonymous. Registered users may contribute posts, comment and vote.

We chose to study Reddit in particular because (1) the user-community is very active, (2) the Web site has a soaring popularity, and (3) *all* posting, comment and aggregate user data is publicly accessible.

Reddit, in particular, is beginning to influence the world in ways that both the mainstream media and research community do not yet fully understand. The Reddit community is able to bring a higher order of organization to online content, and is changing the methods of discourse online. Recent posts by presidents, including Barack Obama, Nobel laureates, A-list actors, singers, astronauts, scientists, CEOs, and so on reinforce this trend.

Before we introduce the experimental dataset, we describe the basic framework for the Reddit system:

Subreddits. Reddit is comprised of thousands of user-created and user-moderated *subreddits*, which are topical forums for content. For example, there is a general POLITICS

ASONAM'13, August 25-29, 2013, Niagara, Ontario, CAN

Capture Dates	7/25/2012 – 11/19/2012
Users	1,154,184
Posts	369,833 (across 25 subreddits)
Post Votes	488,555,185 (58% Upvotes)
Comments	16,540,321
Comment Votes	371,439,104 (79% Upvotes)

TABLE I: Statistics of the Reddit dataset

subreddit as well as CONSERVATIVE, LIBERAL, PROGRESSIVE, etc., subreddits. Any user can create and moderate a subreddit at any time, and Reddit administrators rarely interfere with subreddits. New users are auto-subscribed to a handful of popular subreddits, and other subreddits can be subscribed to according to the user's interests. Certain subreddits have specific rules that determine what can and can not be posted, for example, PICS requires posts to be only pictures. It is unclear, and outside the scope of this paper, if these rules play any part in this study's results. There used to be a general subreddit called REDDIT.COM, but it was removed to encourage topical discussion.

Posts. Regardless of subreddit subscription status, any registered user can contribute to any subreddit by submitting a link to external content or by creating a self-post. Self-posts are Wiki-style text with a generous 10,000 character limit.

Comments. Registered users can also comment on posts. The comment pages of Reddit are hierarchically threaded, *i.e.*, a comment can be in response to the post in general (a root comment), or in reply to another comment. This creates a discussion hierarchy and facilitates discussion subtopics.

Voting. Registered users are able to *upvote* or *downvote* posts and comments; one vote per post/comment per user, +1 point per upvote, -1 point per downvote. Posts and comments are displayed on the site in sorted order according to a time and vote total ranking function. Popular posts may trigger "vote fuzzing", which is an anti-spam mechanism and the only closed-source part of Reddit. According to the Reddit FAQ[1] the vote fuzzing mechanism changes the number of up and down votes; the vote score *i.e.*, upvotes - downvotes, is not changed.

Karma. When a post or comment receives votes, the user who contributed the post or comment receives *karma*. For example, if a user submits a link to an article that receives a total of 10 upvotes and 2 downvotes, then that user will receive 8 karma points. Post-karma and comment-karma are counted separately. Self-posts do not receive karma points. Users with a large amount of karma are allowed to contribute more frequently. This rewards users who contribute high quality content and make insightful, amusing or otherwise interesting comments.

To gather a dataset sufficient for a large-scale exploration, we crawled the Reddit API four times daily: at 0:00, 6:00, 12:00 and 18:00 CST. During each crawl we retrieved the 100 top-scoring posts from the 25 most popular subreddits[2], as well as the 100 top-scoring posts of the day from across all subreddits. From each post we retrieve the 500 top-scoring comments, with a depth limit of 10. Each post and comment has submission time, text, username, and vote totals. To ensure we gather complete voting results, comments, and full set of edits, we initially only *note* the top posts and comments; we actually *collect* the complete text, votes, etc. after 48 hours has elapsed. Results presented later in this paper demonstrate that 48 hours is a sufficient waiting period; in fact, we find that the vast majority of activity occurs within the first 4 hours of a post's life-cycle. We also collect the registration date and aggregate karma scores for each user we encounter. Of course, we would like to collect the full set of data, but Reddit asks that crawlers limit the number of API requests to one per second making this full dataset impossible to collect without violating the terms of service. Table I has statistics of the collected data.

Posts and comments are frequently deleted. However, our data capture system does not make any effort to delete a comment or post from the captured dataset if it has been deleted on Reddit post hoc. Obviously, if a post is deleted before the crawl, then it cannot be captured. However, if a comment received replies before it was deleted prior to the crawl, then the Reddit API will return `[deleted]` as the author and text. Deleted comments are ignored in all evaluations, but children of deleted comments are not ignored.

We mentioned earlier that Reddit has experienced remarkable growth in the past several years. In August 2012 Reddit reported 3.4 billion page views over 42.9 million unique visitors. Our dataset, however, only captures data about a subset of active users that contribute at least one post or comment in a top 25 subreddit during our crawl period.

III. TOPICAL HIERARCHIES AND THE EVOLUTION OF A COMMENT THREAD

Topical clustering algorithms, such as LDA [2] and its hierarchical cousin hLDA [3], have received a lot of recent attention both in research literature and in commercial system development. Hierarchical LDA, in particular, clusters words into hierarchical topics such that general words appear towards the top of the hierarchy, and specific words appear at the leaves of the hierarchy.

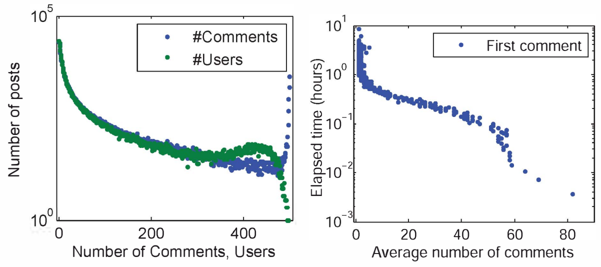
Comment threads on Reddit are structurally hierarchical, that is, a comment can be a reply to the post (a root comment) or a comment can be in reply to another comment. This subsection investigates the extent to which comment hierarchies exhibit a topical hierarchy. If we find that comment threads are topically hierarchical as we expect, then perhaps comment threads could be used to enhance future developments in topic models. On the other hand, if we find little or negative correlation between topic and discussion hierarchies, then we would need to rethink our assumptions about hierarchical topic models, discussion threads or both.

We are also interested in how discussions topics evolve temporally and structurally. In temporal terms, we ask the question: does the discussion diversify as time passes? or does the discussion diversify immediately and then stay topically disjoint? In structural terms, we investigate the effect that a comment's thread depth has on its topical granularity and its ultimate vote score.

Previous studies have examined the structure of comment threads by analyzing the radial tree representation of thread hierarchies [4], via a text classification problem [5], and

[1] http://www.reddit.com/wiki/faq

[2] http://www.reddit.com/reddits/, accessed on 7/24/2012



(a) (b)

Fig. 1: Distribution of the number of comments and users per discussion thread (a). Average number of total comments as a function of the elapsed time to the first comment (b).

by examining discussion *chains* [6]. A relevant study by Kaltenbrunner *et al.* on the hierarchical comments of Slashdot found that the volume of comments over time represented a lognormal distribution [7].

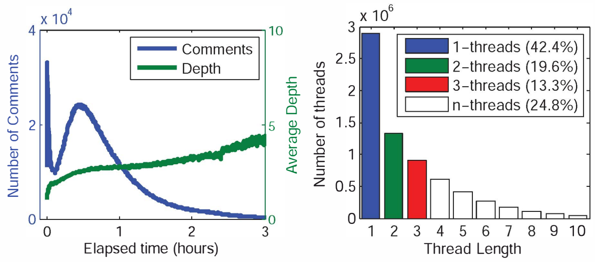
Very little is known about the topical distribution of comment hierarchies. We hypothesize that comment threads are topically similar to the contributed content, and that subtopics emerge as discussion progresses and the thread hierarchies deepens.

A. *Comment Threads over Time*

Recall that our dataset contains the top-scoring posts from the most popular subreddits; thus the values in this section are likely to be inflated in comparison to less popular subreddits. In our dataset, posts receive an average of 53 comments, and half of all posts receive 10 comments or fewer. A small number of highly discussed posts, however, can receive tens-of-thousands of comments, although we only collect the 500 highest scoring comments.

Figure 1a shows the number of distinct users and comments per posting. This figure shows a heavily tailed distribution similar to the findings of Laniado *et. al* on Wikipedia's discussion dataset [6]. However, a major difference is found in the tail of the distribution: there is a drastic uptick in the number of articles having between 450 and 500 comments (blue points). This is explained by the observation that when postings become extremely popular they are displayed on Reddit's front page garnering even more attention, while the majority of posts, not receiving front-page attention, fall somewhere in the long-tail distribution. The number of users per discussion (green points) exhibits a moderate deviation in the tail, that is, there are more discussions with 400 distinct users than with 350 distinct users. This is also a result of popular posts being listing on Reddit's front page. The steep decline in postings with between 480-500 distinct users solely is an artifact of the 500 comment collection limit: it is rare to find a post with 500 comments from 500 distinct individuals.

Figure 1b shows the average number of comments as a function of the elapsed time to the first comment. We find that when the first comment is submitted early-on in the post's life-cycle, then the post is likely to receive a large amount of comments. Conversely, when the first comment is submitted later in the post's life-cycle, then the post is not likely to have a large number of comments. This effect is causal because posts having a large (or small) number of comments must start with the first comment.



(a) (b)

Fig. 2: Number of comments and average comment depth as a function of time (a). Number of discussions at different depths (b). Percentages indicate the proportion of comments in different depths.

As time passes the number of comments ought to increase. Figure 2a shows the rate of commenting as a function of the elapsed time in hours (blue). We see that, in aggregate, there is a spike in extremely early commenting; these early comments come as soon as 1 to 5 seconds after the posting. After the initial surge the comment rate gradually rises and falls over the aggregate lifetimes of all posts. Except for the initial spike, our result are consistent with the lognormal distribution reported by Kaltenbrunner *et al* [7].

The depth of a comment in the discussion hierarchy refers to the number of ancestors the comment has. Also in figure 2a we find that the average depth (green) steadily increases as the discussion progresses. The next subsection discusses the topicality of comments given their time and depth.

The density of discussions at progressive depths is illustrated in figure 2b. Clearly, most comments are situated at the top level (depth of 1), and the number of comments at each successive depth trails off exponentially.

B. *Structure of Comment Threads*

As a comment thread evolves new comments are added in response to parent-comments, and users vote on older comments. The previous subsection showed aggregate statistics for thread depth and timeliness. Figure 3 illustrates a discussion thread for a randomly chosen post. In this illustration bright/red colors indicate early comments while dark/blue colors indicate later comments, and large circles indicate higher vote scores, while smaller circles indicate low (and sometimes negative) vote scores.

We see that many of the first-level comments are early comments, and the comments tend to become darker as their depth increases. Likewise, first-level comments are typically high scoring, and the comments tend have lower vote scores as their depth increases. Figure 3 also hints and an answer to one of our original questions: does the discussion diversify as time passes, or does the discussion diversify immediately? Observations from the radial comment thread illustration and Figure 2a show that subthreads (and presumably their subtopics) are

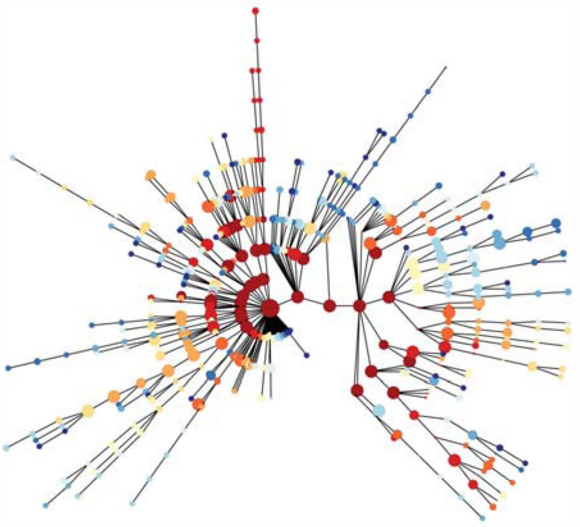


Fig. 3: Structure of a randomly selected comment thread[3]. Early comments are in bright colors, later comments are in dark colors. Node sizes indicate each comment's final vote score.

started early in a post's life-cycle and *also* diversify further, creating sub-subthreads, later in the post's life cycle.

One particular sub-discussion on the right-hand side of the radial comment thread illustration in Figure 3 developed quickly, and has a comparatively broad fanout along with relatively high scores. In general, we find that Reddit discussions typically have one or two sub-threads that receive the most attention, by way of comments and votes, and these high-attention sub-threads usually develop relatively quickly.

C. Topical hierarchies

Previous figures show that as time progresses the average comment depth increases. We believe that this is, in part, an artifact of the nature of online discourse. More concretely, the results from the previous subsection suggests that when an online discussion first begins users contribute top-level comments that often initiate various threads of discourse. Based on these observations we ask: do hierarchical threads, like those on Reddit a) demonstrate a hierarchy of topics; or b) do hierarchical threads present a flat or narrowing topical representation.

For example, an illustration of the two types of threads is found in Table II. The first discussion is a debate between and about climate change skeptics - a relatively narrow topic with back-and-forth rebuttals, etc. The second discussion is more topically diverse, and the topics continue to diversify into subtopics as the comment hierarchy deepens. Specifically, the root comment talks about the article's proposed solution, this topic is then subsumed by discussion on wind and solar energy in one subthread and oil in another subthread, which is further diversified into nuclear alternatives instead of solar/wind, etc.

12 hottest years on record have come in the last 15 years

This is the best site to discredit climate deniers...

The reason people are skeptical is because they should be...

There is not one item in this response that even makes a serious attempt at making an argument...

The problem with skeptics of all kinds is that their approach is...

The [problem] in that argument is that facts show...

Too bad his "solution" is fracking and "clean" coal.

Clean coal lol

And a vast expansion in solar and wind energy over the past several years...

Wind and solar energy are inefficient, nuclear energy is where it is at.

I think people underestimate the influence of big oil over governments.

People also underestimate the influence of big oil over their own lives.

TABLE II: Truncated discussion thread showing topically narrow thread (top) and topically diverse thread hierarchy (bottom).

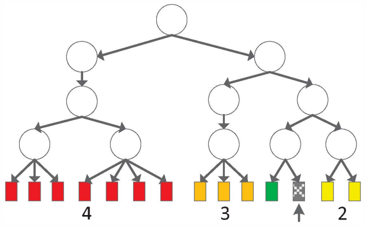


Fig. 4: Illustration of 4 level hLDA output. Green, yellow, orange, red indicate most topically similar to least topically similar.

Unlike this small, truncated example, actual comment threads can contain thousands of comments and deep and broad thread-trees. In this subsection we investigate the extent to which threads trees are topically hierarchical. Fortunately, recent advances in hierarchical topic models allow for a systematic, quantitative evaluation of the topical distributions in text hierarchies.

Latent Dirichlet Allocation (LDA) [2] and its nonparametric/hi-erarchical extension (hLDA) [3] are two commonly used probabilistic topic models. Given a set of documents hLDA hierarchically clusters comments/documents so that topically similar documents share the same topic-parent, less-similar comments share topic-grandparents, etc. In essence, the topical distance between two comments can be measured by the tree-distance in the hLDA output; sibling-comments have more in common than cousins, who have more in common than second-cousins, etc.

Figure 4 illustrates, with respect to a given document/comment (indicated by the arrow at center-right), that documents/comments that are topically similar share an early common ancestor. Likewise, documents/comments that are topically dissimilar share distant ancestors.

The goal, therefore, is to measure if and how topics diverge as discussion threads deepen. This measurement is accomplished by a straightforward methodology. First, we randomly sample 10,000 postings resulting in 429,041 comments. For each post hLDA was run for 5,000 Gibbs iterations at heights of 3, 4, and 5, the sample with the highest log likelihood was captured as the output model. The LDA model, which can loosely be characterized as hLDA with a height of 2, was likewise run for each posting with $k = 25$.

If discussion threads exhibit a topic hierarchy, then topi-

[3] http://redd.it/100icq – "Former National Security Agency official Bill Binney says US is illegally collecting huge amounts of data on his fellow citizens — The Guardian"

[2] Full discussion available at http://redd.it/18l9je/

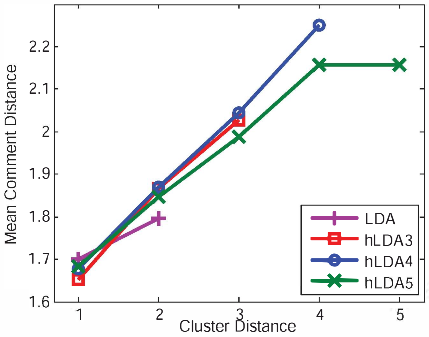


Fig. 5: Average distance between comments as a function of cluster distance

cally similar comments should appear in the same or similar hLDA clusters. To evaluate the comment threads, we determined the distance between each pair of comments in the hLDA output, measured by distance to the least common ancestor, resulting in n^2 total distance measurements. Measurements are averaged across all pairs of comments, and then across all of the 10,000 posts for each level of hLDA. Because the LDA model does not generate a hierarchy, the comments either exist in the same cluster, determined by picking the most probable cluster or each document, or they do not. If comments threads do not exhibit a topical hierarchy, then we expect to find a low correlation between the comment thread distance and the topical cluster distance, and vice versa.

Figure 5 shows the aggregate results of these measurements plotted against the average distance in the actual thread hierarchy. HLDA with a depth of 3 can only show results for cluster distances of 1, 2 and 3 because the maximum cluster distance is 3; similar restrictions exist with hLDA with depths of 4 and 5. Comments that are clustered together (siblings) in the LDA/hLDA output have, on average, a small least common ancestor distance in discussion threads. This shows that, in the aggregate, comments in a discussion thread that share a common ancestor are more topically similar. These results seem to show that thread structures correlate to thread topicality. In other words, thread hierarchies tend to exhibit a topical hierarchy in the general case. We stress that these measurements are for the general case; there are certainly cases in which the opposite is true like in the top thread example of Table II. These initial measurements should serve as motivation for a more complete investigation of topical hierarchies in discussion threads.

IV. Conclusions

We conclude by revisiting the original questions raised at the beginning of this work.

Regarding the structure and evolution of a comment thread, we observe that, in general, hierarchical comment threads consist of top level comments that start a subtopic. We also observe that these top level comments, especially those which receive a large number of replies, are usually created during the early stages of the post's life cycle. From among the early, top-level comments/subtopics further sub-subtopics are created as a natural part of online discourse. Topically, we find that the document clusters found with the word frequency-based hierarchical clustering algorithm hLDA are correlated with the least common ancestor distance within the comment thread tree. In plain terms, we present strong evidence that hierarchical comment threads on Reddit represent a topical hierarchy. An anecdote to topic divergence is the rise of the Internet-slang, *thread hijacking*, in which a group of users deviate so far off topic as to warrant the creation of an entirely new post.

Finally, we encourage readers to use the information presented in this paper to inform their future works. For example, the discussion threads and edit history of Wikipedia have been used in role-finding [8], quality assessment [9], content enhancement [10], and for dozens of other purposes. We believe that the comment threads from Reddit can serve a similar role by annotating its linked-content.

One important aspect of the Reddit site that we did not address in this paper is the topical differences among different subreddits. We believe that the language models in different subreddits can serve as background knowledge for clustering and labeling document clusters.

To help promote Reddit as a dataset for future work, the entire dataset, along with a queryable interface is available at http://web.engr.illinois.edu/~weninge1/reddit.

V. Acknowledgments

We thank Reddit for allowing us to crawl and curate their user data. The authors are not affiliated with Reddit in any way. This work is funded by an NSF and NDSEG Fellowship to the first author. The third author is supported by NSF, U.S. Air Force Office of Scientific Research, and by the U.S. Army Research Laboratory.

References

[1] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *SIGKDD*. ACM Press, Jun. 2009, p. 497.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.

[3] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies," *Journal of the ACM*, vol. 57, no. 2, pp. 1–30, Jan. 2010.

[4] V. Gómez, A. Kaltenbrunner, and V. López, "Statistical analysis of the social network and discussion threads in slashdot," in *WWW*. ACM Press, Apr. 2008, p. 645.

[5] G. Mishne and N. Glance, "Leave a Reply: An Analysis of Weblog Comments," in *WWE*, 2006.

[6] D. Laniado, R. Tasso, Y. Volkovich, and A. Kaltenbrunner, "When the Wikipedians talk: Network and tree structure of Wikipedia discussion pages," in *ICWSM*, 2011, pp. 177–184.

[7] A. Kaltenbrunner, V. Gómez, A. Moghnieh, R. Meza, J. Blat, and V. López, "Homogeneous temporal activity patterns in a large online communication space," *International Journal on WWW/INTERNET*, vol. 6, no. 1, Aug. 2008.

[8] H. T. Welser, D. Cosley, G. Kossinets, A. Lin, F. Dokshin, G. Gay, and M. Smith, "Finding social roles in Wikipedia," in *iConference*. ACM Press, Feb. 2011, pp. 122–129.

[9] A. Kittur and R. E. Kraut, "Harnessing the wisdom of crowds in wikipedia," in *CSCW*. ACM Press, Nov. 2008, p. 37.

[10] J. Schneider, A. Passant, and J. G. Breslin, "Understanding and improving Wikipedia article discussion spaces," in *SAC*. ACM Press, Mar. 2011, p. 808.