Will Too Many Editors Spoil The Tag? Conflicts and Alignment in Q&A Categorization

OMITTED FOR REVIEW

Q&A websites compile useful knowledge through user-generated questions and responses. Many Q&As use collaborative tagging systems to improve search and discovery while distributing the work of categorizing and organization throughout the community. Although early work on collaborative tagging questioned whether consistent categorization schemes could emerge from large groups with little to no coordination, empirical studies have found surprising coherence among users' tags. We build on this research by testing whether coherence emerges in tag usage on Q&As, a more challenging context, focusing in particular on mismatches in the specificity of tags (basic level disagreement). We found that some users shifted toward more specific tag usage over time slightly increasing conflict, but that moderators were instrumental in helping to resolve some of this conflict. This study highlights the importance of learning and moderation in the development of coherence in collaborative tagging systems.

CCS Concepts: •**Human-centered computing → Collaborative content creation; Empirical studies in collaborative and social computing;**

Additional Key Words and Phrases: collaborative tagging; categorization; moderation; distributed cognition; Q&As

ACM Reference format:
Omitted for review. 2016. Will Too Many Editors Spoil The Tag? Conflicts and Alignment in Q&A Categorization. 1, 1, Article 1 (January 2016), 19 pages.
DOI: 0000001.0000001

1 INTRODUCTION

Online question and answer communities (Q&A) have become important sites of knowledge sharing and creation. Knowledge seekers on these sites post questions to be seen by other community members, who reply in the form of comments and answers to provide relevant experience, specific solutions, and links to outside information and resources. The scope and topics of these Q&As range from everyday concerns to highly specialized and technical issues; these communities often generate solutions in a matter of minutes [27]. By asking and answering individual questions, these sites and their dedicated communities of experts have created rich content of lasting value [2, 32].

Successful Q&As have become repositories of knowledge with millions of questions and answers. Stack Overflow, a prominent example, has over 24 million answers to 16 million questions[1]. Tags are one way that users sift through this volume of information. When tags are integrated into platform tools, they can be used to filter and retrieve relevant content; on Stack Overflow tags can be used to set notifications, personalize the question feed appearing on the homepage and search

[1] https://stackexchange.com/sites?view=list#traffic

DOI: 0000001.0000001

for specific content. Tags are also used internally for the algorithms that generate lists of related questions[2].

Collaborative tagging systems allow sites to distribute work of categorization across the entire community. Collaborative tagging systems became popular in the early 2000s with the creation of social bookmarking sites, such as del.icio.us and CiteULike, introduced to label and organize content. These systems have now been adopted by many sites to help organize large online repositories (e.g. photos on flickr; questions on Stack Overflow). For example, on Stack Overflow question askers when posting a question are asked to apply one to five relevant tags to describe their question. Collaborative tagging systems empower users to actively participate in the organization of community content and distribute the effort of categorization content across many users. As a result more content can be labeled, making it (hopefully) easier to discover.

While tagging a document may seem relatively easy, prior work suggests that tagging actually involves a complex set of cognitive tasks [10]. When tagging is done with others in a collaborative tagging system it involves sharing and negotiating concepts across many different people. Researchers have described collaborative tagging both as a collective sensemaking [14] and an example of a distributed cognitive system [10]. In categorizing a document individuals may develop a better sense of how the information in a document is related to nearby knowledge [14, 36, 42]. By sharing tags individuals externalize and share their internal categories.

Conflict can arise in tag usage when individuals have different concepts and categories [14]. However conflicts can also be resolved organically if individuals adopt the most popular tags and learn from each other [10]. While prior work has focused on the organic resolution of conflicts, in this study we also consider the active intervention of moderators tasked with overseeing user behavior in these online communities.

To summarize, collaborative tagging has been adopted by some Q&As to make it easier to label, organize, and find content. This feature distributes the work of tagging, but in order for tags to be consistently applied it requires overcoming differences in categorization across individuals. In this paper we investigate the consistency of tag usage in 5 Q&A communities that use the Stack Exchange platform. In the first part of the paper we investigate whether users naturally tend toward more consistent usage over time as would be expected if users were assimilating conceptual structures [10]. In the second part of the paper we investigate whether moderators intervene to make tag usage more consistent. The Stack Exchange platform allows moderators to add, remove, and replace tags on questions. Researchers have found that content organization, which included adding and removing tags (as well as other actions like merging questions and linking between questions) was an important function that moderators performed [8].

The contribution of this paper is twofold. First, we replicate and extend prior work which has investigated whether tag usage stabilizes over time (e.g. [7, 14]) by studying collaborative tagging in a new context (i.e., Q&As) using a different method. In addition, we extend this prior work to consider how deliberate efforts on the part of moderators impact consistency in tag usage. Second, by investigating the role moderators play in "fixing" tag usage we add to a body of work on moderation in online peer-production communities. Previous work on moderation has focused on how moderators regulate user behavior (e.g. [13]) and ensure high quality work (e.g. [21, 22]); in this paper we investigate whether moderators help to align mismatches in representations across a community.

[2]The following pages describe relevant features on Stack Overflow, some methods were updated July 2018 after the collection period. However, methods remain similar in function: https://meta.stackexchange.com/questions/312180/favorite-tags-is-now-tag-watching, https://stackoverflow.com/help/searching, https://meta.stackexchange.com/questions/20473/how-are-related-questions-selected

This work highlights the challenges and opportunities for supporting complex, interdependent work involving sensemaking using a model of collaboration that involves a large number of users making contributions with little to no coordination.

2 RELATED WORK

2.1 Collaborative Tagging Systems

Collaborative tagging systems build on the long-time practice of using keywords to describe and label content, standard in most libraries and repositories [14, 26]. However, collaborative tagging systems democratize the role of categorization. In place of a central authority, collaborative tagging systems allow everyone in the community to generate tags and label documents. Across a variety of sites tags have been used to organize many types of entities including pictures (e.g., Flickr [29]), websites (e.g., del.icio.us [6, 14]), scholarly references (e.g., CiteULike [33]), and movies (e.g., Movie Lens [38]).

Users are motivated to use tags online for two dominant reasons: to provide organization and as a social activity [29]. On task oriented sites, tags serve organizational functions, such as retrieval, discovery, and sharing. On other sites, tags also serve social functions, such as play, competition, self-presentation, and opinion expression. Tags, incorporated as hashtags in social media, have also become novel symbols for communication and expression [20, 31].

Most prior work on collaborative tagging systems has focused on tags for organization. Creating and applying tags in order to organize content requires the cognitive process of categorization [10, 11]. By applying a tag to an entity a user is asserting that that entity is related to other entities with the same tag, that they belong in the same category. Human categorization is a basic cognitive process in which we group ideas and objects together based on shared characteristics [35]. Categories help us understand objects and ideas better. Prior work has examined, whether tags are applied accurately [39] and consistently [14, 16, 43]; and whether users are influenced by others' tags [11].

2.2 Conflicts in Categories and Tags

Collaborative tagging has the potential to surface conflict and divisions in understanding that exists between users. Researchers have argued that the collective aspect of these tagging systems may exacerbate existing problems in individual categorization [14]. Human categories have fuzzy boundaries [30], which means that non-prototypical objects may not fit cleanly into existing categories. In addition, some objects may belong in more than one category. Individuals' personal experience, expertise, and current knowledge structures affect how they form categories, resulting in differences in categorization across different users. Thus, a central concern of using collaborative tagging systems has been that "no coherent categorization scheme [will] emerge" because tagging is done by a large and open population of users whose work is only loosely coupled [16].

Golder and Hubberman [14] describe three major conflicts in tagging that might emerge.

- ***Polysemy***: Users might use the same tag to describe different categories. For example, "pepper" might mean "bell pepper" for one user and "black pepper" for another.
- ***Synonymy***: Users might apply different tags for the same concept. For example, "chilies" and "chili-peppers".
- ***Basic level variation***: Users might apply tags with different levels of specificity. For example, a question about very hot peppers might be tagged "chili-peppers" but also "carolina-reaper" (a particularly hot variety).

In the current study, we focus on conflict that arises from basic level variation. Early foundational work in cognitive science found that people tend to agree on what level of categorization is most

appropriate and informative; these are known as the "basic level" categories [35] (e.g. "cat"). However, later work showed that people with more domain expertise tend to use subordinate-level categories compared to people with less expertise; as people gain domain experience they develop more differentiated concepts [40] (e.g. a groomer might use "Persian cat").

2.3 Agreement in Categories and Tags

Though tagging conflicts are a real concern, researchers have also proposed several countervailing mechanisms which might drive community members towards more consistent usage. Two of the most important are imitation and learning.

Early models assumed that in tagging, as in language, individuals would imitate the way that others used a tag (e.g. [6, 11, 14]). Golder and Huberman [14] showed that tagging behavior on del.icio.us reflected what you would expect if users were influenced by what tags others had applied to a bookmark when they themselves tagged that same bookmark. They showed that behavior was consistent with the expectation of Polya's urn, a model which suggests that the likelihood of a user applying a tag is based on the existing distribution of tags. Similarly Cattuto, Lorento, and Pietronero [6] found that behavior on CiteULike could be explained by the Yule-Simon Model which assumes users are heavily influenced by currently used tags (i.e. "rich-get-richer"). In an experiment, Fu and colleagues [11] found that when individuals could not see what tags others used they tended to use different tags, but when they could see what tags others used they tended to use similar tags.

In addition to imitation, others have suggested that users may adopt similar tags due to learning (e.g. [10, 24]). Fu [10] describes the development of shared meaning through tags. He argues that collaborative tagging is an example of distributed cognition, in which cognitive processes are distributed across different users and across tools in the environment. Under his theoretical model, in the explore phase users seek out content on a site exploring and navigating using their own internal representations and the tags. Users start out without a fully formed understanding of the knowledge space. In the refine phase, users refine and enrich their understanding of the information based on the content that they have explored and the tags that they have used to navigate this content; in the process they update their internal representations. Thus, when users create new tags or use existing tags to label documents their choices reflect their internal representations as well as the knowledge they gained from others through using the external representations (i.e. tags). Fu's [10] theoretical model of iterative cycles of exploring and refining explain how the use of shared tags can lead to the assimilate conceptual structures across many people.

2.4 Stabilization of Tags

Empirical studies on tag usage have shown that communities stabilize on a consistent vocabulary of tags. The frequency of tags applied to the same bookmark stabilized after the first 100 users on del.icio.us [14]. Both [14] and [6] found tags followed a power law distribution, in which the most frequently used tags were used much more frequently than infrequently used tags, suggesting preference for some tags over others. They also showed that the statistical patterns of tag use followed what would be expected if individuals were imitating currently used tags. Ley and Seitlinger [24] found that the rate at which new tags were created declined over time among students using a social bookmarking system for class assignments, suggesting stabilization. They also found that students began using specific tags at higher rates, suggesting that as students developed more domain expertise they exhibited a basic level shift in categorization.

2.5 Moderation in Online Communities

Prior work has not examined the role of moderation on tagging behavior, despite the ubiquitous adoption of moderation in online communities. Governance–the policies, guidelines, and rules that dictate behavior on a site–is an important attribute of online communities [23, 34]. Moderators are users who are given special privileges in an online community and tasked with enforcing these rules, regulating user behavior, and generally promoting good content. Moderators help communities succeed by facilitating positive interactions between users, discouraging anti-social behavior, and setting standards for work quality [12, 13, 17]. Through editing, moderators improve content and encourage better quality work [2, 9]. Prior work has focused on the impact of moderation on user behavior, coordination of work, and work quality, less attention has been paid to the impact of moderation on aligning mental models and representations across a community.

3 CURRENT STUDY

In the current study we investigate how users apply tags to questions on five technical Q&A sites that are part of the Stack Exchange Network (SE) which also includes Stack Overflow. As task-oriented sites, tags on SE Q&As focus on describing the topic of a question. For technical Q&As these often include programming packages and languages (e.g., r, python, spss), commonly used methods and functions (e.g., machine-learning, hypothesis-testing, clustering), and technical terms and concepts (e.g., confidence-interval, sampling).

Q&As are important sites for knowledge sharing, knowledge production, and learning [2, 27, 32]. We focus on tagging behavior on Q&As for a few reasons. First, given the massive volume of content on Q&As, finding relevant knowledge is of paramount importance, and tagging is the main organizational tool on SE and other Q&As like Quora. On SE, a post's tags are used for core features including notification, filtering, and search. Second, Q&As are learning communities which means that there is mix of experts and novices. From prior work we know individuals at different levels of expertise may have different basic level categories [40], which may create more conflict in Q&As than on other sites.

Although, not unique to Q&As, there are two other advantages to studying tagging on SE. One, prior work has studied tag stabilization over time by examining changes in site level tag vocabulary (e.g., [24]) and/or changes in tags applied to a specific entity, such as a particular bookmark (e.g., [14]). The former approach does not examine whether users consistently apply the same tags to the same entity since it focuses on global usage. The latter approach does not examine whether users are able to apply the same tags to related but different entities, which is a harder task. Q&As allow us to study this more difficult problem by examining whether users apply the same tags to a set of questions on the same topic. Two, like many other online communities, SE Q&As have adopted a moderation system, so that users with sufficient experience are given the privilege of adding, deleting, and replacing tags assigned to a question. This allowed us to examine whether deliberate intervention by moderators improved tag consistency.

3.1 Tag Consistency

In order to conceptually replicate and extend prior work, such as [14], in this study we examined whether users were consistent in the tags that they applied to topically related, but different entities. More specifically, we focused on whether users consistently applied tags at the same level of specificity, since basic level category disagreements are known to occur between users of different levels of expertise and Q&As have a mix of experts and novices. Tags are consistent when different questions about a related topic (e.g., questions about Persian cats) are given the same tag, regardless of whether the preferred tag is more specific (e.g., 'persian-cat') or less ('cat').

We made two predictions, one based on a weaker interpretation of prior work and the other on a stronger interpretation.

For the first prediction, we assumed that there would be some initial conflict in tag usage, with some users preferring the general version of a tag and others using a more specific term. We expected that this level of conflict would get slighlty better or remain about the same over time because some users would continue with their preferred tag while others users might imitate the most popular version of the tag.

> *Prediction 1A: For questions on a particular topic Q&A users will tend toward slightly more consistent usage of tags over time.*

Our second prediction concerns the role that learning plays in reducing tag conflict, based on the idea that more knowledgeable users can better differentiate concepts and hence use more specific tags. Thus, we expected that tag use for a set of related questions would become more consistent and more specific over time (e.g., a specific version of a tag was used 40% of the time in year 1 and 70% of the time in year 7).

> *Prediction 1B: For questions on a particular topic Q&A users will tend toward using more specific versions of tags over time.*

3.2 Effect of Moderation on Tag Consistency

In the current study, we investigated the effect of moderation on consistency in tag usage. On SE Q&As moderators are instructed to fix tags for a question if they believe that the question is inappropriately tagged, however they are not specifically tasked with making tags consistent[3]. Nonetheless we predicted that moderators would make tags more consistent for a few reasons.

First, we expected that moderators would themselves be more consistent in how they used tags. Moderation privileges on SE Q&As (such as the ability to edit tags) are earned by asking and answering questions. Moderators also often read and review many questions. Therefore moderators can be expected to be more familiar with the site and have more domain expertise. As a group, since moderators are likely to share high levels of expertise, they can be expected to prefer more specific versions of tags as basic level categories. Thus, we predicted:

> *Prediction 2: Moderators on Q&A communities will be more consistent in their use of general and specific tag versions than regular users, preferring specific over general tags*

Second, we expected that moderators would try to make tags more consistent in order to standardize tag usage. Users discuss guidelines to make tag use more consistent on meta SE Q&As–companion Q&A sites frequented by experienced users and used to discuss the operation and governance SE sites. For example, on meta some users express the commonly held belief that existing tags should be used instead of creating new tags[4] and other users suggest ways to clean up related tags[5].

> *Prediction 3: Moderators on Q&A communities will intervene to shift tag usage toward more specific versions of tags.*

4 METHOD

4.1 Community Selection & Description

Stack Exchange Inc. (SE) runs a network of over 150 Q&A communities. Several of these communities have become important resources of knowledge in professional communities, including

[3] https://stackoverflow.com/help/tagging

[4] https://meta.stackexchange.com/questions/18878/how-do-i-correctly-tag-my-questions

[5] https://cooking.meta.stackexchange.com/questions/2079/tag-cleanup-chilli-chili-chiles-pepper-peppers-peppercorn

Table 1. Descriptive information of selected communities as of December 31, 2017

Community	Creation Date	Total Questions	Total Users	Selected Tag Pairs
Apple	Aug. 17, 2010	86,102	186,184	32
Statistics	July 19, 2010	113,913	146,004	128
Tex	July 26, 2010	146,868	120,411	85
User Experience	Aug. 9, 2010	24,950	82,469	5
Wordpress	Aug. 11, 2010	82,538	99,303	6

StackOverflow for software developers, Cross Validated for statisticians and data scientists, and MathOverflow for academic mathematicians [1, 41]. We chose SE for our research site because it is the largest Q&A platform in terms of monthly traffic[6], it employs collaborative tagging as the main tool for organization, many of the core functions (e.g., notifications, filtering, search) make use of tags, and it provides moderation privileges, such as adding, removing, and replacing tags, to users who demonstrate enough experience with the site.

In April of 2010 Stack Exchange changed the way new Q&As were created (known as Stack Exchange 2.0[7]). Prior to the change, individuals who wanted to create and operate a SE Q&A has to pay based on traffic to the site. After the change, creating Q&As was free and based on community involvement and interest. All Q&As 2.0 or later went through the same community building steps (e.g. discussion, proposal, commitment, beta). During this process the first set of tags were developed by the community. We choose 5 SE Q&As with the goal of showing generalizability across sites with different users, tags, and topical focus, while controlling for as much heterogeneity as possible (same technical platform, similar formation process, similar age, similar type of topic). For this reason, we selected sites that were created right after the switch to Stack Exchange 2.0 (similar formation process, similar age) and focused on technical topics, as defined as topics involving programming, software, or IT, because non-technical and technical Q&As have different user behavior [4]. Of the dozen or so sites that met these criteria we selected 5 Q&As on topics familiar to the authors of the paper, in order to ensure manual coding of tag pairs was accurate. The final five communities were: Apple, Stats (known as Cross Validated), Tex, User Experience (UX), and Wordpress (Table 1).

On SE, Q&As users are given moderation privileges when they demonstrate sufficient experience by accumulating 2,000 reputation points. Moderators can add, remove, or replace tags, and edit titles, questions and answers[8]. Reaching this level may require substantial effort; users earn reputation points when others in the community endorse their contributions as strong solutions (+10 points/upvote) or interesting questions (+5 points/vote)[9]. Some moderators take an even more active role in the governance of the site by participating in a site's meta Q&A, a site used to discuss guidelines, norms, and governance issues, including tags as well as many other issues.

As well as experience-based moderation, all communities also have a handful of administrator-level moderators who are either elected by the community or appointed by Stack Exchange site

[6]https://www.quantcast.com/top-sites/
[7]https://stackoverflow.blog/2010/04/13/changes-to-stack-exchange
[8]Any user can suggest edits, but they are not visible until reviewed and endorsed by users with enough reputation points (https://stackoverflow.com/help/privileges/edit).
[9]https://stackoverflow.com/help/whats-reputation

operators. In addition to moderation privileges, these administrators are expected to undertake an active role dealing with innapropriate content such as spam or low-quality responses[10].

4.2 Data Collection

Data was collected using Stack Exchange Data Explorer from the start of each of the 5 Q&As through 2017[11]. We first collected aggregate information about tags, such as their frequencies and co-occurrences on questions, in order to identify a set of tag pairs referring to the same concept at different levels of specificity. Next we collected specific tag usage data across a set of questions related to each tag pair.

We developed a method to identify pairs of tags from the five Q&As that represented hierarchically nested categories, that is tags that referred to the same concept, but one tag more specific than the other. Once identified we could measure whether users tended to use the more general or more specific version of the tag when labeling questions related to the concept. For example, a question about paired t-tests might be tagged with the tag "t-test", a more specific term, or the tag "hypothesis-test", a more general term. Either tag is an appropriate label for a question about paired t-tests, and which tag is applied will depend on what the user considers to be the basic level category.

Rather than exhaustively survey all tag pairs in search for all pairs that met our criteria we used the principle of purposeful sampling, a technique common in qualitative research, to focus on a large set of cases (i.e. tag pairs) which unquestionably met our criteria. To identify this set of tag pairs we went through two phases: first we automatically filtered the tag pairs to narrow the potential set of pairs into a manageable number and from the filtered results we manually coded the pairs to identify pairs that exhibited the desired relationship (i.e. referred to the same concept, one term more specific than the other).

Narrowing the set of tag pairs. In order to identify a set of potential tag pairs suitable for this study we gathered all potential tag pairs per Q&A and then performed the following three operations to narrow this pool. First we considered only tag pairs that were likely to be conceptually similar as measured by cosine similarity scores between tag co-occurrence vectors. Second, we looked at tags which had been demonstrated to be good substitutes for each other as indicated by the fact a moderator had swapped out one tag for the other at least 5 times. Finally, we limited our consideration to sufficiently common tags, those used at least 10 times. These filters restricted the intractably large set of all possible tag pairs to a set of plausible pairs, in the process we most likely excluded some relevant pairs, however it met our goal of purposefully sampling (see limitations). We explain the first step in more detail below.

In the first step above, we identified tags representing a similar concept by taking advantage of their patterns of association. Tags which are similar to each other ought to be included on questions with the same tags. For example, *diagrams* and *technical-drawing* are both used to label questions asking about the construction of visuals in latex, and both frequently co-occur with additional tags like *tikz-pgf*, *3D*, *engineering* on the Tex Q&A. Based on this observation we can assign each tag a co-occurence vector, in which each entry counts the number of questions in which two tags overlapped. We then compared these vectors using cosine similarity, a standard method for measuring similarity between vector representations [6, 15, 28, 37]. We limited our set of tag pairs to those that had similarity scores in the top 10% for the Q&A. This restricted our set of tag pairs to a manageable number of pairs, which were more likely to be conceptually similar.

[10]See https://stackoverflow.blog/2009/05/18/a-theory-of-moderation/ for more information on how moderation system works in Stack Exchange communities.
[11]https://data.stackexchange.com/

Selecting specific/general tag pairs. As a final step to ensure that each tag pair exhibited the desired relationship, the remaining tag pairs were manually coded. Initially, the first author coded whether a given pair represented the same concept and then whether it exhibited a specific/general relationship. The second author reviewed the codes and the two authors discussed the disagreements until they reached a consensus. This resulted in 256 tag pairs across the five communities. Each tag pair consists of two related concepts with one was more specific than the other. For example one tag pair identified in the Apple Q&A was (*macos*, *yosemite*). Both tags refer to the Macintosh operating system; *yosemite* is more specific because it refers to a particular version of the operating system, while *macos* is more general because it refers to multiple versions of these operating systems.

Once a set of tag pairs had been identified we extracted tag usage information for each tag, such as time of usage. More details are provided in the next section.

4.3 Statistical Analysis and Variables

The primary level of analysis used for our statistical models was at the tag pair level. Each statistical model included a repeated measures design. For Predictions 1A and 1B we examined tag usage within a tag pair over multiple time periods (Years). For Prediction 2 we examined tag usage within a tag pair across different types of users (User Type). For Prediction 3 we examined tag usage before and after moderation (Moderation Intervention). The following describe our independent variables:

- Tag Pair ID - For each repeated measure analysis we grouped tag usage by tag pair and included Tag Pair ID as a random effect. This allowed us to control for differences between tag pairs.
- Years - To measure change over time we grouped questions into half year increments. Time was included as a quantitative predictor scaled as a fraction of a year (e.g. 0.5 for first half year, 1 for second half year).
- User Type - We compared tag usage of question askers (general users) to three types of moderators those elected as administrators (admin mods), those that communicated using the formal communication back channel known as meta (meta mods), and others who had moderators privileges (general mods). In order to make edits to tags moderators had to earn 2000 reputation points.
- Moderator Intervention - We compared tag usage on questions before any moderation and to after moderation.

We examined the effect of time and moderation on two measures of consistency in tag usage included as dependent variables. Consistency Ratio was agnostic to whether users favored the general or specific tag version within a tag pair. Users were considered consistent within a tag pair and time period if they always used the same tag for a set of related questions. Users were considered inconsistent if they used a mix of specific and general versions of the tag. Specificity Ratio measured the degree to which users favored the specific tag version over the general tag version within a tag pair. To calculate each of these ratios for each tag pair we had to identify a set of related questions that either of these tags could be applied to. For each tag pair we selected the set of questions that included the text of the specific tag in the body of the question (e.g. t-test) and were tagged with at least one of the two tags (e.g. t-test, hypothesis-test), so that the questions were necessarily relevant to both tag versions. For each tag pair we calculated the number of questions that had the specific tag only, the general tag only, or both specific and general tags binned by the repeated measure (e.g. time period).

- Consistency Ratio - The degree to which users use the same tag for a set of related questions. Defined as the absolute difference in the number of related questions tagged with the specific tag versus the general tag divided by the number of related questions binned by the repeated

measure (e.g. time period): $\frac{|\#Specific_{ij} - \#General_{ij}|}{\#Total_{ij}}$ where i is the ith Tag Pair and j the jth repeated measure. We only considered related questions in which users used one of the two tags in the tag pair, since using only one of the tags is both partially consistent and partially inconsistent with using both tags. For example, for questions about Yosemite Macintosh operating system if users always only used the tag *Yosemite* the ratio would be 1, if users always only used the tag *MacOS* the ratio would be 1, and if they used only *Yosemite* 75% of the time and only *MacOS* 25% of the time (or the reverse) the ratio would be 0.5.

- Specificity Ratio - The degree to which users use the specific tag for a set of related questions. Defined as the total number of related questions tagged with the specific tag divided by the number of related questions binned by the repeated measure (e.g. time period): $\frac{\#Specific_{ij}}{\#Total_{ij}}$ where i is the ith Tag Pair and j is the jth repeated measure. We considered both related questions tagged with specific tag only or specific and general tags as using the specific tag, since in both cases users are labeling the question with the differentiated version of the concept. For example, for questions about Yosemite Macintosh operating system if users always included the tag *Yosemite* the ratio would be 1, if they always included only the tag *MacOS* the ratio would be 0.

Given the repeated measure design to test each prediction we constructed linear mixed-effects regression models using the *lmerTest* package in R [3, 19]. We included Tag Pair ID as a random intercept to control this dependency in the data. In addition, we examined the effect of including Tag Pair ID as a random slope for models examining consistency and specificity over time, to test the possibility that tag usage changed differently for different tag pairs. The random slope models outperformed the models without random slopes, suggesting differences across tag pairs. For each prediction we report the comparison of the model with random slopes. We also included Q&A ID as a potential fixed effect, but found no significant difference across Q&As.

5 RESULTS

5.1 Tag Consistency

We examined the degree to which users used tags consistently, whether users became more consistent over time, and whether users became more specific in their tag usage over time.

For every tag pair we examined the degree to which users always used the same tag for a set of related questions as measured by the consistency ratio. Table 3 summarizes the distribution of consistency ratios at the beginning and end of the 7.5 period lifespan of the Q&As. We found that users were often, but not universally consistent in which tag they used for a set of related questions. At the beginning of the Q&As tags were used consistently for 85% of the tag pairs, somewhat consistently for 6% of the tag pairs, and inconsistently for 9% of the tag pairs[12].

5.1.1 Consistency over time. Next, we examined how consistency in tag use changed over time by building a linear mixed effects model to evaluate the effect of years on consistency ratio for each tag pair. We considered the possibility that the community might become more consistent in their use of some tags. To test this possibility we evaluated a model with and without random slopes for tag pairs. We found that the model with random slopes (AIC = -722.76, BIC = -686.48) outperformed the model without random slopes (AIC = -489.71, BIC = -465.52) suggesting that the Q&A users changed consistency in tag usage over time in different ways for different tag pairs. Table 4 summarizes these differences, which we will return to after describing the main effect of the model.

[12]For convenience we refer to consistency ratios between 1-0.75 to be consistent, 0.75-0.5 to be somewhat consistent, and less than 0.5 to be inconsistent.

Table 2. The results of the Linear Mixed Effects models tracking changes in consistency and specificity over time

	Dependent variable:	
	Consistency Ratio	Specificity Ratio
	Coef. (SE)	Coef. (SE)
Intercept	0.86^{***} (0.02)	0.41^{***} (0.03)
Years	-0.02^{***} (0.01)	0.05^{***} (0.01)
# Observations	3,720	3,720
# Tag Pairs	248	248
$R^2_{conditional}$	0.56	0.86
$R^2_{marginal}$	0.004	0.02
Note:		*p<0.1; **p<0.05; ***p<0.01

We expected that over time users would imitate the most popular tags for a given concept and as a result tag consistency would get slightly better or at least stay the same (prediction 1A). We found that the main results did not support this prediction. Instead we found that tag usage for a set of related questions became less consistent over time (Coef. = -0.02, $p < 0.01$; See Table 2). Which means that on average users were more likely to use the same tag for a set of related questions at the beginning of the Q&A, than they were 7.5 years later. However, the degree to which the consistency declined over time was relatively small. According to the model, consistency of tag usage declined by around 2% per year during the 7.5 years the Q&A sites were active. At the beginning 85% of tag pairs were consistent with a ratio between 0.75 and 1, by the end of the 7.5 years 74% were consistent.

The average decline in consistency in tag usage was driven by a small percentage of the tag pairs. Nakagawa and Schielzeth's[13] R^2 suggested that the main effect of time explained only 0.4% of variance in consistency ratios, whereas difference across tag pairs explained 56.2% of the variance in consistency ratios. Table 4 shows that the largest decrease in tag consistency occurred for 11% of the tag pairs (28 out of 248). Many tag pairs exhibited only modest change in consistency over time and 2% of tag pairs increased in tag consistency.

5.1.2 Specificity over time. We examined how specificity in tag usage changed over time by building a linear mixed effects model to evaluate the effect of years on specificity ratio for each tag pair. As before, we found that a model with random slopes outperformed a model without random slopes, suggesting that users changed specificity in tag usage over time in different ways for different tag pairs. Table 4 shows that specificity increased for 76% of the tag pairs.

We expected that as the community and its members deepened their understanding of the material and learned from each other that they would develop more differentiated concepts and as a result shift toward using more specific tags for similar questions (prediction 1B). We found that the main results did support this prediction. As predicted we found that for a related set of questions users used the specific version of a tag at greater frequencies over time (Coef. = 0.05, $p < 0.01$; See Table 2). On average there was a 5% increases in specificity per year. At the beginning

[13]Nakagawa and Schielzeth's R^2 is a commonly used pseudo R^2 for mixed effects models

Table 3. Frequencies of consistency and specificity ratios for the different tag pairs

Consistency ratio	0-0.25 (%)	0.25-0.50 (%)	0.50-0.75 (%)	0.75-1 (%)	Total
# pairs at the start	11 (4)	11 (4)	15 (6)	211 (85)	248
# pairs at the end	18 (7)	24 (10)	23 (9)	181 (74)	246 [a]
Specificity ratio	0-0.25 (%)	0.25-0.50 (%)	0.50-0.75 (%)	0.75-1 (%)	Total
# pairs at the start	135 (54)	13 (5)	11 (4)	89 (36)	248
# pairs at the end	73 (30)	17 (7)	23 (9)	133 (54)	246

[a]Two tag pairs were used only once as original tags during 7 years hence excluded

Table 4. The distribution of rate of change over time (i.e. slopes) across the different tag pairs

Consistency ratio (slopes)	-0.21 to -0.11	-0.11 to 0	0 to 0.11	0.11 to 0.21	Total
# tag pairs	28	99	117	4	248
Specificity ratio (slopes)	-0.44 to -0.22	-0.22 to 0	0 to 0.22	0.22 to 0.44	Total
# tag pairs	1	59	163	25	248

only 36% of tag pairs used the specific version of the tag at high rates, by the end of the 7.5 years 54% of the tag pairs used the specific version of the tag at high rates (see Table 3).

The rise of specific tags may help to explain our unexpected result for Prediction 1A. At the beginning general tags were used more frequently, as some users shifted toward specific tags, there was a greater mix of general and specific tags creating more conflict. If the shift toward specific tags continues, over a longer time period we might observe improvement in tag consistency as specific tags are used at higher rates than general tags.

5.2 Effect of Moderation on Tag Consistency

5.2.1 Consistency across different users. We examined whether moderators were consistent and/or specific in their usage of general and specific variants of tags. We gathered sets of related questions for each tag pair. For each set of related questions (and tag pair) we recorded tag usage for original question authors (general users) and for any edits applied (grouped by type of moderator). These records were used to calculate consistency and specificity ratio per tag pair and user type and entered into two mixed effects regression models with user type as the independent variable and consistency or specificity as the dependent variable. Tag pair ID was included as a random effect.

We expected that moderators would be more consistent in their tag usage than general users. The results showed that certain moderator groups did display more consistent usage than general users (Table 5). General mods and admin mods were significantly more consistent in their use of tags than general users (Coef. = 0.15, $p < 0.001$; Coef. = 0.12, $p < 0.01$ respectively), whereas there was no statistically significant difference in the tag usage between meta mods and general users (Coef. = 0.04, p = 0.11). On average general mods favored one version of the tag to the other 11:1, admin mods favored one version 7:1, meta mods favored one version 7:1 and general users favored one version 6:1. We expected that moderators might be more consistent than general users because they

were more experienced and and involved in governance discussions. This prediction was mostly supported by the results in that we observed higher consistency ratios for all three moderator groups which reached statistical significance for all but one of these groups. The difference in consistency between meta mods and general users did not reach statistical significance, which was surprising because meta mods are mods most active in discussions of governance. This suggests that most gains in consistency are probably due to experience and not because of agreements reached in discussion.

We also expected that moderators would be more specific in their usage of tags. One of the primary reasons moderators were expected to be more consistent in their usage of general and specific tag variants is because they are more experienced. Not only are moderators required to have experience creating good question and answers, moderators also build up experience through reading and reviewing many questions. In gaining experience moderators may develop more differentiated concepts and would be predicted to favor more specific tag variants over general tag variants. We examined the degree to which moderators favored specific tag variants compared to general users (Table 5). All moderators used specific tag variants significantly more than original askers (General mods: Coef. = 0.14, $p < 0.01$; Meta mods: Coef. = 0.12, $p < 0.01$; Admin mods: Coef. = 0.13, $p < 0.01$).

Table 5. The results of Linear Mixed Effects models comparing consistency and specificity between moderators and general users

	Dependent variable:	
	Consistency Ratio	Specificity Ratio
	Coef. (SE)	Coef. (SE)
General users (intercept)	0.72^{***} (0.02)	0.56^{***} (0.02)
General mods	0.15^{***} (0.03)	0.14^{***} (0.03)
Meta mods	0.04 (0.03)	0.12^{***} (0.03)
Admin mods	0.12^{***} (0.03)	0.13^{***} (0.03)
# Observations	725	725
# Tag Pairs	254	254
$R^2_{conditional}$	0.20	0.56
$R^2_{marginal}$	0.04	0.03
Note:		*p<0.1; **p<0.05; ***p<0.01

5.2.2 Impact of moderation on consistency. Next we examined the impact of moderators' actions on the collection of questions as a whole. When moderators change the tags associated with questions, they change the consistency of tag usage across the collection of questions. We considered the degree to which tags were used consistently before any moderator edits to the degree to which tags were used consistently after all moderator edits. We predicted that moderators would improve the consistency of tag usage. We gathered sets of related questions for each tag pair. We found that moderators edited 17% of questions under study. For each set of related questions (and tag pair) we recorded tag usage as originally applied to a question (before moderation) and tag usage after all edits had been made (after moderation) and calculated consistency and specificity ratio for these two groups. Then we entered these numbers into two mixed effects regression models with

moderation intervention as the independent variable and consistency or specificity ratio as the dependent variable. Tag pair ID was included as a random effect.

The results showed that there was a small but significant increase in tag consistency after moderation (Coef. = 0.09, $p < 0.01$). After moderation the favored variant of the tag was used 10:1, whereas before moderation the favored variant of the tag was used only 6:1. In other words before moderation on average we observed a consistency ratio of 0.72 and after moderation the consistency ratio rose to 0.81. In total 3% of the variance in consistency across tag pairs could be explained by the intervention of moderators.

Not only did we predict moderators would make tag usage more consistent, we also predicted that they would shift tag usage toward specific versions of tags (Table 6). The results showed that there was more use of specific tag variants compared to general tag variants after moderation than before moderation (Coef. = 0.06, $p < 0.01$). Before moderation on average 56% of the questions were tagged with the specific variant, after moderation this rose to 62%.

Table 6. The results of Linear Mixed Effects models comparing consistency and specificity of tags before and after moderation

	Dependent variable:	
	Consistency Ratio	Specificity Ratio
	(1)	(2)
Intercept	0.72^{***} (0.02)	0.56^{***} (0.03)
Moderation	0.09^{***} (0.02)	0.06^{***} (0.02)
Observations	487	487
Tag Pairs	248	248
$R^2_{conditional}$	0.57	0.83
$R^2_{marginal}$	0.03	0.01
Note:	*p<0.1; **p<0.05; ***p<0.01	

6 DISCUSSION

In this study, we investigated the degree to which SE Q&A users were consistent in how they used tags, and the degree to which moderation helped to improve consistency. We analyzed basic level disagreements by focusing on specific/general tag pairs from 5 Q&A communities over the span of 7.5 years. The most important findings of this study are: 1) we found mild basic level disagreements that resulted in low level inconsistency in tag usage, 2) moderation helped to reduce inconsistency in tag usage, and 3) in contrast to prior work, tag usage became less consistent over time.

An optimistic perspective on these results suggest that the overall level of conflict in basic level categorization observed on these Q&As was minimal. On average there was 72% consistency before moderation. Some level of inconsistency is inevitable, and the fact that initial consistency was so good suggests high levels of pre-existing agreement in basic level categories; this could be due to an initial shared understanding of categories and/or assimilation of categories and labels through imitation and learning [10, 14]. Nonetheless, Q&As may benefit considerably from reducing even low levels of inconsistency. Q&As now serve as large repositories of knowledge that are important resources for people in technical fields like software engineering [2, 27]. Tags can help as features

in tools to filter and find relevant content [29], and discovering exactly the right question can save users valuable time and effort.

The inconsistencies between general users were reduced by moderation. Moderation improved consistency, raising the average consistency score 9% by smoothing out some disagreements in basic level categories and creating more uniform representations. This highlights the important role moderation can play in knowledge production communities. At a minimum moderation serves to 'mop up' inconsistencies. Moderators perform necessary chores by fixing minor errors in organization, similar to menial chores performed by editors on Wikipedia [5]. More substantially, moderators may help by imposing standardization that is difficult to achieve with many people and little to no coordination. Having one person provide a high level structure helps a crowd produce artifacts that benefit from a standardized vision [18, 25]. This second role may be particularly important in helping to create coherence. By making tags more consistent, moderators may be pushing the community toward a shared understanding of the topic. And because moderation is applied after content is initially tagged, it does not detract from the benefits of decentralization. SE Q&As are rare in allowing and encouraging moderators to edit collaboratively generated tags; other sites and online repositories might benefit from adopting this form of moderation.

Unlike prior work we found that consistency in tag usage got worse over time rather than remaining the same or improving (e.g. [14]). Collaborative tagging may be more challenging in contexts in which it is more difficult to develop a universal categorization scheme, such as Q&As. Early work on collaborative tagging examined tagging on sites, like del.icio.us, CiteULike, and flickr (e.g. [6, 14]). There are a few differences between technical Q&As examined in this study and other sites. First, Q&As are learning environments in which users ask questions to better understand concepts and the material, which means that users tag questions before their internal representations have stabilized. Second, the Q&As studied here focus on technical topics, so that understanding and correctly tagging a question often requires highly specialized knowledge. Together these two characteristics help to explain why tagging starts out with more general tags that shift toward more specific tags over time. Tag consistency may be a bigger problem for sites that use tags to organize highly specialized knowledge and are made up of a community of novice and expert users. Thus, collaborative tagging may work better without intervention when it is used to organize popular content (e.g., images, websites), than when it is used to organize specialized content (e.g., tax, medical, legal information and resources) if these specialized sites attract a broad cross section of people, including those trying to learn more about this specialized content.

In addition, previous work examined easier tagging problems. For example, Golder and Huberman [14] examined the degree to which users were consistent in how they applied tags to the same bookmark over time. For a categorization scheme to be effective it must be applied consistently across related objects as well as the same object. In this study we test the harder case and find that users have some difficulty staying consistent in their categorization across different, but related objects.

6.1 Design Implications

The primary implication of this study is that more sites should allow and encourage moderators to edit community generated tags. We found that moderators were able to improve tag consistency across related questions by making lasting edits to some questions. Online communities have mostly encouraged moderators to take on roles as gate keepers and enforcers. There may be benefits to emphasizing the role moderators can play at standardizing content across a site. Few sites that use collaborative tagging allow other users to edit tags once applied. Even sites in which tagging is primarily personal and individualized might benefit from this type of moderation. For example,

multiple disparate conversations are created when users inadvertently use two different versions of a hashtag to refer to the same topic on Twitter or Instagram.

There are also design choices and automatic tools sites could adopt to help a community resolve tag conflicts organically. In implementing a collaborative tagging system site designers chose whether or not to restrict the number of tags applied to an entity. For example, YouTube allows unlimited number of tags, while Stack Overflow restricts users to 5 tags. Restricting the number of tags may create more tagging inconsistencies. If given the option to apply many tags, users may apply tags at multiple levels of specificity (e.g. 'paired','t-test','hypothesis test'). If restricted they will have to choose a particular level of specificity and conflict may arise when different users chose different levels of specificity (e.g., 't-test' vs. 'hypothesis test'). Thus, many sites may benefit from not restricting the number of tags.

Sites may benefit from having users build explicit models of relations between tags. For example, by creating a taxonomic tree between hierarchically nested tags, such that a more specific tag like the R package 'ggplot2' would be associated as a child of the more general tag 'R'. Within the taxonomy of tags basic level tags could be explicitely declared (e.g. 'ggplot2') which could help users quickly learn which tag to apply. In addition, the taxonomy could be used for other functions that made use of tags, such as search. Thus, a user searching using a too general tag like 'R' could still find questions that were tagged with more specific tags, such as a package name 'ggplot2'. By explicitely building and modeling hierarchical relations between tags site designers could help resolve some inefficiencies associated with basic level variation.

7 LIMITATION AND FUTURE WORK

There were several limitations of this study. First, we examined only one type of conflict in tagging, basic level disagreement, on one platform, Stack Exchange. Basic level disagreement is one important type of conflict, but there are many others (e.g. polysemy, synonyms) that result from a large community of contributors who have different mental models. We expect that some of our findings, such as the increase in conflict associated with a rise in specific tags, and the value of moderators in fixing conflict will generalize to other Q&A platforms and other sites with highly specialized content that requires deep expertise (e.g., online sites with tax, legal, and medical information and resources). However, we also expect that specific features of a site may have a large impact on the degree of conflict. For example, we argued above that restricting the number of tags may increase basic level disagreement. Future research, should investigate basic level disagreements in tags on a variety of sites with a range of different features.

Second, we observed very small effects and a large amount of variance across tag pairs. Future work is needed to investigate why there is more conflict for some tag pairs and why conflict gets worse for some but not all tag pairs. One explanation, is that some concepts are harder to understand and require more expertise; we may see more conflict about these concepts. Another explanation is that external factors, such as dynamic changes in user populations (and their expertise) also may influence tag pairs differentially. Communities sometimes attract users with much more expertise in some areas than others and the make up of the community can shift over time. We used the principle of purposeful sampling to select a set of tag pairs that met our criteria, in doing we probably missed some relevant tag pairs and some of these tag pairs may exhibit different patterns over time.

Third, we used a relatively crude and simplistic approach to identifying questions that we believed should be similarly tagged. This approach obscures a great deal of nuance in questions. It also cannot rule out the possibility that the observed changes over time are due to changes in the nature of the question, such as questions getting more specific over time. We argue that it is important to

evaluate consistency in tag use across multiple entities rather than simply looking at tags applied to the same entity by multiple users. However, there may be better approaches to identifying questions that should have the same tag.

8 CONCLUSION

Q&A websites compile useful knowledge through user-generated questions and answers. Collaborative tagging has been adopted by some Q&As to make it easier to label, organize, and find content. We investigated whether users naturally tended toward more consistent tag usage over time as would be expected if users were assimilating conceptual structures [10]. We found low level inconsistency in tag usage that persisted and got slightly worse over time as some users shifted toward more specific tags. We found that moderation helped to reduce inconsistencies. We argue that there may be widespread benefit to allowing moderators to edit tags.

ACKNOWLEDGMENTS

Omitted for review

REFERENCES

[1] Rabe Abdalkareem, Emad Shihab, and Juergen Rilling. 2017. What do developers use the crowd for? a study using Stack Overflow. *IEEE Software* 34, 2 (2017), 53–60.

[2] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2012. Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*. ACM, New York, NY, USA, 850–858. http://doi.acm.org/10.1145/2339530.2339665

[3] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. , 48 pages. http://dx.doi.org/10.18637/jss.v067.i01

[4] Keith Burghardt, Emanuel F Alsina, Michelle Girvan, William Rand, and Kristina Lerman. 2017. The myopia of crowds: Cognitive load and collective evaluation of answers on Stack Exchange. *PLoS One* 12, 3 (March 2017), e0173610. http://dx.doi.org/10.1371/journal.pone.0173610

[5] Moira Burke and Robert Kraut. 2008. Mopping Up: Modeling Wikipedia Promotion Decisions. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW '08)*. ACM, New York, NY, USA, 27–36. http://doi.acm.org/10.1145/1460563.1460571

[6] Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. 2008. Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In *Proceedings of the 2008 International Semantic Web Conference (ISWC)*. Springer Berlin Heidelberg, 615–631. http://dx.doi.org/10.1007/978-3-540-88564-1_39

[7] Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. 2007. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences* 104, 5 (Jan. 2007), 1461–1464. http://dx.doi.org/10.1073/pnas.0610487104

[8] Joohee Choi and Yla Tausczik. 2017. Content management through distributed moderation. (2017).

[9] Dan Cosley, Dan Frankowski Sara, Sara Kiesler, Loren Terveen, and John Riedl. 2005. How Oversight Improves Member-Maintained Communities. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, New York, NY, USA, 11–20. http://doi.acm.org/10.1145/1054972.1054975

[10] Wai-Tat Fu. 2008. The microstructures of social tagging: a rational model. In *Proceedings of the 2008 ACM conference on Computer Supported Cooperative Work*. ACM, New York, NY, USA, 229–238. http://doi.acm.org/10.1145/1460563.1460600

[11] Wai-Tat Fu, Thomas Kannampallil, Ruogu Kang, and Jibo He. 2010. Semantic Imitation in Social Tagging. *ACM Transactions on Computer-Human Interaction* 17, 3 (July 2010), 12:1–12:37. http://doi.acm.org/10.1145/1806923.1806926

[12] Joaquín Gairín-Sallán, David Rodríguez-Gómez, and Carme Armengol-Asparó. 2010. Who Exactly is the Moderator? A Consideration of Online Knowledge Management Network Moderation in Educational Organisations. *Computers & Education* 55, 1 (Aug. 2010), 304–312. http://dx.doi.org/10.1016/j.compedu.2010.01.016

[13] R Stuart Geiger and David Ribes. 2010. The Work of Sustaining Order in Wikipedia: The Banning of a Vandal. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10)*. ACM, New York, NY, USA, 117–126. http://doi.acm.org/10.1145/1718918.1718941

[14] Scott A Golder and Bernardo A Huberman. 2006. Usage patterns of collaborative tagging systems. *Journal of information science* 32, 2 (April 2006), 198–208. https://doi.org/10.1177/0165551506062337

[15] Wael H Gomaa and Aly A Fahmy. 2013. A survey of text similarity approaches. *International Journal of Computer Applications* 68, 13 (2013), 13–18.

[16] Harry Halpin, Valentin Robu, and Hana Shepherd. 2007. The Complex Dynamics of Collaborative Tagging. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*. ACM, New York, NY, USA, 211–220. http://doi.acm.org/10.1145/1242572.1242602

[17] Andrea Kienle and Carsten Ritterskamp. 2007. Facilitating asynchronous discussions in learning communities: the impact of moderation strategies. *Behaviour & Information Technology* 26, 1 (Jan. 2007), 73–80. https://doi.org/10.1080/01449290600811594

[18] Joy Kim, Justin Cheng, and Michael S Bernstein. 2014. Ensemble: exploring complementary strengths of leaders and crowds in creative collaboration. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 745–755. http://doi.acm.org/10.1145/2531602.2531638

[19] Alexandra Kuznetsova, Per Brockhoff, and Rune Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82, 13 (2017), 1–26. https://www.jstatsoft.org/v082/i13

[20] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web*. AcM, 591–600.

[21] Cliff Lampe and Paul Resnick. 2004. Slash(Dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 543–550. http://doi.acm.org/10.1145/985692.985761

[22] Cliff A C Lampe, Erik Johnston, and Paul Resnick. 2007. Follow the Reader: Filtering Comments on Slashdot. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 1253–1262. http://doi.acm.org/10.1145/1240624.1240815

[23] Jonathan Lazar and Jennifer Preece. 2002. Social considerations in online communities: Usability, sociability, and success factors. In *Cognition in a Digital World*, H van Oostendorp (Ed.). L. Erlbaum Associates Inc, Hillsdale, NJ, USA.

[24] Tobias Ley and Paul Seitlinger. 2015. Dynamics of human categorization in a collaborative tagging system: How social processes of semantic stabilization shape individual sensemaking. *Computers in Human Behavior* 51, A (Oct. 2015), 140–151. http://dx.doi.org/10.1016/j.chb.2015.04.053

[25] Kurt Luther, Casey Fiesler, and Amy Bruckman. 2013. Redistributing Leadership in Online Creative Collaboration. In *Proceedings of the 2013 ACM Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 1007–1022. http://doi.acm.org/10.1145/2441776.2441891

[26] George Macgregor and Emma McCulloch. 2006. Collaborative tagging as a knowledge organisation and resource discovery tool. *Library review* 55, 5 (2006), 291–300.

[27] Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. 2011. Design Lessons from the Fastest Q&a Site in the West. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 2857–2866. http://doi.acm.org/10.1145/1978942.1979366

[28] Benjamin Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, and Gerd Stumme. 2009. Evaluating Similarity Measures for Emergent Semantics of Social Tagging. In *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*. ACM, New York, NY, USA, 641–650. http://doi.acm.org/10.1145/1526709.1526796

[29] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. 2006. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia*. ACM, 31–40.

[30] Michael E McCloskey and Sam Glucksberg. 1978. Natural categories: Well defined or fuzzy sets? *Memory & Cognition* 6, 4 (July 1978), 462–472. https://doi.org/10.3758/BF03197480

[31] Changhoon Oh, Taeyoung Lee, Yoojung Kim, SoHyun Park, and Bongwon Suh. 2016. Understanding Participatory Hashtag Practices on Instagram: A Case Study of Weekend Hashtag Project. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1280–1287.

[32] Chris Parnin, Christoph Treude, Lars Grammel, and Margaret-Anne Storey. 2012. *Crowd documentation: Exploring the coverage and the dynamics of API discussions on Stack Overflow*. Technical Report.

[33] Denis Parra and Peter Brusilovsky. 2009. Collaborative filtering for social tagging systems: an experiment with CiteULike. In *Proceedings of the third ACM conference on Recommender systems*. ACM, 237–240.

[34] Jenny Preece. 2000. *Online Communities: Designing Usability and Supporting Sociability* (1st ed.). John Wiley & Sons, Inc., New York, NY, USA.

[35] Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology* 8, 3 (July 1976), 382–439. https://doi.org/10.1016/0010-0285(76)90013-X

[36] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The Cost Structure of Sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems (CHI '93)*. ACM, New York, NY, USA, 269–276. http://doi.acm.org/10.1145/169059.169209

[37] Hinrich Schütze and Jan O Pedersen. 1997. A Cooccurrence-based Thesaurus and Two Applications to Information Retrieval. *Information Processing & Management* 33, 3 (May 1997), 307–318. http://dx.doi.org/10.1016/S0306-4573(96)

00068-4

[38] Shilad Sen, Jesse Vig, and John Riedl. 2009. Tagommenders: connecting users to items through tags. In *Proceedings of the 18th international conference on World wide web*. ACM, 671–680.

[39] Fabian M Suchanek, Milan Vojnovic, and Dinan Gunawardena. 2008. Social tags: meaning and suggestions. In *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 223–232.

[40] James W Tanaka and Marjorie Taylor. 1991. Object Categories and Expertise: Is the Basic Level in the Eye of the Beholder? *Cognitive Psychology* 23, 3 (1991), 457–482. https://doi.org/10.1016/0010-0285(91)90016-H

[41] Yla R Tausczik, Aniket Kittur, and Robert E Kraut. 2014. Collaborative problem solving: A study of mathoverflow. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 355–367.

[42] Karl E Weick, Kathleen M Sutcliffe, and David Obstfeld. 2005. Organizing and the Process of Sensemaking. *Organization Science* 16, 4 (Aug. 2005), 409–421. https://doi.org/10.1287/orsc.1050.0133

[43] Robert Wetzker, Carsten Zimmermann, Christian Bauckhage, and Sahin Albayrak. 2010. I tag, you tag: translating tags for advanced user models. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 71–80.

Received February 2007; revised March 2009; accepted June 2009