Of Pins and Tweets: Investigating how users behave across image- and text-based social networks

Raphael Ottoni †
rapha@dcc.ufmg.br

Diego Las Casas †
diegolascasas@ufmg.br

João Paulo Pesce †
jpesce@dcc.ufmg.br

Wagner Meira Jr. †
meira@dcc.ufmg.br

Christo Wilson ‡
cbw@ccs.neu.edu

Alan Mislove ‡
amislove@ccs.neu.edu

Virgilio Almeida †
virgilio@dcc.ufmg.br

† Universidade Federal de Minas Gerais, Brazil
‡ Northeastern University, USA

Abstract

Today, it is the norm for online social (OSN) users to have accounts on multiple services. For example, a recent study showed that 34% of all Twitter users also use Pinterest. This situation leads to interesting questions such as: Are the activities that users perform on each site disjoint? Alternatively, if users perform the same actions on multiple sites, where does the information originate? Given the interlinking between social networks, failure to understand activity across multiple sites may obfuscate the true information dissemination dynamics of the social web.

In this study, we take the first steps towards a more complete understanding of user behavior across multiple OSNs. We collect a sample of over 30,000 users that have accounts on both Twitter and Pinterest, crawling their profile information and activity on a daily basis for a period of almost three months. We develop a novel methodology for comparing activity across these two sites. We find that the global patterns of use across the two sites differ significantly, and that users tend to post items to Pinterest before posting them on Twitter. Our findings can inform the understanding of the behavior of users on individual sites, as well as the dynamics of sharing across the social web.

1 Introduction

Online social networks (OSNs) are now a popular way for individuals to connect, communicate, and share content; many now serve as the de-facto Internet portal for millions of users. Because of their massive popularity, data about the users and their communication offer unprecedented opportunities to examine how human society functions at scale.

In recent years, the number and diversity of social networks have both increased beyond traditional OSNs like Facebook; popular examples include sites like Tumblr (short-form blogging), Pinterest (sharing image collections), and FourSquare (sharing user locations). In fact, today, many users have active accounts on multiple such sites (*e.g.,* it has recently been shown that over 34% of Twitter users also have a Pinterest account (Pew Research Center 2013)). While there have been many studies that have examined user behavior on these sites, it remains unknown how users distribute their time and activity *across* multiple sites. For example, do users share the same kinds of content on multiple sites? Are there sites where content tends to originate?

In this paper, we take a closer look at *cross-OSN user behavior*, focusing on the popular sites Twitter and Pinterest. Pinterest is a photo-sharing web site that allows users to share collections ("boards") of photos with others. Pinterest is primarily designed for users to share ("pin") images from other Web sites (unlike sites like Flickr, which are centered around having users upload their own photos). Like Twitter, Pinterest users can "follow" others, meaning the other user's pinned photos will show up in a feed when the user logs in.

Our goals are to understand how user activity is correlated across these two sites. First, we aim to understand whether users share a common *identity* across these sites. Second, we investigate the interests of users on these two different platforms. In Pinterest, user interest is conveyed by the images a user pins and repins in different boards, each of which is tagged with a category. In Twitter, user interest can be deduced from the words in their tweets. Questions that we aim to answer include: where does user interest manifest first, Twitter or Pinterest? Does a user share the same interests in both networks? How does one compare the interests represented by images versus the interests found in tweets?

We face a number of challenges in comparing activity on Twitter and Pinterest. Unlike Twitter, Pinterest is challenging to study, as the content shared is images (as opposed to text) and there is no official API for gathering data. We develop techniques for gather Pinterest data at scale, and collect a set of 30,000 users who have both Pinterest and Twitter accounts. We download all pinned photos and tweets of these users, resulting in a data set of over 2.9 million pins and 7.1 million tweets. Additionally, we develop a novel methodology that is able to compare image-based content (Pinterest) with text-based content (Twitter). We do so using *categories* of content, based on the categories provided by Pinterest; we demonstrate that our approach to categorization on Twitter shows high accuracy on many different types of tweets.

Overall, our analysis makes three contributions: First, we aim to understand the macro-scale patterns of activity across these sites. We find that, despite considering the same set of users on both sites, we see remarkably different global patterns of activity. Second, we use our categorization approach to study how users distribute their activity across the sites; we find that users engage in more categories of content on Pinterest than Twitter, but Twitter categories have greater predictive power. Finally, we explore the pollination of content from one site to the other and find that new content tends to originate on Pinterest before spreading to Twitter. This result suggests that while Twitter is an incredibly popular global communication platform, sites like Pinterest play a crucial role in the generation of new ideas and content.

2 Background

2.1 Pinterest

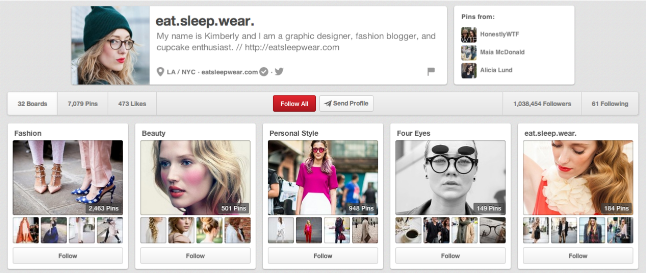
Pinterest is a pinboard-style image sharing social network designed to let users collect and share images and videos in an organized, categorized way. Pinterest was founded in 2010, and boasts a user population of 70 million as of July 2013.[1] It is currently the world's 27th most popular website.[2] and is the fastest growing OSN in both unique visitors and clicks via search engines (Walker 2012) Pinterest users are 80% female (Chafkin 2012; Ottoni et al. 2013) and the average monthly usage time per visitor is 98 minutes, which makes Pinterest the second most used OSN behind Facebook.[3] Pinterest has become especially important as a driver for e-commerce traffic: a recent study showed that Pinterest users are worth almost twice as much as Facebook users.[4]

On Pinterest, users create a personal profile as well as one or more *boards*. Each board is given a name and a description (both freeform text fields) as well as a category. Pinterest pre-defines 33 categories, varying from "Women's Fashion" and "Hair Beauty" to "Geek" and "Tattoos". The basic units of data on Pinterest are the images and videos users *pin* to their boards. Each pin is characterized by a relative timestamp (*e.g.,* "posted 22 hours ago", "posted 5 days ago"), a description (freeform text), and a link to the source of the content (if it originated from a third-party website). The only interactions supported by Pinterest are *repins*, comments on pins, and *likes*. By default, all data on Pinterest is public.

The organization and functions of the Pinterest social network are similar to Twitter. Personal profiles on Pinterest include a profile image, a brief self-description, and lists of the user's boards, pins, likes, followers, and friends (*i.e.,* those who the user follows). Figure 1 shows a typical user profile on Pinterest. Like Twitter, Pinterest users are presented with a chronologically-ordered *timeline* of pins from their friends upon logging in. Unlike Twitter, Pinterest users may follow other users, or follow specific boards. In the former case, all of the friend's pins appear in the timeline; in the latter case only pins from the specific board appear.

2.2 Twitter

Twitter is the de-facto standard for micro-blogging in most of the world. Founded in 2006, Twitter currently has 645 million users who generate over 500 million 140-character *tweets* per day.[5] Twitter is now the 11th most popular website in the world.[2] On Twitter, users create personal profiles that include a profile image, a brief self-description, and lists of the user's tweets, followers, and friends. Each tweet is characterized by an exact timestamp, and may include embedded images, URLs, #hashtags, @mentions, or geotagged coordinates. Users can *retweet* messages from other users to their own followers. Although Twitter allows users to set their profiles as "protected" (in which case only approved people may follow the account and view its tweets), the vast majority of users accept the default policy in which all data are public (Wasserman 2012).



eat.sleep.wear.

My name is Kimberly and I am a graphic designer, fashion blogger, and cupcake enthusiast. // http://eatsleepwear.com

Figure 1: Pinterest profile of a famous designer/blogger.

3 Dataset and Methodology

In this section, we present the dataset we will be using in our analysis. First, we introduce our raw data and describe how it was collected. Second, we briefly compare the users in our dataset against random samples of Pinterest and Twitter users, to quantify any bias in our target population. Third, we describe our methodology for labeling tweets with high-level *categories* (*e.g.,* design, fashion, politics). This categorization facilitates our comparison of Pinterest and Twitter in § 4, since all pins are categorized by default. Finally, we describe how we group tweets and pins into *sessions*, which represent a sequential period of user activity.

3.1 Dataset and Collection Methodology

The goal of our work is to compare and contrast the behavior of users across Pinterest and Twitter. In order to perform this comparison, we need to locate users that have accounts on *both* OSNs. Fortunately, there is a straightforward way to identify such users: when Pinterest first began accepting users, the only way to create an account was log-in with a pre-existing Facebook or Twitter account. Between August 21th and October 9th, 2012, we crawled 2 million users from Pinterest, of which 210,426 had signed-up with their Twitter account (Ottoni et al. 2013). Of these users, 76% have either protected or closed their Twitter account, leaving us with 50,702 users for our study. For clarity, we will refer to these users as the *selected population*.

To analyze user behavior, we need to collect pins and tweets generated by users in the selected population. To gather tweets, we can simply use the Twitter REST API v1.1. However, collecting pins requires addressing two technical challenges: 1) each pin must be gathered within 24-hours after it was generated, and 2) each pin must be gath-

[1] http://bit.ly/1jfjRHL

[2] http://bit.ly/1ksdYHv

[3] http://bit.ly/1dxEJSX

[4] http://bit.ly/OFmPrj

[5] http://bit.ly/18KIwd2

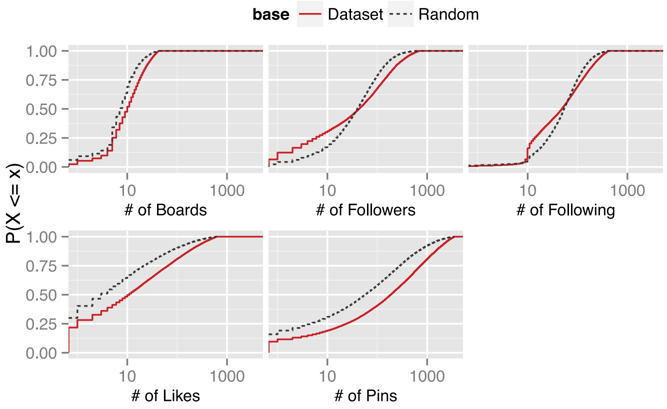


Figure 2: Comparison of our target Pinterest users to a random sample of Pinterest users.

ered individually. The 24-hour requirement stems from the fact that Pinterest displays relative timestamps with decreasing resolution, *i.e.,* a pin from an hour ago will display "posted 1 hour ago," whereas a pin from yesterday will display "posted 1 day ago." Our study necessitates high-resolution timestamps, thus we need to crawl pins while the timestamps are displayed in hours. The second challenge occurs because Pinterest does not have a public API for gathering pins with timestamps in bulk.

To address these challenges, we built a distributed crawler that collected pins from the selected users every day. The Pinterest website is a complex Javascript application, so we leveraged PhantomJS to crawl the site. PhantomJS is a headless version of the WebKit browser engine that can be automated using Javascript. Our crawler recorded the daily pin activity from each user. Unfortunately, it was impossible to perform daily crawls on all 50K users in the selected population due to Pinterest's rate limits. To work around this issue, we focused our crawls on random sample of 30K users from the selected population. We refer to these users as the *sampled population*.

Final Dataset Our final dataset consists of just under three months of pins and tweets from the sampled population. We actively crawled Pinterest between September 12 and December 3, 2013; since we only have accurate timestamps for pins collected during this period, we constrain our focus to tweets that were also generated during this time period. Table 1 shows the number of users from the sampled population that performed specific actions (*i.e.,* pinning or tweeting) during the time window; we observed 23,313 users who had at least one activity. We refer to these 23K users as the *target population*, and we focus on them for the remainder of this study.

	Action	Users	Pins	Tweets
Ignored	No activity	6,687	—	—
Target population	Pinned	15,329	1,778,798	—
	Tweeted	19,107	—	4,427,040
	Pinned & Tweeted	11,123	1,299,266	2,746,472

Table 1: Our final 82-day (09/12/13 to 12/03/13) dataset, comprising our *target population*.

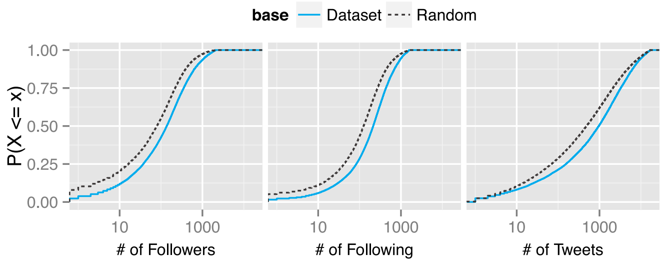


Figure 3: Comparison of our target Twitter users to a random sample of Twitter users.

3.2 Dataset Validation

We now present a brief comparison of the characteristics of users in the target population versus random samples of Pinterest and Twitter users. The purpose of this comparison is to quantify any bias in our target population. To enable this comparison, we selected 30K Pinterest users uniformly at random from our original sample of 2 million Pinterest users. Similarly, we selected 30K Twitter users uniformly at random from the "gardenhose" feed (10K each from September, October, and November 2013).

Figure 2 compares the characteristics of our target and random Pinterest users. The target users are slightly more active: they have more boards, likes, and pins. This suggests that the target users are more active social networkers, which is not surprising given that the target users have accounts on Pinterest and Twitter. Despite this, both populations have similar average numbers of friends and followers.

Figure 3 compares the characteristics of our target and random Twitter users. Much like on Pinterest, we observe that the target users generate slightly more tweets than the random sample. In this case, the target users also have slightly more friends and followers, further confirming that the target users are more active than random OSN users.

Overall, Figures 2 and 3 reveal that there are minor differences between the target and random user populations. However, despite this divergence, we believe that the target population is broadly representative of the active user community on Pinterest and Twitter.

3.3 Labeling Tweets with Categories

The next step in our methodology is to classify tweets from our target users with category labels. This step is necessary in order to make pins and tweets directly comparable: on Pinterest, all pins fall into 1 of 33 pre-defined categories, while tweets are freeform text. Figure 4 depicts the novel process we developed to categorize tweets.

Selecting Categories The first step in our classification process is to select relevant categories. We use the 33 categories provided by Pinterest as a natural starting point,

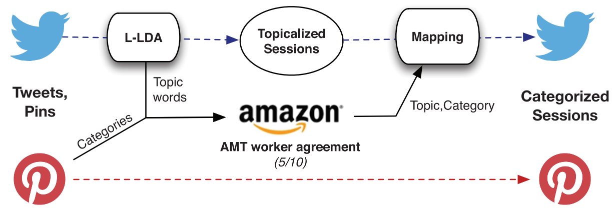


Figure 4: Diagram of our tweet categorization methodology.

and added the following five additional categories, since they have been commonly identified as important in Twitter-related studies (Duh et al. 2011; Hong and Davison 2010; Lee et al. 2011; Zhao et al. 2011): "Politics", "Religion", "Charity", "Twitter-Specific", and "Business". "Twitter-specific" refers to tweets that discuss Twitter culture, *e.g.,* #followfriday.

To determine whether these categories were sufficient, we conducted a brief pilot study on Amazon Mechanical Turk (AMT) where we asked AMT workers to label tweets with one of the aforementioned 38 categories. The results of this pilot study revealed some problems with the categories. For example, the AMT workers had difficulty labeling tweets as "Men's Fashion" or "Women's Fashion", so we merged these two categories together as "Fashion". Similarly, we merged "Health & Fitness" with "Sports", and "Home Decoration" with "DIY". We also generalized "Kids" to "Kids & Family". After this process, we were left with 35 categories.

Identifying Topics on Twitter The second step in our classification process is to extract *topics* from our corpus of tweets. The key idea here is to leverage well-known topic extraction tools as a stepping stone towards categorization: if we can extract topics from tweets, and map the topics into our 35 categories, then we can use the topic labels to also apply category labels to the tweets.

To extract topics from our tweets, we leverage Labeled Latent Dirichlet Allocation (L-LDA) (Ramage et al. 2009). L-LDA uses the same underlying mechanisms as LDA, however each topic is seeded with a label (*i.e.,* a word chosen by the researcher), to help anchor the topic extraction process. We chose L-LDA because prior work has shown it to be more effective than LDA at extracting topics from microblogs (Quercia, Askham, and Crowcroft 2012; Ramage, Dumais, and Liebling 2010).

We ran L-LDA on a random sample of 2 million tweets from our corpus. This was the maximum number of tweets we could process on a machine with 70GB of RAM. We preprocessed the tweets by removing the top 300 words (*i.e.,* the stop words), URLs, and all words that appeared less than 10 times. As shown in Figure 4, we parameterized the L-LDA algorithm with $\alpha = 0.167$, $\beta = 0.001$, and $k = 300$. We selected the 300 most common hashtags from our tweets as topic labels. The final output of L-LDA is 300 topics, each containing $\approx$50 words ranked by frequency in that topic.

Mapping Topics to Categories The third step in our classification process is to map the topics extracted from Twitter to our 35 categories. As shown in Figure 4, we leveraged AMT for this task. We presented each Turker with 20 of the topics extracted by L-LDA, and asked them to select the 3 most applicable categories for each topic. AMT workers were shown the top 25 words associated with each topic. We divided our $300/20$ topics into 15 tasks, and had 10 AMT workers complete each task.[6] AMT workers were paid $0.25 per task, and this process took 2 days to complete.

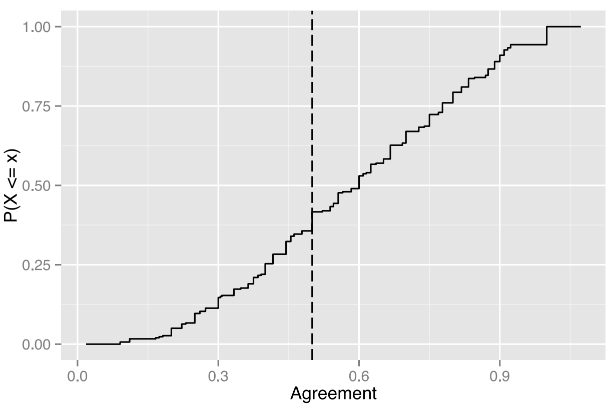


Figure 5: Agreement between AMT workers on the category label for each of the 300 Twitter topics.

After collecting the results from the AMT workers[7], we analyzed the data to determine which of the 300 L-LDA topics could be mapped with high confidence onto our 35 categories. Figure 5 shows the percentage of agreement between the AMT workers on the category for each topic, *i.e.,* whether they listed a particular category as one of their three options. 60% of the topics have $>$50% agreement between the 10 AMT workers, meaning we can say with high confidence that we know the best mapping of those topics to categories. In the remainder of our categorization methodology, we only use the topics that have $>$50% agreement; we discard the remaining low-agreement topics.

Tweet Categorization The fourth step of our classification process is to calculate the most likely category for each tweet based on its topic distribution. Initially, we attempted to use LDA to generate the topic distribution for each tweet (the output of LDA labeling is a topic distribution, representing the likelihood that a given document contains each topic). However, we had three independent raters examine a random sample of the labeled tweets, and discovered that the accuracy of the topic labels was $\approx$30%. This low accuracy stems from the fact that many tweets are very short, and some cannot be interpreted when removed from their conversation context.

To remedy this situation, we group tweets into *sessions*. We define a session as sequentially-generated tweets from a single user, where the time-gap between any two successive

[6] An example task can be found here: `http://dcc.ufmg.br/~rapha/ofpinsandtweets/categorization/?s=1`

[7] The labeling dataset can be found here: `http://dcc.ufmg.br/~rapha/ofpinsandtweets/labelingdataset`

tweets is ≤2 hours. Intuitively, sessions group tweets that are generated in rapid succession, and are therefore likely to fall into the same topic. We chose 2 hours as the time delimiter between sessions on Twitter because 2 hours is the smallest unit of time that we can accurately measure on Pinterest (recall that timestamps on Pinterest are shown in increments of hours, *e.g.,* "posted 4 hours ago"). In § 3.4, we will also group pins into sessions, so that our categorized data from Pinterest and Twitter are directly comparable.

We can now assign a category to each Twitter session by generating a topic distribution for the session. We treat each Twitter session as a document and use LDA to generate the topic distributions. The results from the AMT workers give us a mapping between L-LDA topics and categories, however this is a many-to-one mapping (*i.e.,* many topics may map to the same category). Thus, to determine the category of a session, we subtract $1/number of topics$ (*i.e.,* the baseline probability) from the probability of each topic, and then sum the positive results of all the topics that map to each category. We then assign the category with the highest aggregate probability to each session.

As a final step, we manually validated the categories assigned to 100 random Twitter sessions. As before, three independent raters examined each session to determine the appropriateness of the category. Of the 35 categories, 11 were judged as having accuracy >60% (*i.e.,* the majority of the time, the category accurately described the tweets in the session). These 11 categories are: "Art," "Charity," "Design," "DIY & Crafts," "Fashion," "Food & Drink," "Hair & Beauty," "Health, Fitness & Sports," "Politics," "Technology," and "Weddings". Thus, in § 4, we focus on Twitter sessions from these 11 categories, and ignore all sessions from other categories.

We make our code and data for classifying Twitter sessions into topics available to the research community.[8]

3.4 Sessions on Pinterest and Twitter

At this point, we have grouped tweets from our target users into sessions and categorized each session. After filtering out the sessions with low-accuracy categories, we are left with 100,212 sessions (out of 665,980 total sessions). To ensure that our Pinterest and Twitter data remain comparable, we also grouped pins from the target users into 2 hour sessions. The category for a Pinterest session is simply the most frequent category among the pins in that session.

Figure 6 shows the number of categorized sessions per user in our dataset. Our users tend to have fewer categorized Twitter sessions than Pinterest sessions; this is understandable, given that we have filtered out Twitter sessions that belong to low-accuracy categories. The most important line in Figure 6 is "Both," which plots the number of users with at least k categorized sessions on *both* OSNs. Since our goal is to examine user behavior across Pinterest and Twitter, we must focus on the subset of users who are active on both OSNs. Thus, in § 4, we focus on the users who have ≥7 categorized sessions on both Pinterest and Twitter. In total, this dataset includes 41,928 Pinterest sessions and 33,452 Twitter sessions generated by 1,227 users.

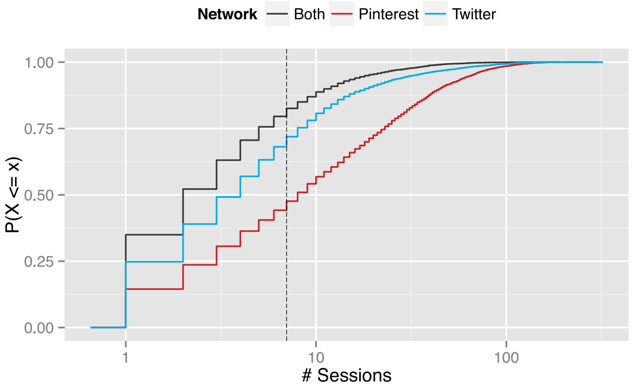


Figure 6: Number of categorized sessions per user by network.

4 Analysis

We now present our analysis on cross-user behavior on Twitter and Pinterest. We first explore global patterns of user behavior (§ 4.1) on the two sites, followed by an examination of the similarity in user attributes across the sites (§ 4.2). We then turn to investigate the content shared on these two sites, based on the categories we discussed in the previous section. We first look at global category popularity (§ 4.3) followed by pollination of content categories across sites (§ 4.4).

4.1 General and Temporal Analysis

We begin our analysis by examining the high-level characteristics of our target users on Pinterest and Twitter. In this section, we examine all 23K users, 1.8M pins, and 4.4M tweets in our 82 day dataset (see Table 1 for details).

First, we examine overall pinning and tweeting behavior. Figure 7 is a scatterplot showing the number of pins and tweets per user in our dataset. Black regions are occupied by ≈50 users, while light grey regions are occupied by <10 users.

Figure 7 reveals that most users in our target population tend to fall in one of two regions of the plot. There is a group of users along the axes who are active on either Pinterest or Twitter. Conversely, there is a group of users in the center of the plot that generate on the order of 100 pins and 100 tweets during our 82 day sample. This demonstrates that our target population does include many users who are equally active on both OSNs.

Next, we examine the temporal dynamics of Pinterest and Twitter users. Figure 8 plots a moving average of the number of pins and tweets per day between October 15 and November 15, 2013. Two conclusions can be drawn from Figure 8: first, our target users generate around two times as many tweets per day as pins. Second, Pinterest and Twitter exhibit different long-term temporal dynamics. Twitter use peaks during the week, while Pinterest use peaks later in the week, heading into the weekend. One deviation occurs the

[8] http://dcc.ufmg.br/~rapha/ofpinsandtweets

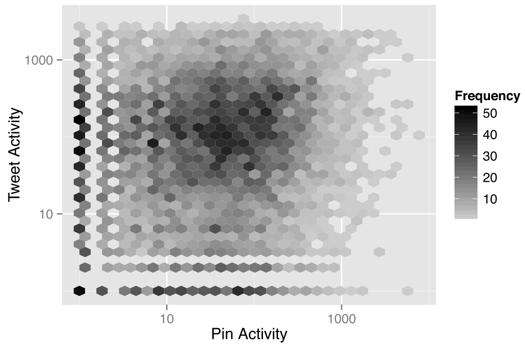


Figure 7: Scatter plot of activity by user in 82 days.

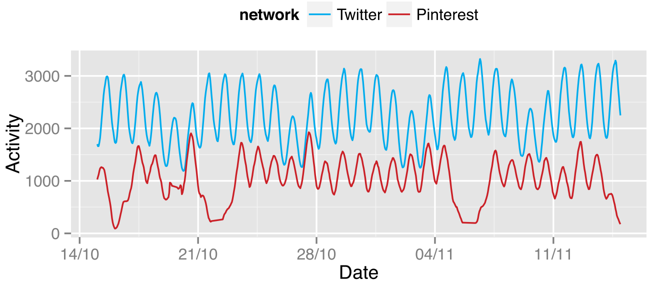


Figure 8: Average pins and tweets per day over one month.

week of October 28: we hypothesize that Pinterest saw increased activity this week due Halloween-related pins.

Figure 9 plots the daily activity patterns of our target users, broken down by individual days of the week. Our target users are primarily located in the U.S., Figure 9 is normalized to Eastern Standard Time. Unsurprisingly, the usage of both OSNs troughs in the early-morning. However, Twitter users frequently tweet late at night, and there is a pronounced reduction in tweeting on Saturday and Sunday (see Figure 8). In contrast, Pinterest use tends to slowly rise over the course of each day, especially on Sundays.

4.2 User and Linguistic Identity

In this section, we examine the personal profiles and linguistic features of our 23k target users. We pose the question: *do users have unique* identities *on Pinterest and Twitter, or do they share one identity across both OSNs*? To quantify user identity, we examine three textual features: usernames, descriptions in personal profiles, and words used in pin descriptions and tweets. We focus on textual features because prior work has been able to successfully measure individual emotional and cognitive factors using automated text analysis (Tausczik and Pennebaker 2009; Kahn et al. 2007; Veltman 2006).

First, we compare the usernames chosen by individuals on Pinterest and Twitter. Studies have shown that the usernames chosen by people on online systems are often reflective of their individual personalities, the way they want to be perceived by others, and their cultural environment (Suler 2002; Bechar-Israeli 1995). Thus, we hypothesize that if users choose the same username on Pinterest and Twitter, this indicates a homogeneous identity across both social platforms.

Figure 10(a) plots the Levenshtein Ratio (LR) between each user's username on Pinterest and Twitter. LR is defined as $LR(s_i, s_j) = 1 - ED(s_i, s_j)/\max(|s_i|, |s_j|)$, where s_i and s_j are strings, $ED(s_i, s_j)$ is the Levenshtein edit-distance between s_i and s_j, and $|s|$ is the length of string s. Intuitively, $ED(s_i, s_j) = 1$ means the strings are identical, while $ED(s_i, s_j) = 0$ means they are completely different. As shown in Figure 10(a), a significant number of our target users use the same, or very similar, usernames on both OSNs. This suggests that users share a common identity across Pinterest and Twitter.

Second, we compare the textual descriptions that user's write in their personal profiles. We analyze each description with LIWC (Pennebaker, Francis, and Booth 1999; Pennebaker et al. 2007), which produces a vector of linguistic attributes for each description. We then plot the cosine similarity between users' LIWC vectors on Pinterest and Twitter in Figure 10(b). Similar to Figure 10(a), the majority of users in our target population have the same, or very similar, linguistic attributes in descriptions across both OSNs. This reaffirms our finding that many users share the same personal identity on Pinterest and Twitter. However, unlike Figure 10(a), there are a significant fraction of users with zero similarity, indicating that some users may use Pinterest and Twitter for different purposes, or perhaps desire to separate their identities on these two social platforms.

Third, we compare the language used in pin descriptions and tweets. Although Pinterest is primarly on image-based OSN, users may write (typically brief) freeform textual descriptions for each pin. This leads to the question: *do users share the same linguistic style across pin descriptions and tweets?* To answer this question, we concatenated all of each users pin descriptions into a single document, and analyzed the document with LIWC. We also performed the same process for each user's tweets. This produces a pin-LIWC vector and a tweet-LIWC vector for each user. In order to compare each user's vectors to the language usage of the general population, we also created *network* LIWC vectors by gathering a random sample of pin descriptions/tweets, concatenating them into two respective documents, and analyzing these two documents with LIWC.

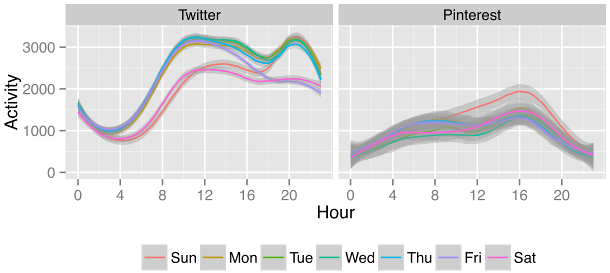


Figure 9: Daily activity on Pinterest and Twitter.

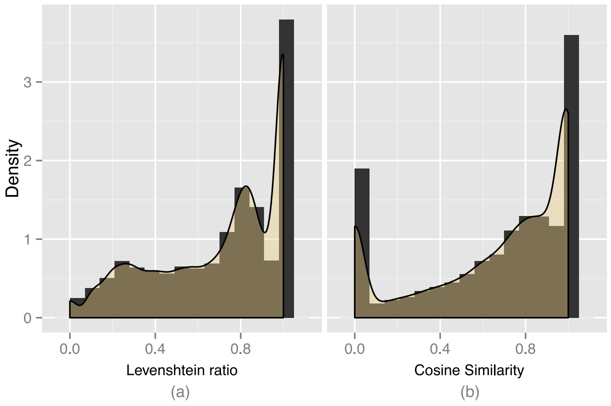


Figure 10: (a) The Levenshtein Ratio between each user's username on Pinterest and Twitter. (b) The cosine similarity between the LIWC vectors derived from each user's profile description on Pinterest and Twitter.

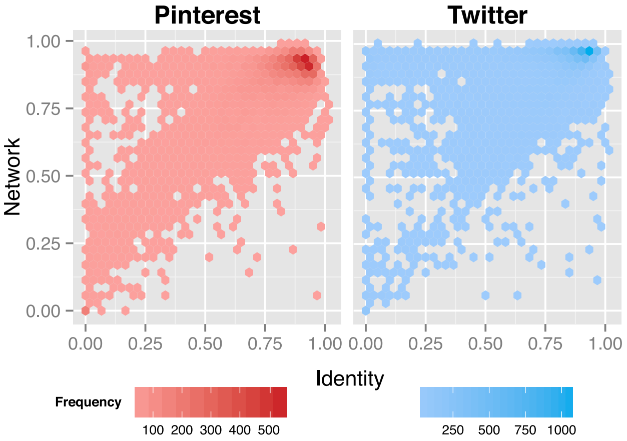


Figure 11: Linguistic identity of users across Pinterest and Twitter, versus similarity when compared to the linguistic norms of each respective OSN.

Figure 11 present the results of our pin and tweet LIWC analysis. Each of the two x-axes captures the same quantity: the cosine similarity between each user's Pinterest and Twitter LIWC-vector. We refer to this as the "Identity" axis. The y-axis of the Pinterest scatter plot captures the cosine similarity between each user's Pinterest LIWC-vector and the network LIWC-vector for Pinterest. The y-axis for the Twitter scatter plot is the same, except it substitutes the individual and network LIWC-vectors for Twitter. We refer to these as the "Network" axes. For example, a user at position (0, 0) in the Pinterest plot would have 1) zero similarity between the LIWC of their own pins and tweets, and 2) zero similarity between the LIWC of their pins and the pins of other Pinterest users. Conversely, a user at position (1,1) in the Pinterest plot would have 1) exact similarity between the LIWC of their own pins and tweets, and 2) exact similarity between the LIWC of their pins and the pins of other Pinterest users. The intensity of color in Figure 11 captures the number of users with a given set of similarity scores.

There are several takeaways from Figure 11. First, the majority of users cluster in the top-right of both plots, meaning that these individuals use similar language across both OSNs, and this language is similar to other users. This suggests that both OSNs have similar linguistic norms, and that most users adopt these conventions. Second, there are many users in the upper-left half of each graph, but not in the lower-right. These users tend to use different language constructs on Pinterest and Twitter. However, although these users diverge from themselves, they still tend to conform to the linguistic norms of the platform. This suggests that although an individual's linguistic identity may vary across OSNs, they still tend to adopt the conventions of each platform. Prior work has observed similar linguistic adaption on online forums (Danescu-Niculescu-Mizil et al. 2013).

4.3 Categories on Pinterest and Twitter

In this section, we analyze the characteristics of categories on Pinterest and Twitter. For this analysis, we leverage the 41,928 Pinterest and 33,452 Twitter categorized sessions introduced in § 3.4. For clarity, we organize this section around four high-level questions.

The first question we ask is very simple: *which categories are the most popular overall on Pinterest and Twitter?* To answer this, Figure 12 plots the number of sessions classified amongst the top 10 categories on Pinterest on Twitter. The top 2 categories ("Fashion" and "Food & Drink") are shared across both sites, while "Design, Decorations, & Crafts" and "Hair & Beauty" are more popular on Pinterest, and "Health & Sports" is more popular on Twitter. Although other categories, such as "Technology" and "Politics," are very popular amongst the general Twitter population, it is unsurprising that these categories are not as popular with our target users. Our target population intentionally contains many avid Pinterest users, which is reflected in the popularity of the design, food, and beauty categories.

The second question we ask is: *how do users distribute their content across categories?* Are users highly focused (*i.e.,* most of their content is in a few categories), or are users more varied? To answer this question, we calculate the *Shannon Entropy* of each user from the distribution of categories across their sessions. Shannon Entropy is defined as

$$H(p) = -\sum_{i=1}^{n} p_i \log(p_i) \tag{1}$$

where p is the user's distribution of sessions, n is the number of categories present in the user's sessions, and p_i is the probability that a given session from u will be of category i. A user specializing in a single category will have $H(p) = 0$, while a user whose sessions are evenly distributed across all categories will have maximum entropy.

Figure 13(a) shows the CDF distribution of *session entropy* for our target users. Figure 13(a) reveals that even though the Pinterest platform has a wider range of popular categories (see Figure 12), users tend to specialize more on Pinterest than on Twitter. This result seems to agree with the

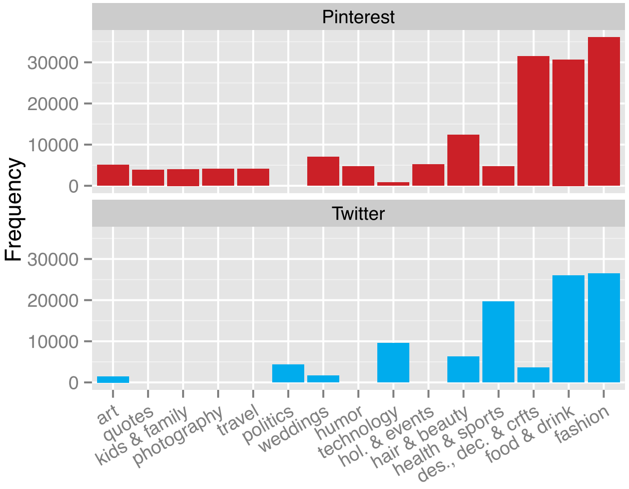


Figure 12: Popularity of the top 10 most frequent categories on Pinterest and Twitter.

prevailing understanding of Twitter as a general communication platform, versus Pinterest which is organized around particular interest groups (*e.g.,* fashion, food, photography).

The third question we ask stems naturally from our second question: *can you predict the categories of a user's tweets based on their pins, or vice-versa?* To answer this question, we use the metric *Cross Entropy*, defined as:

$$H(p, m) = -\sum_{i=1}^{n} p_i \log(m_i), \tag{2}$$

where p and m are the user's distributions of sessions either in Pinterest or Twitter. *Cross Entropy* is used to compare two probability distributions: if $p = m$ then $H(p) = H(p, m)$, therefore the closer the $H(p, m)$ is to the true entropy $H(p)$ (eq. 1), the better m is as an approximation of p. Intuitively, *Cross Entropy* scores closer to zero (*i.e.,* low entropy) indicate that the model distribution m is a good predictor of the observed distribution p.

Figure 13(b) plots the CDF of *Cross Entropy* for our target users. The red line represents $H(Pinterest, Twitter)$ (*i.e.,* Twitter is the model distribution), while the blue line represents $H(Twitter, Pinterest)$. Figure 13(b) reveals that for our users, Twitter categories are a better predictor of Pinterest categories than vice-versa. This finding may stem from the fact that the set of popular Twitter categories amongst our target users is more constrained than the set of Pinterest categories (see Figure 12), thus giving the Twitter categories more predictive power.

The final question we ask is: *does each user's session stream exhibit* locality of interest*?* A user with high locality engages in the same categories over many sequential sessions, while a user with low locality spreads their sessions over many different categories with random interleaving. To quantify locality of interest, we use the *stack distance* metric (Cascaval and Padua 2003; Almeida et al. 1996) to analyze each user's session timeline. A stack distance of zero indicates that sessions of each category appear close to each other in the timeline.

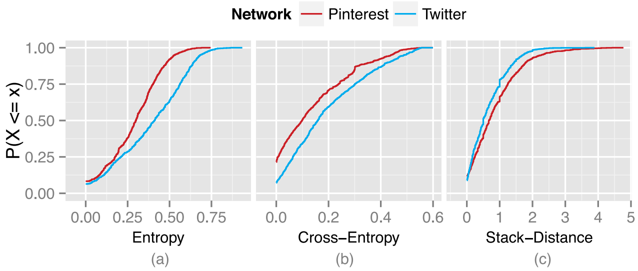


Figure 13: Distributions of *Entropy*, *Cross-entropy* and *Stack-distance* for each user's categorized sessions on Pinterest and Twitter.

Figure 13(c) plots the distribution of *stack distances* for our target users. Both OSNs show similar a similar trend, with >60% of users having stack distance ≤1. This shows that users on both OSNs tend to have many subsequent sessions about the same category of content. This finding suggests that not only do user's interests cross-over between OSNs, but that those interests tend to remain stable for relatively long periods of time.

4.4 Pollination of Categories

In this section, we examine the temporal flow of ideas from Pinterest to Twitter, and vice-versa. Specifically, we pose the question: *where do new ideas first appear: on Pinterest, or on Twitter?* This question is fundamental to the study of information dissemination across OSN boundaries. To answer this question, we again focus on the categorized session introduced in § 3.4.

To begin to address this question, Figure 14 plots a heatmap of transitions from category to category across all of our target users. For each user, we take all of their session on Pinterest *and* Twitter and put them into a single, chronological timeline. We then examine the transitions between categories and OSNs over time.

The red region of Figure 14 represents intra-Pinterest transitions, while blue is intra-Twitter, and grey is inter-OSN. The brighter the color, the more frequently we observe that transition. The transitions happen from column to row, *e.g.,* if the user's current session is about technology in Pinterest, the probability that the next session will be about technology in Twitter is ≈15%.

There are three noteworthy features of Figure 14. First, there are many self-transitions (*i.e.,* category c to category c) on both OSNs, as shown on the diagonal of the heatmap. This reinforces the finding from the *stack distance* plot (Figure 13(c)). Second, we observe many transitions from all categories into the most popular categories on Pinterest and Twitter, which manifest as the horizontal bands of color in the heatmap. Although transitions from one OSN into a popular category on the other OSN do occur, they are less frequent than intra-OSN transitions.

Third, there are specific inter-OSN self-transitions that

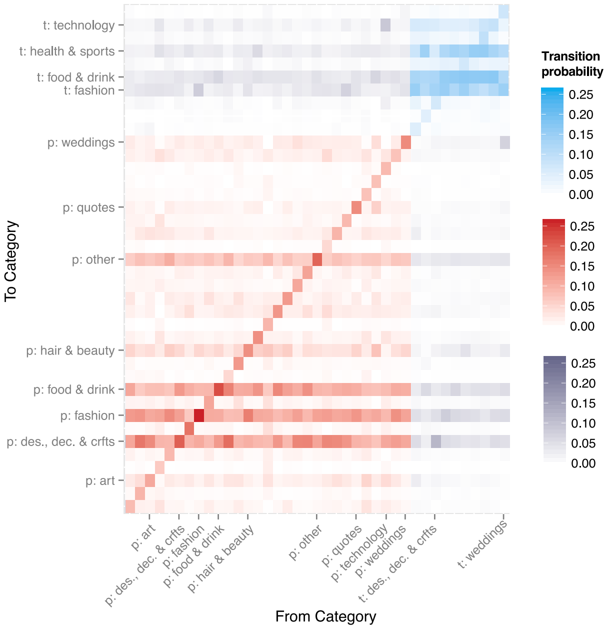


Figure 14: Heatmap of inter and intra-OSN category transitions.

occur more frequently than others. Examples include "Design, Decorations, & Crafts" and "Weddings" from Twitter to Pinterest, and "Fashion" and "Technology" from Pinterest to Twitter. These cases epitomize *cross-pollination*, or information transfer, from one OSN to another. One question that arises from Figure 14 is: *why are inter-OSN transitions less frequent than intra-OSN transitions?* In future work, we plan to examine whether the proliferation of mobile social apps plays are role in this, since it is more difficult to multi-task between apps on a smartphone than on a desktop.

Network Leaning Although Figure 14 shows which inter-OSN category transitions are common between Pinterest and Twitter, it does not tell us the strength of information transfer in either direction. In other words, *which OSN drives innovation: do ideas tend to originate on Pinterest and then move to Twitter, or vice-versa?*

To answer this question, we define a metric called *Cross Network Precedence* (CNP) as the number of times a session of category c on OSN X at start time t precedes another session of category c on OSN Y at time $t' > t$ for a specific user. Simply put, CNP counts the number inter-OSN self-transitions for each user. Now, we define the *Network Leaning* (NL) score of a user as the *relative difference* between the sum of all CNP of his/her sessions across the OSNs:

$$NL(u) = \frac{\sum_{i=1}^{S_P} CNP(i) - \sum_{i=1}^{S_T} CNP(i)}{\sum_{i=1}^{S_P} CNP(i) + \sum_{i=1}^{S_T} CNP(i)}, \qquad (3)$$

where S_P and S_T are the user's sessions in *Pinterest* and *Twitter*, respectively. NL varies from -1 to 1, with values close to -1 indicating stronger influence of Twitter over Pinterest, while values close to 1 indicate the contrary.

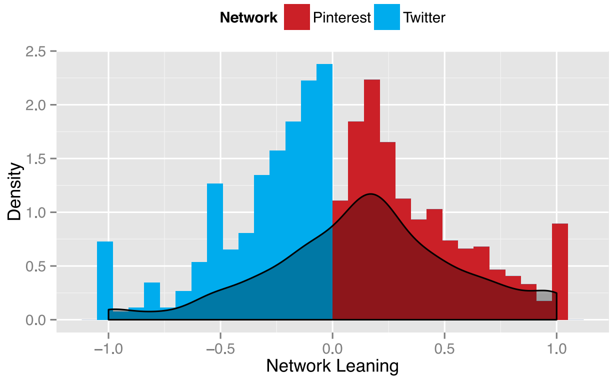


Figure 15: Histogram of Network Leaning scores for our target users. (*mean* = 0.15, *median* = 0.16, *skew* = -0.2)

Figure 15 plots a histogram of the NL scores for our target population. The majority of users fall somewhere in the middle of the range (*i.e.,* mild preference for one OSN over the other), while a defiant subset of users have -1 and 1 scores (*i.e.,* strong preference for one OSN over the other). However, the mean of the NL scores is 0.15 and the median is 0.16 (with skew of -0.2), demonstrating that overall, users tend to start generating content on Pinterest, and then transition to Twitter. This finding demonstrates that although Twitter may have a large, vibrant user community, smaller OSNs like Pinterest play a key role in fostering new ideas, which can then transition onto and trend on Twitter.

5 Related work

We now briefly summarize existing studies of cross-OSN user behavior, as well as studies of user activity on Twitter and Pinterest. The most related work to ours explores how users curate content in Pinterest and Last.fm (Zhong et al. 2013). While the authors focus on the popularity of items that users choose, we focus on how users post across multiple sites; thus, the two studies are highly complementary.

There have been a number of approaches to categorizing Twitter content, *e.g.,* (Duh et al. 2011; Hong and Davison 2010). Twitter itself categorizes its trending hashtags/topics as *trends* in reviews published yearly. For example, the 2012 Review[9] labels classes of hashtags as politics, sport, tech, and food. Unfortunately, Twitter's classification methodology is not public. Other work (Lee et al. 2011) proposes a method to map Twitter's *trending topics* to higher level categories. The result is similar to our methodology, with most of the categories overlapping with our Pinterest-based categories. In the domain of news, Zhao et al. (Zhao et al. 2011) topicalize tweets based on news labels from the New York Times, with some additions of their own. Finally, other work (Hong and Davison 2010) uses categories from Twitter Suggestions, the former official user recommendation tool; almost all of these categories have an equivalent in Pinterest.

[9] https://2012.twitter.com/en/trends.html

As Pinterest has grown, there have been a number recent studies (*e.g.,* (Feng et al. 2013)) that focus on quantifying and analyzing Pinterest user behavior. Recent work (Zoghbi, Vulić, and Moens 2013) verifies that the images posted to sites like Pinterest are an accurate reflection of a user's interest, and can be used to recommend relevant products. Other work (Ottoni et al. 2013) has shown that behavior on Pinterest differs significantly by gender.

Finally, the use of AMT to obtain ground truth for experiments has been examined by a number of studies in the past. For example, recent work (Komarov, Reinecke, and Gajos 2013) studies different setting designs in AMT and compares the results with lab-based settings. They find no significant differences between lab and AMT experiments, suggesting that using AMT for ground truth is likely to yield high accuracy.

6 Concluding Discussion

Today, online social networks (OSNs) are extremely popular; many users actively maintain accounts of a variety of sites. While these sites have enabled significant research into the functioning of society at scale, most prior work has focused on user activity within a single site. It remains unclear how the OSN ecosystem fits together, what roles the various sites play, and how users are choosing to distribute their activities across the wide variety of sites today.

We took a first step towards answering these questions in this paper, focusing on users who have accounts on both Twitter and Pinterest. We observe that even though many users seem to maintain a single identity and interests across the two sites, they show markedly different global patterns of activity. Using a novel tweet categorization methodology, we investigate how users distribute their content across the two sites; we find that users tend to engage in more categories of content on Pinterest than on Twitter, and (crucially) that new content tends to germinate on Pinterest, then transfer to Twitter. This underscores the notion that while Twitter may be an incredibly popular communication platform, smaller topic-specific sites like Pinterest play a key role in the generation of new ideas and content. However, because users' activities are more narrowly focused on Twitter, it can actually serve as a good predictor of activity on smaller OSNs like Pinterest.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This research was supported in part by NSF grants CNS-1054233 and CNS-1319019, ARO grant W911NF-12-1-0556, and an Amazon Web Services in Education Grant.

References

Almeida, V.; Bestavros, A.; Crovella, M.; and De Oliveira, A. 1996. Characterizing reference locality in the www. In *TPDS*.

Bechar-Israeli, H. 1995. From <bonehead> to <clonehead>: Nicknames, play, and identity on internet relay chat. *Journal of Computer-Mediated Communication* 1(2).

Cascaval, C., and Padua, D. A. 2003. Estimating cache misses and locality using stack distances. ACM.

Chafkin, M. 2012. Can ben silbermann turn pinterest into the world's greatest shopfront? `http://bit.ly/OgmHYl`.

Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *WWW*.

Duh, K.; Hirao, T.; Kimura, A.; and Ishiguro, K. 2011. Creating Stories: Social Curation of Twitter Messages. In *ICWSM*.

Feng, Z.; Cong, F.; Chen, K.; and Yu, Y. 2013. An empirical study of user behaviors on pinterest social network. In *WI-IAT*.

Hong, L., and Davison, B. D. 2010. Empirical study of topic modeling in Twitter. In *SOMA*.

Kahn, J. H.; Tobin, R. M.; Massey, A. E.; and Anderson, J. A. 2007. Measuring emotional expression with the linguistic inquiry and word count. *The American Journal of Psychology* 120(2).

Komarov, S.; Reinecke, K.; and Gajos, K. Z. 2013. Crowdsourcing performance evaluations of user interfaces. In *CHI*.

Lee, K.; Palsetia, D.; Narayanan, R.; Patwary, M. M. A.; Agrawal, A.; and Choudhary, A. 2011. Twitter Trending Topic Classification. In *ICDM Workshop*.

Ottoni, R.; Pesce, J. P.; Casas, D. L.; Franciscani Jr., G.; Meira Jr., W.; Kumaraguru, P.; and Almeida, V. 2013. Ladies first: Analyzing gender roles and behaviors in pinterest. In *ICWSM*.

Pennebaker, J. W.; Chung, C. K.; Ireland, M.; Gonzales, A.; and Booth, R. J. 2007. The development and psychometric properties of liwc2007 the university of texas at austin. *Development* 1(2).

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 1999. *Linguistic Inquiry and Word Count*. Lawrence Erlbaum.

Pew Research Center. 2013. Social media update. `http://bit.ly/1fXokJy`.

Quercia, D.; Askham, H.; and Crowcroft, J. 2012. TweetLDA: Supervised Topic Classification and Link Prediction in Twitter. In *WebSci*.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*.

Ramage, D.; Dumais, S.; and Liebling, D. 2010. Characterizing Microblogs with Topic Models. In *ICWSM*.

Suler, J. R. 2002. Identity management in cyberspace. *Journal of Applied Psychoanalytic Studies* 4(4).

Tausczik, Y. R., and Pennebaker, J. W. 2009. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1).

Veltman, B. R. 2006. *Linguistic analysis of the semantic content of the Rorschach Inkblot Test.* Ph.D. Dissertation.

Walker, T. 2012. State of us internet in q1 2012. `http://bit.ly/V9L10Y`.

Wasserman, T. 2012. Twitter user id numbers cross into the billions. `http://on.mash.to/1c5ifeM`.

Zhao, W. X.; Jiang, J.; Weng, J.; He, J.; and Lim, E.-p. 2011. Comparing Twitter and Traditional Media Using Topic Models. In *ECIR*, 338–349.

Zhong, C.; Shah, S.; Sundaravadivelan, K.; and Sastry, N. 2013. Sharing the Loves: Understanding the How and Why of Online Content Curation. In *ICWSM*.

Zoghbi, S.; Vulić, I.; and Moens, M.-F. 2013. I pinned it. where can i buy one like it?: Automatically linking pinterest pins to online webshops. In *DUBMOD*, 9–12. New York, NY, USA: ACM.