ABSTRACT

Although analyzing user behavior within individual communities is an active and rich research domain, people usually interact with multiple communities both on- and off-line. How do users act in such multi-community environments? Although there are a host of intriguing aspects to this question, it has received much less attention in the research community in comparison to the intra-community case. In this paper, we examine three aspects of multi-community engagement: the sequence of communities that users post to, the language that users employ in those communities, and the feedback that users receive, using longitudinal posting behavior on Reddit as our main data source, and DBLP for auxiliary experiments. We also demonstrate the effectiveness of features drawn from these aspects in predicting users’ future level of activity.

One might expect that a user’s trajectory mimics the “settling-down” process in real life: an initial exploration of sub-communities before settling down into a few niches. However, we find that the users in our data continually post in new communities; moreover, as time goes on, they post increasingly evenly among a more diverse set of smaller communities. Interestingly, it seems that users that eventually leave the community are “destined” to do so from the very beginning, in the sense of showing significantly different “wandering” patterns very early on in their trajectories; this finding has potentially important design implications for community maintainers. Our multi-community perspective also allows us to investigate the “situation vs. personality” debate from language usage across different communities.

Categories and Subject Descriptors: J.4 [Computer Applications]: SOCIAL AND BEHAVIORAL SCIENCES; H.2.8 [Database Applications]: Data Mining

General Terms: Algorithms, Experimentation

Keywords: multiple communities; lifecycle; language; Reddit; DBLP

1. INTRODUCTION

How people behave within a given community is a profound and broad question that has inspired work ranging from basic social-sciences research (e.g., [24]) to the design of online social systems (e.g., [21]). However, many settings offer an array of multiple possible interest sub-groups for users to engage in. In the offline world, for example, within the bounds of a single college campus, students can get involved with a variety of clubs, organizations, and social circles. And in the online case, there are many multi-community sites, such as Reddit, 4chan, Wikia, and StackExchange, all of which host a slew of topic-based sub-discussion forums. As the results in this paper show, multi-community settings exhibit many interesting and useful properties that are not manifested in within-community situations, and so our main goal is to demonstrate that multi-community engagement is an exciting and underexplored research area: we believe that such work will shed additional light on human behavior and on the design of social-media systems.

To demonstrate, we first tackle a seemingly foregone conclusion: that, analogously to the human life course [5][14], a person first passes through an “adolescent” phase of trying out many different interests before “settling down”. Indeed, the best-paper award at WWW 2013 was given to an excellent within-community study [9] demonstrating (among other things) that users’ language use becomes more inflexible and out-of-step with the community’s over time. But, contrary to this expectation, we find that even people with long histories of participation in a global community continually try out new sub-communities. Figure 1 depicts this for two very different settings: for Reddit and for the universe of computer-science conferences given by DBLP, the latter choice inspired by [3]. Note that despite their very different timescales (one can post to Reddit at any time, but submission deadlines only roll around every so often) and barriers to entry (conferences have gate-keepers, whereas posting on Reddit can be done essentially at will), they exhibit the same qualitative behavior. On average, Redditors post to 5 communities in their first 10 posts and then post to 2.5 new communities every 10 posts (Fig. 1a and 1b). These exploration trends continue over the users’ lifetimes (Fig. 1c and 1d). Thus, while within a single community “all users die old” [9], it seems that a multi-community setting keeps users young by offering them choices to explore as an alternative to opting out entirely.

Having established the prevalence of “wandering” behavior, we are led to investigate a host of related phenomena. We believe that these phenomena are interesting in their own right, and at times quite surprising. Moreover, we also demonstrate that our findings inspire new kinds of features that are strongly predictive of users’ future level of activity.

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author’s site if the Material is used in electronic media.

WWW 2015, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741661.

ABSTRACT

Although analyzing user behavior within individual communities is an active and rich research domain, people usually interact with multiple communities both on- and off-line. How do users act in such multi-community environments? Although there are a host of intriguing aspects to this question, it has received much less attention in the research community in comparison to the intra-community case. In this paper, we examine three aspects of multi-community engagement: the sequence of communities that users post to, the language that users employ in those communities, and the feedback that users receive, using longitudinal posting behavior on Reddit as our main data source, and DBLP for auxiliary experiments. We also demonstrate the effectiveness of features drawn from these aspects in predicting users’ future level of activity.

One might expect that a user’s trajectory mimics the “settling-down” process in real life: an initial exploration of sub-communities before settling down into a few niches. However, we find that the users in our data continually post in new communities; moreover, as time goes on, they post increasingly evenly among a more diverse set of smaller communities. Interestingly, it seems that users that eventually leave the community are “destined” to do so from the very beginning, in the sense of showing significantly different “wandering” patterns very early on in their trajectories; this finding has potentially important design implications for community maintainers. Our multi-community perspective also allows us to investigate the “situation vs. personality” debate from language usage across different communities.

Categories and Subject Descriptors: J.4 [Computer Applications]: SOCIAL AND BEHAVIORAL SCIENCES; H.2.8 [Database Applications]: Data Mining

General Terms: Algorithms, Experimentation

Keywords: multiple communities; lifecycle; language; Reddit; DBLP

1. INTRODUCTION

How people behave within a given community is a profound and broad question that has inspired work ranging from basic social-sciences research (e.g., [24]) to the design of online social systems (e.g., [21]). However, many settings offer an array of multiple possible interest sub-groups for users to engage in. In the offline world, for example, within the bounds of a single college campus, students can get involved with a variety of clubs, organizations, and social circles. And in the online case, there are many multi-community sites, such as Reddit, 4chan, Wikia, and StackExchange, all of which host a slew of topic-based sub-discussion forums. As the results in this paper show, multi-community settings exhibit many interesting and useful properties that are not manifested in within-community situations, and so our main goal is to demonstrate that multi-community engagement is an exciting and underexplored research area: we believe that such work will shed additional light on human behavior and on the design of social-media systems.

To demonstrate, we first tackle a seemingly foregone conclusion: that, analogously to the human life course [5][14], a person first passes through an “adolescent” phase of trying out many different interests before “settling down”. Indeed, the best-paper award at WWW 2013 was given to an excellent within-community study [9] demonstrating (among other things) that users’ language use becomes more inflexible and out-of-step with the community’s over time. But, contrary to this expectation, we find that even people with long histories of participation in a global community continually try out new sub-communities. Figure 1 depicts this for two very different settings: for Reddit and for the universe of computer-science conferences given by DBLP, the latter choice inspired by [3]. Note that despite their very different timescales (one can post to Reddit at any time, but submission deadlines only roll around every so often) and barriers to entry (conferences have gate-keepers, whereas posting on Reddit can be done essentially at will), they exhibit the same qualitative behavior. On average, Redditors post to 5 communities in their first 10 posts and then post to 2.5 new communities every 10 posts (Fig. 1a and 1b). These exploration trends continue over the users’ lifetimes (Fig. 1c and 1d). Thus, while within a single community “all users die old” [9], it seems that a multi-community setting keeps users young by offering them choices to explore as an alternative to opting out entirely.

Having established the prevalence of “wandering” behavior, we are led to investigate a host of related phenomena. We believe that these phenomena are interesting in their own right, and at times quite surprising. Moreover, we also demonstrate that our findings inspire new kinds of features that are strongly predictive of users’ future level of activity.

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author’s site if the Material is used in electronic media.

WWW 2015, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741661.
they receive from other members of the community (§3.3). Consistently, we see that — again, in contrast to the “older people become less adventurous” hypothesis — our users appear to continually seek out new and different communities, and adopt the language characteristics of the new communities. Another interesting point, albeit arguably less surprising, is that they tend to move to smaller communities (a fact noted by Redditors), which might be a signal to site designers to make sure to offer a menu of narrowly-targeted options for users to choose from (or to ensure that sub-groups can arise organically). Finally, a complete surprise is that for users who made at least 50 posts, the patterns exhibited by those who end up staying by their first 10 posts. The fact that future abandonment can be detected so early should be of interest to administrators of social-media systems. But, there is an unexpected factor potentially making this discrimination difficult: in our data, the eventually departing users are often most similar not to the least active users in our study, but to the most active users. We conjecture that our “dying” users are actively striving to remain engaged, but are not quite managing to explore enough to make their overall posting experience satisfactory. A design implication might be to include mechanisms in one’s site that more proactively suggest new, diverse sub-communities for posting.

In Section 4 we show that the aforementioned differences in patterns are not “mere” correlations, but do indeed serve as features that are effective at predicting future activity level.

Again, our overall goal is to encourage further work on multi-community settings. As a spur to the imagination, and as a demonstration that this research domain is rich with possibilities, we discuss in sections 5 and 6 two additional questions that arise. First, what makes a user abandon a community and move on to new ones? We see that the positivity of initial feedback correlates with what groups users choose to return to, a finding that contradicts recent results on the power of negative feedback [8], albeit for commenting instead of posting. Second, we make a foray into the “status vs. personality” debate in psychology [20][12]: how much of our behavior is determined by fixed personality traits, versus how much is variable and influenced by the specific situation at hand? We consider this question from a linguistic perspective, and determine that even after topic-specific vocabulary is discarded (after all, it wouldn’t be interesting to find that people use gym-related words at the gym that they don’t use at work), users do employ different language patterns in different communities. This means that they are able to adapt even into “maturity” — a positive note to end on.

2. EXPERIMENTAL SETUP

In the following, we first describe the data that we use and then propose an analysis framework for capturing the temporal dynamics of multi-community engagement.

2.1 Datasets.

The main dataset used in this paper is drawn from Reddit, a very active community-driven platform for submitting, commenting on, and rating posts [10]. Reddit is organized into thousands of topic-based, user-created discussion forums called “subreddits”, which users can post to essentially at will (modulo spam filtering, rate limits, and deletion of posts by moderators). Other users can “up-vote” or “downvote” posts; the difference between the number of

---

1 One comment: “the longer you are on reddit, the more you get pulled into smaller subs”.

2 A Reddit post consists at a minimum of a title that serves as anchor-text for a link. The link may be to an offsite item (“link post”) or to some text that the post’s author places on Reddit (“text post”). The dataset with more detailed explanation is available at [https://chenhaot.com/pages/multi-community.html](https://chenhaot.com/pages/multi-community.html)
upvotes and the number of downvotes, a difference that we henceforth refer to as feedback, is readily available.

Relying primarily on RedditAnalytics, in February 2014 we collected all 76.6M posts ever submitted to Reddit since its inception, together with their associated feedback values. We discarded the last month of posts, since their feedback values might not have had sufficient time to converge.

Since we need our users’ community trajectories to be long enough to be able to exhibit significant wandering (whether or not they actually do), the set of users we consider are those who have made at least 50 posts, following the choice in . We focus on the 157K 50+ posters who first posted between January 2008 and January 2012 so that we have at least two years’ worth (2012-2014) of observations for each of them. We chose to start from January 2008 because users were granted the ability to create their own subreddits at will then. Not only are the 50+ posters good objects of study because we have a lot of data on their behavior, but they also play a major role in determining the character of Reddit because they made 63% of the posts written by users who first posted in the time period under consideration.

In order to ensure that our findings generalize beyond Reddit, we also consider a (more) physical-world multi-community situation: the set of conferences in computer science. Conferences generally correspond to topic areas within CS, and each can be thought of as representing a social group, at least to some degree. In this setting, we take “posting” to mean publishing a paper. We use the DBLP database to find what papers appeared in which conferences, and refer to the resultant dataset as “DBLP”. For DBLP, we do not consider an analog of Reddit’s feedback, although citation or download counts could be used in future work.

It is important to note that program committees play a huge role in determining an author’s conference trajectory. This property makes DBLP a less suitable domain for the questions of user choice that we focus on in this paper. We thus place our DBLP trajectory results in the Appendix (§9).

Statistics on the 50+ posters in Reddit and DBLP are given in Table 1.

### Note 1: how we define “posting”.

In this paper, we use the term posting to refer to submitting an item to be voted or commented upon. We distinguish posting from commenting in multi-community environments is an interesting way that comments are presented on Reddit makes scraping the useful that commentary is subject to: Reddit influences commenting by

Second, posting is not affected by a confounding factor age since the site will presumably die without fresh conversation-starters. The actual number of upvotes or downvotes is purposely inaccessible.

http://bit.ly/1xrcIQY

Except that we filter out bots and banned users.

Cross-posting (posting the same URL to multiple subreddits, with or without a title change) accounts for only 3% of the posts from the users that we consider in this paper — only 1.77% if we only consider their first 50 posts.

Figure 2: Illustration of windows and stages for window size $w = 10$, number of stages $S = 5$, number of posts $T = 150$, number of windows $T_w = 15$. $W_i$ is a window; $S_i$ is a stage.

### 2.2 Analysis framework.

We now set up terminology and concepts that facilitate discussion of users’ trajectories among communities.

For each post by a given user, we store the timestamp, time, and the community (sometimes $C$ for short). For Reddit data, we also store the post’s feedback as of February 2014 and its words (the anchor-text plus any text written by the user, all tokenized and part-of-speech tagged using the Stanford NLP package).

Several of the questions we are interested in pertain to properties of subsequences of trajectories. For example, suppose we want to know whether users are visiting a broader set of communities over time; one way to check is to look at how many communities they engaged with in their first $w$ posts versus in their last $w$ posts. Therefore, a basic element in our analysis is a window. Let variable $t$ index the posts made by a user $u$, and suppose $u$ has made $T$ posts altogether. We split the entire index sequence $1, \ldots, T$ into non-overlapping consecutive windows $W_i$ of size $w$, where $i$ ranges from 1 to $T_w \overset{def}{=} [T/w]$. For example, in Fig. $W_6$ would be the integers in the range $[51, 60]$. We use $w = 10$ throughout this paper. Our Reddit results were insensitive to choices of $w$.

We define functions $F$ on windows $W_i$ to summarize properties of that window and track how these properties change over time. We use two ways to define $F$. One way is to directly define $F$ based on the entire window, for example, $F(W_i) = |\{C_t : t \in W_i\}|$, the number of unique communities in $W_i$. The other way is to define a function $f$ for each index $t$ — for example, $f(t)$ could be the number of words in the $t$th post — and let $F(W_i)$ be induced by $f$’s average value over the indices in $W_i$, $F(W_i) = \frac{1}{|W_i|} \sum f(t)$.

Given a window size $w$ and a function of interest, $F$, we take two perspectives to track the trajectory of $F$: a full-life view (all the user’s posts) and a fixed-prefix view (50 posts). The rationals are as follows:

- The first perspective, full-life, tracks users’ entire lifetimes. Because the value of some functions is affected by choice of window size (e.g., the number of unique communities), we still fix the window size in the full-life view, but set an additional parameter $S$ of the number of life stages that we want to examine, where each life stage contains the same number of windows, as depicted in Fig. For each stage, we compute the average value over the windows in that stage.

A slight problem with the full-life view is that for different users, the value of the same life stage (say, the first 10% of one’s life)
Figure 3: Number of unique communities per window. x-axis: each of the first 5 windows. y-axis: number of unique communities appearing in the corresponding window. In Fig. 3b and Fig. 3c, users are categorized by their future state after the initial 50 posts. Standard-error intervals are depicted, but very small.

Note 2: y-axes scales, and other considerations regarding subsequent figures. Since many of the figures in Section 3 tend to support the same overall point as in Figure 3, we make the subsequent figures relatively small (labeling the y-axes in the captions), but use the same x-axis, legends, and line styles in all of them. As in Figure 3, each of the other figures in Section 3 consists of three sub-figures. In each, we scale the y-axes according to the corresponding data’s range in order to show significant changes (all figures show standard-error bars, which are tiny). But it should be noted that the lines when averaging over all users (leftmost sub-figure in the figures) would usually look flatter if plotted on the graphs that divide users by departure status (middle sub-figures) or activity quartile (rightmost sub-figures).

may be based on a significantly different number of posts (say, 10 for one user but 100 for another). The full-life view also includes information about the entirety of the user’s life, and thus is not appropriate for prediction settings (for example, one does not ordinarily know at the time what percent of one’s life has already passed). Thus we also take a fixed-prefix view, where only the initial 50 posts are examined. (Recall from the caption of Fig. 1 that this encompasses a long time span on average.) Thus, the same amount of data is used for every user and the induced features are valid for predicting future behavior. For space reasons, in the main paper we will focus on the fixed-prefix view, and place some full-life-view results in the Appendix (§9).

Future activity level. We further relate our analysis to users’ future activity level, since future activity level is a useful quantity to predict. We employ two different ways to categorize users’ future commitment: the two-way classification of whether a user eventually abandons the global community altogether or not, and a 4-way split based on the relative number of posts that a user eventually makes over his/her lifetime, as follows.

- Departing status. To determine which users should be considered to have abandoned the entirety of the user’s life, we define a date (specifically, 6 months before January 2014) as the start-of-future (SOF). We define departing users as those who stopped posting as of SOF; we define lasting users as non-departing users who additionally post at least once in the first 3 months and at least once in the second 3 months since SOF, so that they are consistently “active”. There are 43,910 departing users and 75,708 lasting users. Note that they all made at least 50 posts before SOF.

- Activity quartile. We split users into four quartiles based on the number of posts that they make in their entire life after the initial 50 posts. (As it happens, the lasting/departing ratio is higher in the the higher-activity quartile.)

3. TRAJECTORY PROPERTIES

We have established in Fig. 1 that users do constantly “wander around” in multi-community environments. In this section, we apply the framework proposed in §2 to explore three aspects of this wandering process: (§3.1) the communities users post to; (§3.2) the language users employ in each community; (§3.3) the feedback that users receive from other community members. In §3 we will further validate the effectiveness of features based on these properties in prediction tasks.

3.1 Multi-community aspects

We have shown in §1 that users on average consistently post to 2.5 new communities every 10 posts (Fig. 1). But what else characterizes their patterns of movement among communities? The answers to this question have the design implications outlined in §1.

Section summary. We find that over time, users span more communities every 10 posts, “jump” more, and concentrate less. They enter smaller and less similar communities. Eventually-departing users seem consistently less “adventurous” than lasting users even, notably, from the very beginning. Curiously, eventually-departing users act similarly to users in the top activity quartile.

In the following, we explain the metrics for understanding these properties and discuss related theories.

Users span more and more unique communities in a window, but relatively speaking, departing users span fewer unique communities. Figure 3 shows the per-window number of unique communities that users post to. The actual number is interesting: in Fig. 1 users post to 2.5 new communities every 10 posts; here on average, users post to around 5 communities every 10 posts, and thus only around 2.5 of them are ones that they have ever posted to. Given that users have more potential communities to go back
to over time, this suggests that they do not tend to return to some previous communities. More discussion as to why users return to certain communities will be presented in §5.

Users “jump” between communities more and more “frequently”, but departing users do so at around half the “rate”. (Fig. 3) To understand how often users “jump”, we count the number of “jumps” that users make per window. Formally, define $F(W_t) = \sum_{t \in W,} I(C_t \neq C_{t+1})$, where $I(x)$ is the indicator function: $I(x) = 1$ if $x$ is true, 0 otherwise.

Note that the number of unique communities in a window of 10 does not determine how often users “jump”. Given a window size of 10, users can jump as many as 9 times; given that users on average span 5 communities in a window, users can jump as few as 4 times. In fact, users make around 5.8 “jumps” per 10 posts.

![Figure 4: Number of “jumps”](image)

Users spread their posts out more and more evenly, but relatively speaking, departing users focus more. (Fig. 5) We employ entropy as a metric for concentration, following [1]. Entropy is based on the probability of a community appearing in a window $W_i$, $p_c = \frac{1}{|W_i|} \sum_{t \in W_i} I(C_t = c)$, and is defined as $- \sum_c p_c \log_2 p_c$ for $W_i$. It is an information-theoretic measure that grows as the intra-window community-posting distribution approaches the uniform distribution (minimum concentration) [25]. The same qualitative results hold if we use the Gini-Simpson index $(1 - \sum_c p_c^2)$, a commonly used metric in ecology for species concentration [16] [29].

![Figure 5: Entropy of community-posting distribution.](image)

Users enter smaller-looking communities (fewer posts per month), but relatively speaking, departing users prefer larger communities. (Fig. 6) Engaging with different communities entails a choice between communities of different sizes. A large community can encompass diverse community purposes and member preferences, leading to broader appeal, but at the same time, a large size may dilute personal connection and lead to more conflicts [24]. Or, size might not have any effect at all. To study this question, we set $f(t)$ to log of the number of posts made by the user in the community in month $t$ as a simple metric of how “large” the active portion of a community looks to an incoming user. [15]

![Figure 6: Average log$_2$number of monthly posts in communities that a user posts to.](image)

We note that with respect to this metric of community size, the full-life view, shown in the Appendix (Fig. 16a), differs from the fixed-prefix perspective plotted above. In the full-life view, the higher-activity quartile users eventually enter smaller communities than lower-activity quartile users. It seems that they just move more slowly to such communities.

Users post to less similar communities over time, but relatively speaking, departing users prefer more similar ones. (Fig. 7) One hypothesis for how people select new communities is that they explore similar communities to those they have visited in the past, because they want more exposure to topics that they are already interested in. On the other hand, perhaps they choose new communities because their interests have changed, implying that they would choose more different communities.

We measure the dissimilarity between communities $C_1$ and $C_2$ based on poster overlap, restricting attention to just those communities with at least 1000 posts to ensure sufficient data. Denoting the set of users who ever posted in a community $C$ as $U_C$, our measure is $1 - \frac{|U_{C_1} \cap U_{C_2}|}{|U_{C_1} \cup U_{C_2}|}$. Note that the dissimilarity between two communities is computed based on their eventual poster set, since we want to capture the “actual”, eventual relationship between the two, and so does not change over time. For a window $W_i$, the overall community dissimilarity $F(W_i)$ is defined as the average of all the pairwise dissimilarities between the communities that the user posted at during that window $W_i$.

The same trends hold if we measure language dissimilarity between communities using the KL-divergence between community language models.

**Different activity quartiles.** For all of the above metrics, users of different future activity quartiles manifest significant differences even in their very earliest behavior, although the differences are not as dramatic as those between departing users and lasting users. The curves for the different quartiles always appear in either the

---

11 An alternative hypothesis regarding the difference in activity quartiles is that there isn’t really a difference, but perhaps users in the higher-activity quartile make several posts in a single community where a lower-activity user makes just one, e.g., $C_1C_2C_3C_4C_5C_2C_5C_2$ vs. $C_1C_2$. If this were so, we would observe a lower entropy simply due to accidentally choosing a window size that is small relative to the average burst size. However, we verified that this “burstiness” hypothesis does not hold, since the higher-activity users only change communities about 0.5 fewer times than lower-activity ones.

12 Reddit does not provide directly applicable metrics: the number of subscribers or those “online now” can consist mostly of passive observers. The number of users who posted in a month is not presented at all, but we observe similar trends when extracting that as the metric.
order 1, 2, 3, 4 or 3, 2, 1, and the highest-activity quartile curves are always the closest to those for departing users.

### 3.2 Language aspects

The second aspect that we examine is the language that users employ within communities. This examination, and the formulation we apply below, are inspired by [9], which found that in single-community settings, users first pass through an “adolescent” phase where they learn linguistic norms, but after this phase stop adapting to new norms and become increasingly distant from the community. Our results indicate that this is not the case in the multi-community setting. Rather, with respect to part-of-speech tags or stopwords, users do not move farther and farther away from the community distribution; and when (frequent) content words are included, users seem to “stay young”, continuously growing closer to the community’s language. Surprisingly, departing users are better mimics of the community’s language than lasting users are. The bulk of this section provides the experimental evidence, based on various forms of cross-entropy, from which we draw these conclusions.

Additionally, we, like [9], find that the usage of 1st-person-singular pronouns (e.g., I, me) declines over time [1] which has been argued to indicate a greater sense of community affiliation [7, 28]. However, upon closer inspection, the fact that departing posters use these words less frequently than those users who end up staying seems problematic for such theories — although one could speculate that the cause is that our departing users start out with strong affiliation needs but become disappointed. These results are shown in Figure[8]

**Cross-entropy with vocabulary-varying language models.** We use cross-entropy to measure the distance between (a language model constructed from) a user’s $t^{th}$ post and a language model built from all the posts in the corresponding community, $C$, in that same month $m(t)$. Importantly, we will compute these models based on various choices of vocabulary $V$; this will reveal that although users’ topical-word usage grows closer and closer to that of the community’s, their usage in part-of-speech tags and stopwords stabilizes in terms of distance from the community’s.

The first step of our $V$-dependent language-model construction is to replace every instance of any word not in $V$ with the new token $<$RARE$>$

“$<$RARE$>$”. Next, we define the community-based language model to be the distribution over $v \in V \cup \{<$RARE$>\}$ given by setting $p^C$ to the relative frequency of $v$ in the concatenation words$_C^m(m(t)$ of all the posts in $C$ during the month $m(t)$. Then, we measure the cross-entropy by

$$f(t) = \frac{1}{|words_s|} \sum_{v \in words_s} \log_2 \frac{1}{p^C(v)}.$$  

(This equation shows why we do not need to smooth the community language model: since words$_s$ is a component of words$_C^m(m(t)$, $p^C(v) > 0$ for $v \in \text{words}_s$.)

With all of this in hand, Figure[9] depicts representative evidence for the conclusions we drew at the beginning of this section. Specifically, the evidence consists of cross-entropy values for $V$ chosen to be 46 parts-of-speech tags, the most frequent 100 words in Reddit, or the most frequent 1000 words in Reddit. Trends for $V$ set to the 500 or 5000 most frequent words are similar to the most frequent 1000 words.

**Technical aside: the potentially confounding factor of rare words interacting with community posting volume.** We also used a “full” vocabulary that contains all words that appear more than 100 times in Reddit (180K types), but do not show the results here. This is due to the fact that for large vocabulary sizes, what appears to be differences in language matching can actually be merely a side-effect of one class of users posting in more-voluble communities. The argument runs as follows. The full vocabulary allows for many words $v'$ with low frequency in the community — say, 1 — to contribute to the cross-entropy computation. The probability

![Figure 7: Community dissimilarity based on poster overlap.](image)

![Figure 8: Percentage of first singular person pronouns.](image)

![Figure 9: Distance from the community language model. The rows indicate different choices of vocabulary $V$.](image)
3.3 Feedback aspects

A final question that Reddit data allow us to easily answer is, how are users received by other members of the community? For each post, Reddit provides the difference between the number of upvotes and number of downvotes. Because the average value of this difference can vary among different communities, we measure the feedback that users get by the relative position of this difference among all posts in the community for that month, i.e., how often the posts made by a user outperform the “median post” in a community. For each index \( t \), we define \( f(t) \) as \( I(feedback_i > \text{median}(C_i, m(t))) \), where \( \text{median}(C, m) \) represents the median vote difference in community \( C \) in month \( m \).

Surprisingly, the feedback that 50+ posters receive is continually growing more positive, although the rate slows over time (Fig. 10). However, the growth is small compared to the drastic differences between departing users and lasting users. Even departing users get more-positive feedback over time, but the increase is not as great as for lasting users. Users in the top activity quartile also fare worse, although as shown in the relative perspective (Fig. 15b), they catch up in the later stages of their life. The results are consistent if we measure how often posts outperform 75% of the community’s posts.

3.4 Recap

In all three aspects that we examined, users with different future activity levels manifest significant differences in their trajectories of multi-community engagement. Interestingly, users that eventually depart seem “destined” to do so even from the very beginning, since the curves for the departing vs. lasting users generally start out apart and maintain or increase that distance over time. Meanwhile, there are smaller but significant differences in these metrics between users at different activity quartiles. It is important to note that some metrics can be correlated (e.g., number of unique communities and entropy). However, none of the metrics determines another, so we believe discussing each one of them was valuable.

Another interesting phenomenon we consistently observe is that for all our metrics, users in the top activity quartile are the closest to the departing users in the first 50 posts (a direct comparison for language is shown in Fig. 11).

---

14 This concern cannot be alleviated simply by sub-sampling a community’s posts, since the true root of the problem is rare words, not just the length and number of posts in the community per se.

15 Alternatively, one could set \( w_i = 50 \), thus extracting features from all 50 posts in a single batch. This approach turns out to be poorer than using 5 windows because trend information is not captured.
and the full vocabulary as defined in §3.2. Additionally, we include the proportion of 1st-person-singular pronouns and post length in words.

- Feedback aspects. This includes the fraction of posts that outperform 50% and 75% of all of the corresponding month’s worth of the community’s posts in terms of positivity of feedback. Refer back to §3.3 for more information.

For entropy, Gini-Simpson index, and number of unique communities, we include the value for all 50 posts, since for these features, the values for all 50 posts are not simply the average of the values from 5 windows of 10 posts. We also use the index of the window with the largest value and the smallest value as features, following §3. All features are linearly scaled to [0, 1] based on training data.

**Experiment protocol.** In both tasks, we perform 30 randomized trials. In each trial, we randomly draw 20,000 users from our dataset as training data and a distinct set of 5,000 users as testing data. We use 5,000 users from the training data as validation set. We use LIBLINEAR [15] in all prediction tasks. For significance testing, we employ the paired Wilcoxon signed rank test (p < 0.001).

The standard procedure for generating learning curves would be to only look at the first x posts as x varies, x = 10, 20, 30, 40, 50. A non-obvious but ultimately fruitful idea we introduce here is to contrast the effectiveness of the information in the early part of each 50-post instance with that of the late part of the 50-post instance. That is, we compare the performance if we use the first (‘fst’ in our plots) x posts with the performance of using the last x posts. (One might expect later periods to be more predictive, given that they are more recent. But surprisingly, we will see that when we predict departure status, we find that earlier information is more useful, which again suggests that departing users are “destined” to leave from the very beginning.)

### 4.1 Predicting departing status

**Basic comparisons.** (Figure 12a) Using all features outperforms a strong baseline that uses time-gap features by 18.3% — the difference between an F1 of .699 and an F1 of .591 — which shows the effectiveness of features drawn from multi-community engagement.

The performance of the first x posts is always above that of the last x posts. This suggests that the initial information is more predictive of eventual departure. Note that for 50+ posters, departure is quite “far away” from the initial posts. In fact, using all features drawn from only the first 10 posts outperforms time-gap features extracted from all 50 posts. Thus it may be very important for designers of social systems to make sure that users start well, perhaps through positive feedback or by recommending communities to visit based on training data.

**Feature-set analysis.** (Figure 12b) In predicting departure, it is most useful to know how well users match a community’s language. The second most useful features are the patterns of community visitation, language-matching, community-trajectory, and community-feedback features all outperform time-gap information, which suggests that how users interact with different communities is more predictive than activity rate in predicting whether 50+ users will leave.

### 4.2 Predicting activity quartile

**Comparisons with the baseline.** (Figure 13a, 13b) In contrast to the case just discussed of predicting departure status, time-gap between posts is a much stronger feature in predicting future total number of posts. This is plausible because for these 50+ posters, time-gaps in posting determine how many posts that people can physically make. However, adding all the features based on multi-community engagement still improves the performance over time information to a statistically significant degree. Prior work has shown that adding language features can lead to big improvements over time-based features (p < 0.001).

![Figure 13a: Window comparison](image1)

![Figure 13b: Feature comparison](image2)

**Figure 13:** Results for predicting log₂(future total number of posts). y-axis: RMSE, the smaller the better. The line styles are the same as in Fig. 12. “Average” shows a baseline that always predicts the mean value in the training data. All differences for 50 posts are statistically significant according to the Wilcoxon signed rank test (p < 0.001).

Similarly, we can frame the “how much” issue succinctly by asking the following question. Suppose we partition the set of communities a user visits into (1) those that he or she abandons after just a single post, and (2) those that he or she posts at least twice to. Which set — the single-post communities or multiple-posts communities, is larger, on average? We claim that the answer is not a priori obvi-
6. **DO USERS SPEAK DIFFERENTLY IN DIFFERENT COMMUNITIES?**

So far we have revealed interesting and sometimes arguably counterintuitive properties of multi-community engagement, and demonstrated that they are effective cues in predicting a user’s future activity level. But an additional fascinating and orthogonal question is: when users participate in multiple communities, to what degree are their actions stable across settings? To look at this question is to contribute another piece of evidence to the “situation vs. personality” debate: how much of our behavior is determined by fixed personality traits, versus how much is variable and influenced by the specific situation at hand? Or, to put it a bit more dramatically, are you fundamentally the same person at work as you are at the gym?

Here, we study the question with respect to language use. The overall message is, *even after topic-specific vocabulary is discarded* (after all, it wouldn’t be interesting to find that people use gym-related words at the gym that they don’t use at work), individuals *do* employ different language patterns in different communities. The way we determine this is conceptually straightforward: we check whether it’s possible to tell which community a user’s posts come from based just on the distribution of stopwords or non-content-words within their posts.

The rest of this section gives a quick sketch of our experiments. (Space constraints preclude a full discussion of the details.)

If we fix some vocabulary $V$ of non-content words, then we can create classification instances from the 227K triples that exist in our data consisting of (1) a user $u$, (2) words of $u$’s first 25 posts in some community $C_1$, and (3) words of $u$’s first 25 posts in a different community $C_2$. Then, we compute the cross entropy of each post against the corresponding monthly language models, over the restricted vocabulary $V$, constructed from each of the two communities $C_1$ and $C_2$.

Add-1/$|V|$ smoothing is applied to all language models concerned. We then use these non-content-word cross-entropies as features to guess which of (2) and (3) came from community $C_1$.

We run experiments for several choices of $V$: parts-of-speech, the 100 most frequent words in Reddit, and the 500 most frequent words in Reddit. The first two choices definitely do not include topic-specific words, and the latter will not include many (there are 180K words in the full Reddit vocabulary), and so these choices may be taken to represent a user’s language style.

If the user’s style does not change from community to community, then the cross-entropy features mentioned above will not be helpful for determining that item (1) comes from $C_1$ and not $C_2$; thus, accuracy at matching language model to community would be 50%. But, as shown below, the average accuracies, utilizing logistic classification, of 30 random-split experiments (10K tuples for training and development, 2500 for testing) for each choice of $V$ are (statistically) significantly above 50%:

<table>
<thead>
<tr>
<th>$V$</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>parts of speech</td>
<td>62.3%</td>
</tr>
<tr>
<td>most frequent 100 words</td>
<td>56.0%</td>
</tr>
<tr>
<td>most frequent 500 words</td>
<td>61.4%</td>
</tr>
</tbody>
</table>

7. **RELATED WORK**

Anthropologists, psychologists and sociologists have looked at some questions regarding multi-community engagement, often in the context of interaction with new social circles or cultures. Recently, computer scientists have turned to examining multi-community engagement data available online. Our work differs by focusing on the following specific problems: (a) characterizing full community-trajectory sequences, as opposed to looking at pairwise community transitions; (b) revealing how properties of these trajectories correlate with a user’s...
future cross-community activity — we incorporate but also go beyond language-based features, as inspired by previous within-community work [9, 25], and timing-based features [13]. (c) considering the effect of each community’s positive and negative feedback, which may shed light on why users choose some communities over others.

Researchers have also been working on predicting users’ survival (also known as churn prediction) [10, 13, 34] and activity level [11, 37]. They focus on the single-community setting. A number of studies examined community-level evolution or the success of individual communities (often websites) [18, 19, 23, 33, 36], whereas our work focuses on the life cycle of users.

8. CONCLUDING DISCUSSION

Summary. We have investigated properties of multi-community engagement; this is a setting that has not received much computational research attention before, and yet is important because it encompasses many online and physical situations. In this first large-scale study of the phenomenon, we have found a number of sometimes counterintuitive but robust properties — some involving choice of community, some involving language use within communities, and some involving feedback from communities — revolving around the discovery that users “wander” and explore communities to a greater extent than might have been previously suspected.

Limitations and further directions. We focused on posting, but commenting and other related behaviors are very interesting subjects for future study. Our study is quantitative and observational. Qualitative studies, or controlled experiments regarding the design implications in [11] can further improve our understanding.

It is important to note that our study is limited to “50+ posters” so that we would have enough history per user to observe a relatively long trajectory. This is an unusually engaged group of users that comprises 5.9% of our users. We have not addressed the question of how multi-community engagement is exhibited by users who are not as active.

The notion of considering users to exist in a multi-community setting can in principle be extended to looking at user behavior across multiple websites or apps. With the advent and adoption of multiple-website services such as OpenID, observing users at that scale of multi-community engagement may well become quite important in the future.

There are many more challenging questions that arise from taking a multi-community perspective. For example, are the particularly nomadic treated differently? What is multi-community engagement like in real life, considering the cost of switching? How can we extend current theories and principles in community design to a multi-community setting? Further understanding of these questions is crucial for on- and off-line community design and an exciting direction for future work.

9. APPENDIX

Full-life view for users in Reddit. In general, the overall trends and differences between departing users and staying users are the same as in the fixed-prefix view. But in terms of activity quartiles, there are some interesting differences. For example, the ordering of the activity quartiles with respect to mean log$_{10}$ (number of posts that month) completely reverses itself (compare Fig. 16a to Fig. 5c).

For feedback, as users receive better feedback over time, users in the top activity quartile receive worse feedback in the beginning and catch up later in their life (Fig. 16b). These results are natural consequences of the trend developing over time. This suggests that the trends that we observe are robust over user life.

Fixed-prefix view for researchers in DBLP. In DBLP, authors span more conferences per window over time (Fig. 17a) in an increasingly scattered fashion (Fig. 17b), but in contrast to Reddit, there is saturation in the last two windows. Perhaps this suggests that as researchers become very senior, they publish more papers in some favorite set of venues.

When a very small window size is considered ($w=5$), the number of unique conferences and within-window entropy first increase and then decrease (Fig. 17c and 17d). But, changing the window size does not affect our central observation in Fig. 1 that 50+ researchers are publishing in new conferences at a relatively consistent rate over the years.

Acknowledgments. “Not all those who wander are lost” (J. R. R. Tolkien). We would have been lost without C. Danescu-Niculescu-Mizil, J. Hessell, A. Katiyar, J. Kleinberg, B. Pang, F. Radlinski, A. Sharma, K. Sridharan, A. Swaminathan, Y. Yue, the Cornell NLP seminar participants and the reviewers for their comments, and Jason Baumgartner for redditanalytics.com. This work was supported in part by NSF grant IIS-0910664 and a Google Research Grant.
The effect of wording on message propagation: 
Topic- and author-controlled natural experiments on Twitter

Chenhao Tan  
Dept. of Computer Science  
Cornell University  
chenhao@cs.cornell.edu

Lillian Lee  
Dept. of Computer Science  
Cornell University  
llee@cs.cornell.edu

Bo Pang  
Google Inc.  
bopang42@gmail.com

Abstract

Consider a person trying to spread an important message on a social network. He/she can spend hours trying to craft the message. Does it actually matter? While there has been extensive prior work looking into predicting popularity of social media content, the effect of wording per se has rarely been studied since it is often confounded with the popularity of the author and the topic. To control for these confounding factors, we take advantage of the surprising fact that there are many pairs of tweets containing the same url and written by the same user but employing different wording. Given such pairs, we ask: which version attracts more retweets? This turns out to be a more difficult task than predicting popular topics. Still, humans can answer this question better than chance (but far from perfectly), and the computational methods we develop can do better than both an average human and a strong competing method trained on non-controlled data.

1 Introduction

How does one make a message “successful”? This question is of interest to many entities, including political parties trying to frame an issue (Chong and Druckman, 2007), and individuals attempting to make a point in a group meeting. In the first case, an important type of success is achieved if the national conversation adopts the rhetoric of the party; in the latter case, if other group members repeat the originating individual’s point.

The massive availability of online messages, such as posts to social media, now affords researchers new means to investigate at a very large scale the factors affecting message propagation, also known as adoption, sharing, spread, or virality. According to prior research, important features include characteristics of the originating author (e.g., verified Twitter user or not, author’s messages’ past success rate), the author’s social network (e.g., number of followers), message timing, and message content or topic (Artzi et al., 2012; Bakshy et al., 2011; Borghol et al., 2012; Guerini et al., 2011; Guerini et al., 2012; Hansen et al., 2011; Hong et al., 2011; Lakkaraju et al., 2013; Milkman and Berger, 2012; Ma et al., 2012; Petrović et al., 2011; Romero et al., 2013; Suh et al., 2010; Sun et al., 2013; Tsur and Rappoport, 2012). Indeed, it’s not surprising that one of the most retweeted tweets of all time was from user BarackObama, with 40M followers, on November 6, 2012: “Four more years. [link to photo].”

Our interest in this paper is the effect of alternative message wording, meaning how the message is said, rather than what the message is about. In contrast to the identity/social/timing/topic features mentioned above, wording is one of the few factors directly under an author’s control when he or she seeks to convey a fixed piece of content. For example, consider a speaker at the ACL business meeting who has been tasked with proposing that Paris be the next ACL location. This person cannot on the spot become ACL president, change the shape of his/her social network, wait until the next morning to speak, or campaign for Rome instead; but he/she can craft the message to be more humorous, more informative, emphasize certain aspects instead of others, and so on. In other words, we investigate whether a different choice of words affects message propagation, controlling for user and topic: would user BarackObama have gotten significantly more (or fewer) retweets if he had used some alternate wording to announce his reelection?

Although we cannot create a parallel universe
in which BarackObama tweeted something else,[4] unfortunately, a surprising characteristic of Twitter allows us to run a fairly analogous natural experiment: external forces serendipitously provide an environment that resembles the desired controlled setting (DiNardo, 2008). Specifically, it turns out to be unexpectedly common for the same user to post different tweets regarding the same URL — a good proxy for fine-grained topic[5] — within a relatively short period of time.[6] Some example pairs are shown in Table[1]; we see that the paired tweets may differ dramatically, going far beyond word-for-word substitutions, so that quite interesting changes can be studied.

Looking at these examples, can one in fact tell from the wording which tweet in a topic- and author-controlled pair will be more successful? The answer may not be a priori clear. For example, for the first pair in the table, one person we asked found $t_1$’s invocation of a “scandal” to be more attention-grabbing; but another person preferred $t_2$ because it is more informative about the URL’s content and includes “fight media portrayal”. In an Amazon Mechanical Turk (AMT) experiment ($\S$1), we found that humans achieved an average accuracy of 61.3%: not that high, but better than chance, indicating that it is somewhat possible for humans to predict greater message spread from different deliveries of the same information.

Table 1: Topic- and author-controlled (TAC) pairs. Topic control = inclusion of the same URL.

<table>
<thead>
<tr>
<th>author</th>
<th>tweets</th>
<th>#retweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>halfsecuritycnn</td>
<td>First ON CNN: After Petraeus scandal, Paula Broadwell looks to recapture ‘normal life.’ [same URL]</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>First on CNN: Broadwell photos shared with Security Clearance as she and her family fight media portrayal of her [same URL]</td>
<td>2</td>
</tr>
<tr>
<td>ABC</td>
<td>$t_1$: Workers, families take stand against Thanksgiving Day. [same URL]</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>$t_2$: Staples, Medieval Times Workers say Opening Thanksgiving Day Crosses the Line [same URL]</td>
<td>2</td>
</tr>
<tr>
<td>cactus_music</td>
<td>$t_1$: I know at some point you’ve been saved from hunger by our rolling food trucks friends. Let’s help support them! [same URL]</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>$t_2$: Food trucks are the epitome of small independently owned LOCAL businesses! Help keep them going! Sign the petition [same URL]</td>
<td>13</td>
</tr>
</tbody>
</table>

First on CNN: After Petraeus scandal, Paula Broadwell looks to recapture ‘normal life.’ http://t.co/zg9jwA5j

Staples, Medieval Times Workers say Opening Thanksgiving Day Crosses the Line http://t.co/qy7GGuYW

I know at some point you’ve been saved from hunger by our rolling food trucks friends. Let’s help support them! http://t.co/ezfjwAS

Food trucks are the epitome of small independently owned LOCAL businesses! Help keep them going! Sign the petition http://t.co/ezfjwAS

Twitter-specific features of more successful phrasings. [5.1] applies hypothesis testing (with Bonferroni correction to ameliorate issues with multiple comparisons) to investigate the utility of features like informativeness, resemblance to headlines, and conformity to the community norm in language use. [5.2] further validates our findings via prediction experiments, including on completely fresh held-out data, used only once and after an array of standard cross-validation experiments.[7]

We achieved 66.5% cross-validation accuracy and 65.6% held-out accuracy with a combination of our custom features and bag-of-words. Our classifier fared significantly better than a number of baselines, including a strong classifier trained on the most- and least-retweeted tweets that was even granted access to author and timing metadata.

2 Related work

The idea of using carefully controlled experiments to study effective communication strategies dates back at least to Hovland et al. (1953). Recent studies range from examining what characteristics of New York Times articles correlate with high re-sharing rates (Milkman and Berger, 2012) to looking at how differences in description affect the spread of content-controlled videos or images (Borghol et al., 2012; Lakkaraju et al., 2013; Simmons et al., 2011) examined the variation of quotes from different sources to examine how textual memes mutate as people pass them along, but did not control for author. Predicting the “success” of various texts such as novels and movie quotes has been the aim of additional prior work not already mentioned in [1] (Ashok et al., 2013; Louis and Nenkova, 2013; Danescu-Niculescu-Mizil et al., 2012; Piter and Nenkova, 2008; McIntyre and Lapata, 2009). To our knowledge, there have been no large-scale studies exploring wording effects in a both topic- and author-controlled setting. Employing such controls, we find that predicting the more effective alternative wording is much harder than the previously well-studied problem of pre-

[5.1] Cf. the Music Lab “multiple universes” experiment to test the randomness of popularity (Salganik et al., 2006).
[5.2] Although hashtags have been used as coarse-grained topic labels in prior work, for our purposes, we have no assurance that two tweets both using, say, “#Tahrir” would be attempting to express the same message but in different words. In contrast, see the same-URL examples in Table[1].

[6] Moreover, Twitter presents tweets to a reader in strict chronological order, so that there are no algorithmic-ranking effects to compensate for in determining whether readers saw a tweet. And, Twitter accumulates retweet counts for the entire retweet cascade and displays them for the original tweet at the root of the propagation tree, so we can directly use Twitter’s retweet counts to compare the entire reach of the different versions.

dicting popular content when author or topic can freely vary.

Related work regarding the features we considered is deferred to §5.1 (features description).

3 Data

Our main dataset was constructed by first gathering 1.77M topic- and author-controlled (henceforth TAC) tweet pairs differing in more than just spacing. We accomplished this by crawling timelines of 236K user ids that appear in prior work (Kwak et al., 2010; Yang and Leskovec, 2011) via the Twitter API. This crawling process also yielded 632K TAC pairs whose only difference was spacing, and an additional 558M “unpaired” tweets; as shown later in this paper, we used these extra corpora for computing language models and other auxiliary information. We applied non-obvious but important filtering — described later in this section — to control for other external factors and to reduce ambiguous cases. This brought us to a set of 11,404 pairs, with the gold-standard labels determined by which tweet in each pair was the one that received more retweets according to the Twitter API. We then did a second crawl to get an additional 1,770 pairs to serve as a held-out set.

The corresponding tweet IDs are available online at http://chenhaot.com/pages/wording-for-propagation.html (Twitter’s terms of service prohibit sharing the actual tweets.) Throughout, we refer to the textual content of the earlier tweet within a TAC pair as $t_1$, and of the later one as $t_2$. We denote the number of retweets received by each tweet by $n_1$ and $n_2$, respectively. We refer to the tweet with higher (lower) $n_i$ as the “better (worse)” tweet.

Using “identical” pairs to determine how to compensate for follower-count and timing effects. In an ideal setting, differences between $n_1$ and $n_2$ would be determined solely by differences in wording. But even with a TAC pair, retweets might exhibit a temporal bias because of the chronological order of tweet presentation ($t_1$ might enjoy a first-mover advantage (Borghol et al., 2012) because it is the “original”; alternatively, $t_2$ might be preferred because retweeters consider $t_1$ to be “stale”). Also, the number of followers an author has can have complicated indirect effects on which tweets are read (space limits preclude discussion).

We use the 632K TAC pairs wherein $t_1$ and $t_2$ are identical† to check for such confounding effects: we see how much $n_2$ deviates from $n_1$ in such settings, since if wording were the only explanatory factor, the retweet rates for identical tweets ought to be equal. Figure 1(a) plots how the time lag between $t_1$ and $t_2$ and the author’s follower-count affect the following deviation estimate:

$$D = \sum_{0 \leq n_1 < 10} |\hat{E}(n_2|n_1) - n_1|,$$

where $\hat{E}(n_2|n_1)$ is the average value of $n_2$ over pairs whose $t_1$ is retweeted $n_1$ times. (Note that the number of pairs whose $t_1$ is retweeted $n_1$ times decays exponentially with $n_1$; hence, we condition on $n_1$ to keep the estimate from being dominated by pairs with $n_1 = 0$, and do not consider $n_1 \geq 10$ because there are too few such pairs to estimate $\hat{E}(n_2|n_1)$ reliably.) Figure 1(a) shows that the setting where we (i) minimize the confounding effects of time lag and author’s follower-count and (ii) maximize the amount of data to work with

---

†No data collection/processing was conducted at Google.

‡The total excludes: tweets containing multiple URLs; tweets from users posting about the same URL more than five times (since such users might be spammers); the third, fourth, or fifth version for users posting between three and five tweets for the same URL; retweets (as identified by Twitter’s API or by beginning with “RT @”); non-English tweets.
is: when \( t_2 \) occurs within 12 hours after \( t_1 \) and the author has more than 5,000 followers. Figure 1(b) confirms that for identical TAC pairs, our chosen setting indeed results in \( n_2 \) being on average close to \( n_1 \), which corresponds to the desired setting where wording is the dominant differentiating factor.

**Focus on meaningful and general changes.**

Even after follower-count and time-lapse filtering, we still want to focus on TAC pairs that (i) exhibit significant/interesting textual changes (as exemplified in Table 1 and as opposed to typo corrections and the like), and (ii) have \( n_2 \) and \( n_1 \) sufficiently different so that we are confident in which \( t_i \) is better at attracting retweets. To take care of (i), we discarded the 50% of pairs whose similarity was above the median, where similarity was tf-based cosine. \(^8\) For (ii), we sorted the remaining pairs by \( n_2 - n_1 \) and retained only the top and bottom 5%. \(^9\) Moreover, to ensure that we do not overfit to the idiosyncrasies of particular authors, we cap the number of pairs contributed by each author to 50 before we deal with (ii).

### 4 Human accuracy on TAC pairs

We first ran a pilot study on Amazon Mechanical Turk (AMT) to determine whether humans can identify, based on wording differences alone, which of two topic- and author-controlled tweets is spread more widely. Each of our 5 AMT tasks involved a disjoint set of 20 randomly-sampled TAC pairs (with \( t_1 \) and \( t_2 \) randomly reordered); subjects indicated “which tweet would other people be more likely to retweet?”, provided a short justification for their binary response, and clicked a checkbox if they found that their choice was a “close call”. We received 39 judgments per pair in aggregate from 106 subjects total (9 people completed all 5 tasks). The subjects’ justifications were of very high quality, convincing us that they all did the task in good faith. \(^10\) Two examples for the third TAC pair in Table 1 were: “[\( t_1 \) makes] the cause relate-able to some people, therefore showing more of an appeal as to why should they click the link and support” and, expressing the opposite view, “I like [\( t_2 \)] more because [\( t_1 \)] starts out with a generalization that doesn’t affect me and try to make me look like I had that experience before”.

If we view the set of 3900 binary judgments for our 100-TAC-pair sample as constituting independent responses, then the accuracy for this set is 62.4% (rising to 63.8% if we exclude the 587 judgments deemed “close calls”). However, if we evaluate the accuracy of the majority response among the 39 judgments per pair, the number rises to 73%. The accuracy of the majority response generally increases with the dominance of the majority, going above 90% when at least 80% of the judgments agree (although less than a third of the pairs satisfied this criterion).

Alternatively, we can consider the average accuracy of the 106 subjects: 61.3%, which is better than chance but far from 100%. \(^11\) This result is noticeably lower than the 73.8%-81.2% reported by Petrović \textit{et al.} (2011), who ran a similar experiment involving two subjects and 202 tweet pairs, but where the pairs were not topic- or author-controlled. \(^12\)

We conclude that even though propagation prediction becomes more challenging when topic and author controls are applied, humans can still to some degree tell which wording attracts more retweets. Interested readers can try this out themselves at [http://chenhaot.com/retweetedmore/quiz](http://chenhaot.com/retweetedmore/quiz).

### 5 Experiments

We now investigate computationally what wording features correspond to messages achieving a broader reach. We start ([5.1]) by introducing a set of generally-applicable and (mostly) non-Twitter-specific features to capture our intuitions about what might be better ways to phrase a message. We then use hypothesis testing ([5.1]) to evaluate the importance of each feature for message propagation.

---

\(^8\) We also computed the Pearson correlation between \( n_1 \) and \( n_2 \), even though it can be dominated by pairs with smaller \( n_1 \). The correlation is 0.853 for “\( > 5K \) f.ters, \( < 12hrs \)”, clearly higher than the 0.305 correlation for “otherwise”.

\(^9\) Tf weighting was not employed because changes to frequent words are of potential interest. Urls, hashtags, @-mentions and numbers were normalized to [url], [hashtag], [at], and [num] before computing similarity.

\(^10\) For our data, this meant \( n_2 - n_1 \geq 10 \) or \( \leq -15 \). Cf. our median number of retweets: 30.

\(^11\) We also note that the feedback we got was quite positive, including: “...It’s fun to make choices between close tweets and use our subjective opinion. Thanks and best of luck with your research” and “This was very interesting and really made me think about how I word my own tweets. Great job on this survey!”.

\(^12\) The accuracy range stems from whether author’s social features were supplied and which subject was considered.
agitation and the extent to which authors employ it, followed by experiments on a prediction task ([§5.2]) to further examine the utility of these features.

### 5.1 Features: efficacy and author preference

What kind of phrasing helps message propagation? Does it work to explicitly ask people to share the message? Is it better to be short and concise or long and informative? We define an array of features to capture these and other messaging aspects. We then examine (i) how effective each feature is for attracting more retweets; and (ii) whether authors prefer applying a given feature when issuing a second version of a tweet.

First, for each feature, we use a one-sided paired t-test to test whether, on our 11K TAC pairs, our score function for that feature is larger in the better tweet versions than in the worse tweet versions, for significance levels $\alpha = .05, .01, .001, 1e-20$. Given that we did 39 tests in total, there is a risk of obtaining false positives due to multiple testing ([Dunn, 1961] [Benjamini and Hochberg, 1995]). To account for this, we also report significance results for the conservatively Bonferroni-corrected (“BC”) significance level $\alpha = 0.05/39=1.28e-3$.

Second, we examine author preference for applying a feature. We do so because one (but by no means the only) reason authors post $t_2$ after having already advertised the same URL in $t_1$ is that these authors were dissatisfied with the amount of attention $t_1$ got; in such cases, the changes may have been specifically intended to attract more retweets. We measure author preference for a feature by the percentage of our TAC pairs where $t_2$ has more “occurrences” of the feature than $t_1$, which we denote by “%(f$_2 >$ f$_1$)”.

We use the one-sided binomial test to see whether %(f$_2 >$ f$_1$) is significantly larger (or smaller) than 50%.

---

**Table 2: Notational conventions for tables in 5.1**

<table>
<thead>
<tr>
<th>One-sided paired t-test for feature efficacy</th>
<th>One-sided binomial test for feature increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\uparrow\uparrow\uparrow\uparrow$: p&lt;1e-20 $\downarrow\downarrow\downarrow\downarrow$: p&gt;1e-20</td>
<td>YES: $t_2$ has a higher feature score than $t_1$, $\alpha = .05$</td>
</tr>
<tr>
<td>$\uparrow\uparrow$: p&lt;0.001 $\downarrow\downarrow$: p&gt;0.999</td>
<td>NO: $t_2$ has a lower feature score than $t_1$, $\alpha = .05$</td>
</tr>
<tr>
<td>$\uparrow$: p&lt;0.01 $\downarrow$: p&gt;0.99</td>
<td>(%)$: %(f_2 &gt; f_1)$, if sig. larger or smaller than 50%</td>
</tr>
<tr>
<td>$\uparrow$: p&lt;0.05 $\downarrow$: p&gt;0.95</td>
<td></td>
</tr>
<tr>
<td>*: passes our Bonferroni correction</td>
<td></td>
</tr>
</tbody>
</table>

---

**Table 3: Explicit requests for sharing (where only occurrences POS-tagged as verbs count, according to the Gimpel et al. (2011) tagger)**

<table>
<thead>
<tr>
<th>effective?</th>
<th>author-preferred?</th>
</tr>
</thead>
<tbody>
<tr>
<td>rt</td>
<td>YES (59%)</td>
</tr>
<tr>
<td>spread</td>
<td>YES (56%)</td>
</tr>
<tr>
<td>please</td>
<td>---</td>
</tr>
<tr>
<td>pls</td>
<td>---</td>
</tr>
<tr>
<td>plz</td>
<td>---</td>
</tr>
</tbody>
</table>

---

**Table 4: Informativeness**

<table>
<thead>
<tr>
<th>effective?</th>
<th>author-preferred?</th>
</tr>
</thead>
<tbody>
<tr>
<td>length (chars)</td>
<td>YES (54%)</td>
</tr>
<tr>
<td>verb</td>
<td>YES (56%)</td>
</tr>
<tr>
<td>noun</td>
<td>YES (56%)</td>
</tr>
<tr>
<td>adjective</td>
<td>YES (51%)</td>
</tr>
<tr>
<td>adverb</td>
<td>YES (55%)</td>
</tr>
<tr>
<td>proper noun</td>
<td>NO (45%)</td>
</tr>
<tr>
<td>number</td>
<td>NO (48%)</td>
</tr>
<tr>
<td>hashtag</td>
<td>---</td>
</tr>
<tr>
<td>@-mention</td>
<td>YES (53%)</td>
</tr>
</tbody>
</table>

---

Not surprisingly, it helps to ask people to share. (See Table 3; the notation for all tables is explained in Table 2). The basic sanity check we performed here was to take as features the number of occurrences of the verbs ‘rt’, ‘retweet’, ‘please’, ‘spread’, ‘pls’, and ‘plz’ to capture explicit requests (e.g. “please retweet”).

**Informativeness helps.** (Table 4) Messages that are more informative have increased social exchange value ([Homans, 1958]), and so may be more worth propagating. One crude approximation of informativeness is length, and we see that length helps. In contrast, [Simmons et al. (2011)] found that shorter versions of memes are more likely to be popular. The difference may result from TAC-pair changes being more drastic than the variations that memes undergo.

A more refined informativeness measure is counts of the parts of speech that correspond to content. Our POS results, gathered using a Twitter-specific tagger (Gimpel et al., 2011), echo those of [Ashok et al. (2013)] who looked at predict-
being the success of books. The diminished effect of hashtag inclusion with respect to what has been reported previously (Suh et al., 2010; Petrović et al., 2011) presumably stems from our topic and author controls.

**Be like the community, and be true to yourself (in the words you pick, but not necessarily in how you combine them).** (Table 5) Although distinctive messages may attract attention, messages that conform to expectations might be more easily accepted and therefore shared. Prior work has explored this tension: Lakkaraju et al. (2013), in a content-controlled study, found that the more upvoted Reddit image titles balance novelty and familiarity; Danescu-Niculescu-Mizil et al. (2012) (henceforth DCKL’12) showed that the memorability of movie quotes corresponds to higher lexical distinctiveness but lower POS distinctiveness; and Sun et al. (2013) observed that deviating from one’s own past language patterns correlates with more retweets.

Keeping in mind that the authors in our data have at least 5000 followers, we consider two types of language-conformity constraints an author might try to satisfy: to be similar to what is normal in the Twitter community, and to be similar to what his or her followers expect. We measure a tweet’s similarity to expectations by its score according to the relevant language model, $\frac{1}{|T|} \sum_{x \in T} \log(p(x))$, where $T$ refers to either all the unigrams (unigram model) or all and only bigrams (bigram model). We trained a Twitter-community language model from our 558M unpaired tweets, and personal language models from each author’s tweet history.

**Imitate headlines.** (Table 6) News headlines are often intentionally written to be both informative and attention-getting, so we introduce the idea of scoring by a language model built from New York Times headlines.17

**Use words associated with (non-paired) retweeted tweets.** (Table 7) We expect that provocative or sensationalistic tweets are likely to make people react. We found it difficult to model provocativeness directly. As a rough approximation, we check whether the changes in $t_2$ with respect to $t_1$ (which share the same topic and author) involve words or parts-of-speech that are associated with high retweet rate in a very large separate sample of unpaired tweets (retweets and replies discarded). Specifically, for each word $w$ that appears more than 10 times, we compute the probability that tweets containing $w$ are retweeted more than once, denoted by $rs(w)$. We define the retweet score of a tweet as $max_{w \in T \& \text{tag}(w)=z} rs(w)$, where $T$ is all the words in the tweet, and the retweet score of a particular POS tag $z$ in a tweet as $max_{w \in T \& \text{tag}(w)=z} rs(w)$.

**Include positive and/or negative words.** (Table 8) Prior work has found that including positive or negative sentiment increases message propagation (Milkman and Berger, 2012; Godes et al., 2005; Heath et al., 2001; Hansen et al., 2011). We measured the occurrence of positive and negative words as determined by the connotation lexicon of Feng et al. (2013) (better coverage than LIWC). Measuring the occurrence of both simultaneously was inspired by Riloff et al. (2013).

**Refer to other people (but not your audience).** (Table 9) First-person has been found useful for success before, but in the different domains of scientific abstracts (Guerini et al., 2012) and books (Ashok et al., 2013).

---

15 This is not an artificial restriction on our set of authors; a large follower count means (in principle) that our results draw on a large sample of decisions whether to retweet or not.

16 The tokens [at], [hashtag], [url] were ignored in the unigram-model case to prevent their undue influence, but remained in the bigram model to capture longer-range usage (“combination”) patterns.

---

17 To test whether the results stem from similarity to news rather than headlines per se, we constructed a NY′T-text LM, which proved less effective. We also tried using Gawker headlines (often said to be attention-getting) but pilot studies revealed insufficient vocabulary overlap with our TAC pairs.
We group the features introduced in §5.1 into 16 lexicon-based features (Table 3, 8, 9, 10), 9 informativeness features (Table 4), 6 language model features (Table 5, 6), 6 rt score features (Table 7), and 2 readability features (Table 11). We refer to all 39 of them together as custom features. We also consider tagged bag-of-words (“BOW”) features, which includes all the unigram (word:POS pair) and bigram features that appear more than 10 times in the cross-validation data. This yields 3,568 unigram features and 4,095 bigram features, for a total of 7,663 so-called 1,2-gram features. Values for each feature are normalized by linear transformation across all tweets in the training data to lie in the range [0, 1]\(^{18}\)

For a given TAC pair, we construct its feature vector as follows. For each feature being considered, we compute its normalized value for each tweet in the pair and take the difference as the feature value for this pair. We use L2-regularized logistic regression as our classifier, with parameters chosen by cross validation on the training data. (We also experimented with SVMs. The performance was very close, but mostly slightly lower.)

A strong non-TAC alternative, with social information and timing thrown in. One baseline result we would like to establish is whether the topic and author controls we have argued for, while intuitively compelling for the purposes of trying to determine the best way for a given author to present some fixed content, are really necessary in practice. To test this, we consider an alternative binary L2-regularized logistic-regression classifier that is trained on unpaired data, specifically, on the collection of 10,000 most retweeted tweets (gold-standard label: positive) plus the 10,000 least retweeted tweets (gold-standard label: negative) that are neither retweets nor replies. Note that this alternative thus is granted, by design, roughly \(\text{twice}\) the training instances that our classifiers have, as a result of having roughly the same number of tweets, since our instances are pairs. Moreover, we additionally include the tweet author’s follower count, and the day and hour of posting, as features. We refer to this alternative classifier as \(\neg \text{TAC}\neg \text{ff}\neg \text{time}\). (Mnemonic: “f” is used in bibliographic contexts as an abbreviation

| Table 8: Sentiment (contrast is measured by presence of both positive and negative sentiments). |
|---|---|
| effective? | author-preferred? |
| positive | ↑↑ * | YES (52%) |
| negative | ↑↑ * | YES (51%) |
| contrast | ↑↑ * | YES (52%) |

| Table 9: Pronouns. |
|---|---|
| effective? | author-preferred? |
| 1st person singular | | YES (51%) |
| 1st person plural | | YES (52%) |
| 2nd person | | YES (57%) |
| 3rd person singular | ↑↑ | YES (55%) |
| 3rd person plural | ↑ | YES (58%) |

| Table 10: Generality. |
|---|---|
| effective? | author-preferred? |
| indefinite articles (a,an) | ↑↑ * | YES (52%) |
| definite articles (the) | | YES (52%) |

| Table 11: Readability. |
|---|---|
| effective? | author-preferred? |
| reading ease | ↑ | YES (52%) |
| negative grade level | ↑ | YES (52%) |

Generalitiy helps. (Table 10) DCKL’12 posited that movie quotes are more shared in the culture when they are general enough to be used in multiple contexts. We hence measured the presence of indefinite articles vs. definite articles.

The easier to read, the better. (Table 11) We measure readability by using Flesch reading ease (Flesch, 1948) and Flesch-Kincaid grade level (Kincaid et al., 1975), though they are not designed for short texts. We use negative grade level so that a larger value indicates easier texts to read.

Final question: Do authors prefer to do what is effective? Recall that we use binomial tests to determine author preference for applying a feature more in \(t_2\). Our preference statistics show that author preferences in many cases are aligned with feature efficacy. But there are several notable exceptions: for example, authors tend to increase the use of @-mentions and 2nd person pronouns even though they are ineffective. On the other hand, they did not increase the use of effective ones like proper nouns and numbers; nor did they tend to increase their rate of sentiment-bearing words. Bearing in mind that changes in \(t_2\) may not always be intended as an effort to improve \(t_1\), it is still interesting to observe that there are some contrasts between feature efficacy and author preferences.

5.2 Predicting the “better” wording

Here, we further examine the collective efficacy of the features introduced in §5.1 via their performance on a binary prediction task: given a TAC pair \((t_1, t_2)\), did \(t_2\) receive more retweets?

Our approach. We group the features introduced in §5.1 into 16 lexicon-based features (Table 3, 8, 9, 10), 9 informativeness features (Table 4), 6 language model features (Table 5, 6), 6 rt score features (Table 7), and 2 readability features (Table 11). We refer to all 39 of them together as custom features. We also consider tagged bag-of-words (“BOW”) features, which includes all the unigram (word:POS pair) and bigram features that appear more than 10 times in the cross-validation data. This yields 3,568 unigram features and 4,095 bigram features, for a total of 7,663 so-called 1,2-gram features. Values for each feature are normalized by linear transformation across all tweets in the training data to lie in the range [0, 1]\(^{18}\)

For a given TAC pair, we construct its feature vector as follows. For each feature being considered, we compute its normalized value for each tweet in the pair and take the difference as the feature value for this pair. We use L2-regularized logistic regression as our classifier, with parameters chosen by cross validation on the training data. (We also experimented with SVMs. The performance was very close, but mostly slightly lower.)

| Table 10: Generality. |
|---|---|
| effective? | author-preferred? |
| indefinite articles (a,an) | ↑↑ * | YES (52%) |
| definite articles (the) | | YES (52%) |

| Table 11: Readability. |
|---|---|
| effective? | author-preferred? |
| reading ease | ↑ | YES (52%) |
| negative grade level | ↑ | YES (52%) |

We also tried normalization by whitening, but it did not lead to further improvements.
Figure 2: Accuracy results. Pertinent significance results are as follows. In cross-validation, custom+1,2-gram is significantly better than ~TAC+ff+time (p=0) and 1,2-gram (p=3.8e-7). In heldout validation, custom+1,2-gram is significantly better than ~TAC+ff+time (p=3.4e-12) and 1,2-gram (p=0.01) but not unigram (p=0.08), perhaps due to the small size of the heldout set.

Baselines. To sanity-check whether our classifier provides any improvement over the simplest methods one could try, we also report the performance of the majority baseline, our request-for-sharing features, and our character-length feature.

Performance comparison. We compare the accuracy (percentage of pairs whose labels were correctly predicted) of our approach against the competing methods. We report 5-fold cross validation results on our balanced set of 11,404 TAC pairs and on our completely disjoint heldout data of 1,770 TAC pairs; this set was never examined during development, and there are no authors in common between the two testing sets.

Figure 2(a) summarizes the main results. While ~TAC+ff+time outperforms the majority baseline, using all the features we proposed beats ~TAC+ff+time by more than 10% in both cross-validation (66.5% vs 55.9%) and heldout validation (65.6% vs 55.3%). We outperform the average human accuracy of 61% reported in our Amazon Mechanical Turk experiments (for a different data sample); ~TAC+ff+time fails to do so.

The importance of topic and author control can be seen by further investigation of ~TAC+ff+time’s performance. First, note that it yields an accuracy of around 55% on our alternate-version-selection task even though its cross-validation accuracy on the larger most- and least-retweeted unpaired tweets averages out to a high 98.8%. Furthermore, note the superior performance of unigrams trained on TAC data vs ~TAC+ff+time — which is similar to our unigrams but trained on a larger but non-TAC dataset that included metadata. Thus, TAC pairs are a useful data source even for non-custom features. (We also include individual feature comparisons later.)

Informativeness is the best-performing custom feature group when run in isolation, and outperforms all baselines, as well as ~TAC+ff+time; and we can see from Figure 2(a) that this is not due just to length. The combination of all our 39 custom features yields approximately 63% accuracy in both testing settings, significantly outperforming informativeness alone (p<0.001 in both cases). Again, this is higher than our estimate of average human performance.

Not surprisingly, the TAC-trained BOW features (unigram and 1,2-gram) show impressive predictive power in this task: many of our custom features can be captured by bag-of-word features, in a way. Still, the best performance is achieved

---

19To construct this data, we used the same criteria as in written by authors with more than 5000 followers, posted within 12 hours, $n_2 - n_1 \geq 10 \text{ or } \leq -15$, and cosine similarity threshold value the same as in cap of 50 on number of pairs from any individual author.

20One might suspect that the problem is that ~TAC+ff+time learns from its training data to overly rely on follower-count, since that is presumably a good feature for non-TAC tweets, and for this reason suffers when run on TAC data where follower-counts are by construction non-informative. But in fact, we found that removing the follower-count feature from ~TAC+ff+time and re-training did not lead to improved performance. Hence, it seems that it is the non-controlled nature of the alternate training data that explains the drop in performance.
by combining our custom and 1,2-gram features together, to a degree statistically significantly better than using 1,2-gram features alone.

Finally, we remark on our Bonferroni correction. Recall that the intent of applying it is to avoid false positives. However, in our case, Figure 2(a) shows that our potentially “false” positives — features whose effectiveness did not pass the Bonferroni correction test — actually do raise performance in our prediction tests.

Size of training data. Another interesting observation is how performance varies with data size. For \( n = 1000, 2000, \ldots, 10000 \), we randomly sampled \( n \) pairs from our 11,404 pairs, and computed the average cross-validation accuracy on the sampled data. Figure 2(b) shows the averages over 50 runs of the aforementioned procedure. Our custom features can achieve good performance with little data, in the sense that for sample size 1000, they outperform BOW features; on the other hand, BOW features quickly surpass them. Across the board, the custom+1,2-gram features are consistently better than the 1,2-gram features alone.

Top features. Finally, we examine some of the top-weighted individual features from our approach and from the competing −TAC+ff+time classifier. The top three rows of Table 12 show the best custom and best and worst unigram features for our method; the bottom two rows show the best and worst unigrams for −TAC+ff+time. Among custom features, we see that community and personal language models, informativeness, retweet scores, sentiment, and generality are represented. As for unigram features, not surprisingly, “rt” and “retweet” are top features for both our approach and −TAC+ff+time. However, the other unigrams for the two methods seem to be a bit different in spirit. Some of the unigrams determined to be most poor only by our method appear to be both surprising and yet plausible in retrospect: “icymi” (abbreviation for “in case you missed it”) tends to indicate a direct repetition of older information, so people might prefer to retweet the earlier version; “thanks” and “sorry” could correspond to personal thank-yous and apologies not meant to be shared with a broader audience, and similarly @-mentioning another user may indicate a tweet intended only for that person. The appearance of [hashtag] in the best −TAC+ff+time unigrams is consistent with prior research in non-TAC settings (Suh et al., 2010 Petrović et al., 2011).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>best 15 custom</td>
<td>twitter bigram, length (chars), rt (the word), retweet (the word), verb, verb retweet score, personal unigram, proper noun, number, noun, positive words, please (the word), proper noun retweet score, indefinite articles (a,an), adjective</td>
</tr>
<tr>
<td>best 20 unigrams</td>
<td>rt, retweet, [num], breaking, is, win, never, ., people, need, official, officially, are, please, november, world, girl, !!!, god, new</td>
</tr>
<tr>
<td>worst 20 unigrams</td>
<td>: [at], icymi, also, comments, half, ?, earlier, thanks, sorry, highlights, bit, point, update, last, helping, peek, what, haven’t, debate</td>
</tr>
</tbody>
</table>

6 Conclusion

In this work, we conducted the first large-scale topic- and author-controlled experiment to study the effects of wording on information propagation.

The features we developed to choose the better of two alternative wordings posted better performance than that of all our comparison algorithms, including one given access to author and timing features but trained on non-TAC data, and also bested our estimate of average human performance. According to our hypothesis tests, helpful wording heuristics include adding more information, making one’s language align with both community norms and with one’s prior messages, and mimicking news headlines. Readers may try out their own alternate phrasings at [http://chenhaot.com/retweetedmore/](http://chenhaot.com/retweetedmore/) to see what a simplified version of our classifier predicts.

In future work, it will be interesting to examine how these features generalize to longer and more extensive arguments. Moreover, understanding the underlying psychological and cultural mechanisms that establish the effectiveness of these features is a fundamental problem of interest.

Acknowledgments. We thank C. Callison-Burch, C. Danescu-Niculescu-Mizil, J. Kleinberg, P. Mahdab, S. Mullainathan, F. Pereira, K. Raman, A. Swaminathan, the Cornell NLP seminar participants and the reviewers for their comments; J. Leskovec for providing some initial data; and the anonymous annotators for all their labeling help. This work was supported in part by NSF grant IIS-0910664 and a Google Research Grant.
References


Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media? In Proceedings of WWW.

Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. 2013. What’s in a name? Understanding the interplay between titles, content, and communities in social media. In Proceedings of ICWSM.

Annie Louis and Ani Nenkova. 2013. What makes writing great? First experiments on article quality prediction in the science journalism domain. Transactions of ACL.

Zongyang Ma, Aixin Sun, and Gao Cong. 2012. Will this #hashtag be popular tomorrow? In Proceedings of SIGIR.


Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruiliong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In Proceedings of EMNLP.

Daniel M. Romero, Chenhao Tan, and Johan Ugander. 2013. On the interplay between social and topical structure. In Proceedings of ICWSM.


