Jared Lorince
Indiana University
Department of Psychology,
1101 E. 10th Street
Bloomington IN 47401

jlorince@indiana.edu

Date: 17 March, 2015

Editor
The Journal of Web Science

Dear Dr. Wolfgang Nejdl:

Please consider our revised manuscript entitled "The Wisdom of the Few? “Supertaggers” in Collaborative Tagging Systems" (original submission #12) for publication in *The Journal of Web Science*.

We thank the reviewers for their comments and suggestions and are very grateful for the amount of thought and effort that has been put into improving the quality of our manuscript. Please find an itemized list of our responses to the comments and suggestions below (the reviewer’s comments appear in regular font, and our replies appear in bold font).

We thank you for inviting a revision of this manuscript for further consideration for publication.

Sincerely,

Jared Lorince (Department of Psychological & Brain Sciences, Indiana University)
Sam Zorowitz (Division of Neurotherapeutics, Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School)
Jaimie Murdock (School of Informatics & Computing, Indiana University)
Peter M. Todd (Department of Psychological & Brain Sciences, Indiana University)

Reviewer B:

This article investigates whether supertaggers, i.e., those who tag more objects than other “normal” users, according to authors’ criteria, have quantitatively different behaviors when compared to their counterparts. The authors test their hypotheses in three different large datasets.

The article is in general well written and motivated. However there are many typos and grammatical errors throughout the text. These should be fixed. (In fact, running a spellchecker before submission would help a lot on this matter).

We have extensively edited and proofread to address the typos and other errors.

More importantly, the authors should make clearer what are the differences between this work and their previous papers. This is not clear at all in the article. A clear differentiation is necessary mainly in the introduction and/or in the related work.

A footnote in Section 1 now explicitly describes what has been added to this new version of the paper.

Regarding the statistics on the datasets the authors should check whether their numbers are consistent with other characterizations of similar tag-based datasets, such as the one presented in [1]. This is important since there is a lot of argumentation defending the way the data was gathered.

This is a reasonable point on which we now comment (Section 3.1), and have also included the suggested reference.

Regarding their definition of supertaggers, as the authors themselves discuss, any definition may be considered as arbitrary. And this is the case in this article. Their definition seems very arbitrary. They really should make a harder effort to check the impact of their threshold choice in some of their analyses, even if in a few ones and in a qualitative way.

We agree that this is an important point, but after substantial consideration have elected to maintain the threshold used in the paper. It is unclear what precisely would be gained by examining any other particular threshold, and our goal was mainly to use one that effectively segregated the most prolific taggers from the other, less prolific ones. We have added text (last paragraph of Section 4) to further explain this threshold choice:

"We reiterate that the particular threshold used here is arbitrary, in that there is no special behavioral shift that occurs at this point. In fact, various behavioral measures show relatively smooth changes as we consider users with progressively more annotations, and as such, we present measures as a function of users’ total annotation counts (rather than simply comparing averages for supertaggers and non-supertaggers) wherever possible. Nonetheless, in those analyses that directly compare the annotations of the two groups we have defined, the 50 percent split is both clearly interpretable (in that it compares “typical” users to the most prolific ones) and analytically convenient, as it normalizes all analyses of the sub-folksonomies such that the total number of annotations in each is constant."

They should also try to be a little less quantitative and try to conjecture the reasons for some of the identified behaviors. Some effort towards this goal has been done in the discussion in the conclusion, but the authors should try to raise some hypotheses earlier while analyzing the data, for instance, in the end of Section 5.2.

Our focus in the paper is more on identifying what differences exist between supertaggers and non-supertaggers than on determining why these differences exist, something we have clarified in the introduction. We have, however, attempted to expand on the speculation for *why* such differences exist, particularly in the conclusion.

There are also some other arbitrary decisions such as using only the tags with 10000 or more total annotations to calculate SPEAR. These choices should be better justified.

We have adjusted this threshold such that we analyze the top 10,000 most popular tags overall that have at least 10 unique users, and included a more detailed justification for the threshold used (Section 6.2.1).

There are also some conjectures that may contradict each other. For instance, in the end of Section 5.1 the authors say that supertaggers deviate from the "normal" users with respect to the most idiosyncratic and singleton tags, but latter they propose a metric of expertise based on how much supertaggers agree with the "wisdom of the crowd". This needs a better explanation.

We acknowledge that this is somewhat confusing. Although we believe this issue is partly addressed ("In cases where a user has assigned multiple tags to the same item, we only include the single highest expertise score for that item in the user's mean score. In this way, we capture whether or not the user knows the 'best' tag for an item, without penalizing her if she additionally assigns other tags to it."), we have expanded on this paragraph to clarify the measure: "This, combined with the low weighting of items tagged only a few times, sidesteps the issue that supertaggers tend to use many idiosyncratic tags (which necessarily are not 'expert' tags under any consensus-based measure). In effect, this analysis allows us to determine whether supertaggers show expertise with relatively popular tags, while not considering their usage of idiosyncratic tags." We additionally ran exploratory analyses (mentioned in a footnote) demonstrating that qualitative results held even when excluding idiosyncratic tags from the analysis.

Finally, the authors comment that tagging consensus may have been achieved because supertaggers may be the first to tag new objects. However such interesting analysis is not present in the paper. It is not clear why it was not done, since they have information about the time and who tagged the objects.

We ran such an analysis, which did in fact show that more prolific taggers tend to tag items earlier than other users. However, we do not include the results explicitly in the paper because the SPEAR results imply exactly this (because SPEAR rewards users for being among the first to tag an item). We do, however, add a footnote addressing this in the conclusion: "This is confirmed directly in the case of Delicious by supplemental analyses (not presented here) examining only the average point at which supertaggers annotate items relative to other users. As expected, more prolific taggers do tend to tag items earlier than others. In the case of Last.fm, however, the low temporal resolution of our data makes the results of such an analysis unreliable and difficult to interpret."

Regarding article organization, the authors should make an effort to approximate the figures to the point where they are mentioned by the first time. Some figures are two pages away from the first mention.

As much as possible, we have edited the .tex to accomplish this.

Reviewer C:

The present work is well written and methodologically as well as statistically very sound. By operationalizing and analyzing the tagger-supertagger dichotomy and by relating it to already established variables, the work continues existing research on collaborative tagging and helps refining our understanding of constructs, such as the describer-categorizer distinction. Beyond that, the results provide practical insights into cognitive and behavioral consequences of interactionparadigms in different tagging systems (Delicious, Flickr and LastFM).

We thank the reviewer for the positive remarks here and below.

However, due to – in my eyes – inconclusive arguments and interpretations in the prominent chapter "Are Supertaggers Expert Taggers", which is also the third principal contribution, I recommend revising the current version. In this chapter, the authors introduce a new metric that seeks to distinguish expert from non-expert users. To my surprise, the suggested metric is consensus-based and quantifies whether a user's tag choices "coincide with the eventual consensus of other users for that item" (p. 11).

Research on human cognition (e.g., Rogers & Patterson, 2007; Rosh et al.,1976) suggests that linguistic consensus emerges around labels/words indexing categories of an intermediate level of abstraction: On average, people prefer so-called basic-level terms (e.g., "tree") over super- and sub-ordinate terms (e.g., "plant" and "oak", respectively) to refer to an object. In CONTRAST to the consensus, experts of a given domain tend to deviate from this verbal behavior reliably (e.g., Tanaka & Taylor, 1991) and tend to apply more specific (sub-ordinate) labels. Hence, a metric that takes the extent of alignment with consensus as an index for expertise appears to be a contradiction in itself. Maybe, my claim is also in line with one of the authors' statements that "somewhat surprisingly, the results show an inverse-u shape" (Figure 11). Summarizing, I do not agree with the statement that "the measures used here assume the intuitive connection of expertise to consensus" (Section 7, page 13).

We agree that this is a more psychologically grounded conception of expertise, and have developed a new expertise measure that attempts to capture what the reviewer describes (we have also added the relevant references and background to Section 2). The details now appear in Section 6.2.3, and involve using an existing method for extracting a taxonomy from folksonomic data. From the resultant tree, we can determine a depth score for all tags considered (greater depth corresponds to a more subordinate term), and thereby examine if more prolific taggers utilize more sub-ordinate terms. We keep the original two expertise measures in the paper to facilitate comparison with past approaches to expertise and discussion of new approaches (that is, we now present three in total: SPEAR, our consensus-based measure, and the new depth-based measure). The results of the new measure are not as compelling as we had hoped, but are a useful addition to the paper.

[Note that the expertise-dependent indexing pattern is also discussed in Golder & Huberman (2006) referenced by the authors.]

In this context, a study of Fu & Dong (2010) who applied the SPEAR algorithm to differentiate experts and non-experts could be related work. Based on the LDA-model, these authors compared the probability distribution function (PDF) of the predictive probabilities of tags between experts and non-experts. They found that the center of the experts' PDF lies to the right of the center of the non-experts' PDF, i.e., experts' tags are more predictive for a particular topic and therefore, more specific than non-experts' tags. From my viewpoint, it should be conclusive to explore the assumption (maybe in future work) that differences between taggers and super-taggers are associated with differences in the PDFs of tags' predictive probabilities.

This is a very interesting study, but unfortunately we cannot replicate their analysis. Fu and Dong performed LDA on the actual resources tagged, and examined how predictive tags were of the resultant topics. We cannot perform such an analysis on our data because we either do not have the raw text content required to run the LDA (for Delicious we only know arbitrary numeric IDs for

each resource tagged) or else the content tagged is non-linguistic (Flickr photos, music on Last.fm). We have added a reference to the paper, however, as it represents a relevant methodology for the related work section of the paper.

Finally, I would like to emphasize that the current paper could be a very strong contribution that should be published if the aforementioned inconsistencies can be mitigated.

Reviewer D:

In this paper, authors study whether tagging is a large scale collaborative process or is it dominion of select few users.

In the introduction, authors do not provide a motivation for their problem. They should specifically consider answer the "so what?" question.

To help make this clearer, we have slightly reorganized the introductory section, and have also added some clarifying text and emphasis.

It is a well known fact that the power law distribution governs the online participation of users. As a result the conclusion of the work is not very surprising.

This is a fair criticism, and we have added clarification in the paper (last paragraph of introduction) that the power law distribution of user participation is not a surprising finding of the paper. We have added emphasis that the interesting finding are the difference in tagging patterns between more and less prolific users.

Since this paper is an extension of the prior work of the authors, they must specify clearly the extensions made in this work (preferrably in bullet or a subsection). It would be hard to determine otherwise whether the extension are of trivial or substantial nature. Also does the extension provide novel insights into the super-tagging behavior.

We have added a footnote addressing this to Section 1.

Overall the paper is an a interesting read, dataset is comprehensive and the experiments are carried out well.

We thank the reviewer for this positive evaluation.