"I Need to Try This!": A Statistical Overview of Pinterest

Eric Gilbert, Saeideh Bakhshi
School of Interactive Computing
Georgia Institute of Technology
[gilbert, sbakhshi]@cc.gatech.edu

Shuo Chang, Loren Terveen
Dept. of Computer Science & Engineering
University of Minnesota
[schang, terveen]@cs.umn.edu

ABSTRACT
Over the past decade, social network sites have become ubiquitous places for people to maintain relationships, as well as loci of intense research interest. Recently, a new site has exploded into prominence: *Pinterest* became the fastest social network to reach 10M users, growing 4000% in 2011 alone. While many Pinterest articles have appeared in the popular press, there has been little scholarly work so far. In this paper, we use a quantitative approach to study three research questions about the site. What drives activity on Pinterest? What role does gender play in the site's social connections? And finally, what distinguishes Pinterest from existing networks, in particular Twitter? In short, we find that being female means more repins, but fewer followers, and that four verbs set Pinterest apart from Twitter: *use*, *look*, *want* and *need*. This work serves as an early snapshot of Pinterest that later work can leverage.

Author Keywords
social network sites; cmc; pinterest; twitter

ACM Classification Keywords
H.5.3 [Group and Organization Interfaces]: Asynchronous interaction - Web-based interaction;

INTRODUCTION
Over the past decade, online social networks have become ubiquitous places for people to maintain and create social relationships, exchange messages with friends and family, share jokes and news, ask and answer questions, etc. Facebook is the largest online social network, with a staggering 1 billion registered users from all around the world. In addition, Twitter, Google+, and LinkedIn all have over 100 million users, and the two most popular Chinese networks—Weibo and Renren—each have over 300 million users.

Not surprisingly, social network sites have become a locus of intense interest for researchers from a wide range of disciplines. Using a variety of methods, researchers study issues such as whether social media traces can predict social network tie strength [16], and build and maintain social capital [7, 14]. They have also developed new techniques, including new approaches for recommending content [12] and friends [11].

In the past year, a new social network site has exploded into prominence: *Pinterest* became the fastest social network to reach 10 million users, growing 4000% in 2011 alone [26]. Pinterest revolves around the metaphor of a "pin board:" users "pin" photos they find on the web and organize them into topical collections, such as hobbies, sports, fashion, etc. Pinterest users—called pinners—can follow one another and also "re-pin", "like", and comment on other pins.

In Pinterest's own words, the site's "goal is to connect everyone in the world through the 'things' they find interesting. We think that a favorite book, toy, or recipe can reveal a common link between two people." [1]. This focus on "things" has made Pinterest of great interest to online retailers and marketers: for example, a recent market survey showed that a higher proportion of Pinterest users click through to e-commerce sites, and when they go there, they spent significantly more money than people who come from sites like Facebook or Twitter [24]. Another frequently discussed property of Pinterest is its demographics. In the United States, about 80% of its users are women; interestingly, however, fewer than 45% of Pinterest users in the United Kingdom are female [2].

While there is a large amount of writing about Pinterest in the popular, trade, and marketing press, there has been little scholarly work so far. As researchers, we want to systematically investigate impressions of Pinterest that come from anecdotal observations and market surveys. We defined three questions to guide our research:

R1-Activity: What drives user activity on Pinterest: what about a pin—and its pinner—"grabs the attention" of other users? For example, what role does gender play? Do pins from women receive more, less, or equal attention than those from men?

R2-Connection: What is the structure of social connection on Pinterest? For example, are women or men more connected?

R3-Comparison: How does behavior on Pinterest compare to behavior on other social network sites? Letting Twitter be our point of reference, do Pinterest and Twitter users systematically differ in what they talk about?

[1] http://pinterest.com/about

We used a quantitative approach to study these questions, performing a statistical analysis of data obtained from a web crawl of Pinterest. In addition to our specific findings, our results serve as an early snapshot of Pinterest that subsequent work will be able to use as a baseline.

In the remainder of the paper, we survey related work, describe the data we analyzed and how we obtained it, explain our statistical methods, present our results, and discuss their implications. Finally, we close with a brief summary.

RELATED WORK

There is a vast research literature on online communities and social networks. Researchers have applied a wide range of methods to study diverse types of research questions. For example, Ellison and Lampe have done a number of survey studies of Facebook users, showing that people use Facebook to maintain existing offline connections [20], identifying motivations for using Facebook, and showing that using Facebook builds social capital [14]. Other research has analyzed traces of interaction in social sites, identifying interaction patterns that: relate to social capital [7]; predict strength of social ties [16]; identify categories of status updates [22]; and characterize the practice of passing along messages from others ("retweeting") [4].

The prior research most relevant to our research questions and methods consists of quantitative studies of *what makes content interesting to members of a social network* and the *role of gender in online social interaction*.

What Makes Content Interesting

Much research attention has gone into investigating what makes content in an online community interesting to its members, that is, what they pay attention to. Two primary contexts in which these investigations have been done are discussion forums and Twitter.

For discussion forums, this problem takes the form of understanding and predicting what types of posts are most likely to receive replies. In a series of studies conducted on Usenet newsgroups, researchers from CMU investigated properties of the post, poster, and the newsgroup itself that influenced likelihood of reply. Early findings included: writing explicit requests, including personal "testimonials" relating one's connection to the group, and staying on-topic increased the probability of receiving a reply; and newcomers to a group were less likely to receive a reply than more established members [1]. There were several follow-ups to this work. Burke et al. [5] studied the role of self-disclosing introductions, signaled by first-person pronouns, mention's of the poster's age, and an acknowledgement that this is the poster's first post. These factors were found to increase reply probability. Burke and Kraut [6] also studied the effect of the politeness of a post, with the interesting finding that politeness leads to more replies in certain types of groups, while in others rudeness actually leads to more replies.

Other researchers have studied factors that predict not just getting a reply, but the quality of the reply. For example, Harper [17] studied Q&A sites and found that the type of a question (e.g., for personal advice or technical assistance) and the question's topic influenced reply quality.

Turning to Twitter, researchers have used retweeting—reposting someone else's tweet—as a measure of community interest, and have investigated what features of tweets and users predict retweeting. Suh [27] found that the presence of URLs and hashtags in tweets predicted more retweeting, as did a richer connection with the community, both in followers (a larger audience) and followees (perhaps indicating access to more diverse information). A more general perspective is that of influence, or flow of information through a network. For example, Cha [10] computed three different measures of influence in Twitter—in-degree in the Twitter social network, retweets of content, and mentions in tweets—and then examined the extent to which the different measures were correlated. Lerman and Ghosh [21] studied the flow of information in both the social news site Digg and in Twitter (where they took retweets as the mechanism of influence).

Our work explores some of the same issues as prior work, but on a new social network site. In particular, we investigated which properties of a pin and its pinner lead the Pinterest community to pay it more attention.

Gender

How gender is expressed in and influences online social interaction is a topic of great research interest. Herring wrote an overview of the literature prior to the dawn of online social networks [18]. A common theme was to study the extent to which gender played a role in online interaction similar to what is seen offline, and to examine whether factors like moderation and anonymity changed its role.

This research was done in an era when online interaction was—or at least was perceived to be—dominated by men and male norms of communication. This began to change as more online spaces were designed by and for women. And the rise of social network sites has changed this picture dramatically: for example, recent data [28] shows that women form a majority of Facebook and Twitter users, as well as dominating Pinterest; however, men are the majority of users on Google+ and LinkedIn.

For our purposes, studies that use quantitative methods to analyze behavioral data from social networks are most relevant. Caverlee and Webb analyzed data from MySpace, finding that women were twice as likely as men to make their profiles private, and that the most popular terms used by men and women in their profiles were starkly different [9]. Thelwall also analyzed MySpace data, with similar findings to Caverlee & Webb; additionally, he found that women had more friends than men, and that both men and women had more female than male friends [29]. Cunha found that male and female Twitter users differed in what hashtags they used for common topics [13].

Spurred by reports in the popular media, several recent studies used quantitative methods to investigate gender disparities in Wikipedia. In particular, the findings of Lam [19] include: women editors focused on different topics than men, topics of most interest to women had lower quality coverage, and that new female editors are significantly more likely to have their edits reverted than are new male editors.

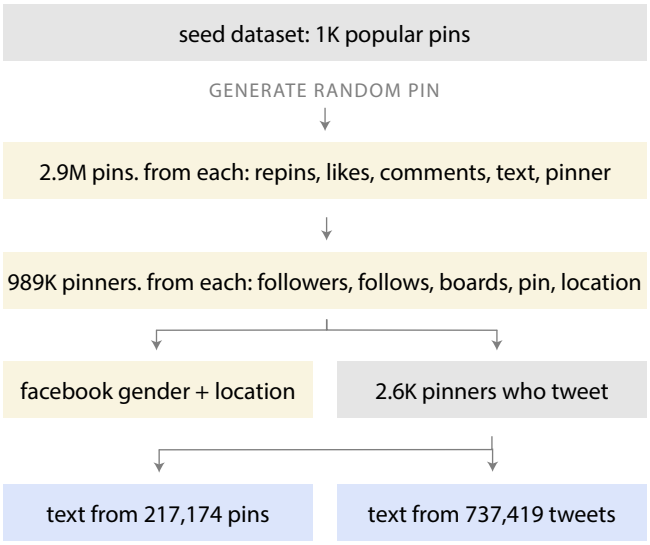


Figure 1. A flowchart of the steps taken in this paper to prepare data for modeling. We cover these steps in detail in the body of the paper, but include this figure for overview and reference. Yellow signifies data we use to answer research questions R1 and R2, while blue signifies data that answer R3.

In our research on Pinterest, we examine some similar issues to prior work, notably whether the gender of a pinner predicts the amount of attention the community gives to the user's pins, as well as the existence of gender-based differences in number of social connections.

DATA AND METHODS

We took a quantitative approach to investigate our research questions. We describe the data we collected and the statistical methods we used[2]. We summarize the limits of our approach at the end of the paper. Figure 1 presents an overview of the data presented in this paper.

Data

The data we needed from Pinterest comprised *pins* and *pinners*. For pins, we were interested in data that signaled interest from the Pinterest community:

repins: how many times this pin was repinned;

likes: how many other users liked this pin;

comments: textual comments included by the pin's pinner;

text: the pin's description and comments by others.

For pinners, we were interested in data indicating their activity on Pinterest and relevant personal characteristics:

followers: how many pinners follow this pinner;

follows: how many users this pinner follows;

boards: how many boards this pinner has created;

pins: how many pins this pinner has created;

Pinterest suggests to pinners that they include their Facebook and Twitter handles in their profiles. This let us obtain three additional data elements (where available):

location: a free-text description of a pinner's location, via either Pinterest or Facebook; if they specified it in both places, we used the Pinterest location;

gender: a pinner's self-reported gender via their Facebook profile page;

tweet text: via their linked Twitter handle, the text of all tweets written by this pinner in the last year.

Obtaining the data

While our goal was to obtain a random sample of pins and pinners, there is no publicly available means to do so. Instead, we developed a web crawler to approximate a random sample. Pinterest does not have a page that lists all pins or pinners; however, it does provide a list of "popular" pins[3]. In June 2012, we ran our web crawler to collect popular pins using 884 machines allocated via PlanetLab[4]. This gave us a "seed set" of approximately 1,000 pins.

We next wanted to expand our sample beyond "popular pins." We experimented with a variety of strategies, such as snowball sampling (which brings with it certain complications [3]). We eventually discovered, however, that we could reverse-engineer a pin's URL identifier structure to generate an apparently random view of the universe of Pinterest pins. We observed that adding small integers to the identifier in a pin's URL would find new pins created slightly later. Via trial and error, we discovered that the trailing five digits of the identifier encoded a pin's creation time. Therefore, we generated new pins by adding random integers in the range [1, 10,000] to our seed set of popular pins. We found no correlations among these pins in terms of boards (i.e., categories), pinners, popularity, location, etc. Obviously, we cannot make any absolute claims about the randomness of this sample. However, by inspection it looked like a "public timeline" view of Pinterest, and seemed preferable to a snowball sample crawl. Using this method, we collected a total of 2.9M pins and their associated data as noted above. We also collected the pinners for these pins, ending up with a set of 989,355 pinners.

For every pinner who specified their Facebook ID, we checked whether they had a public Facebook profile. If so, we obtained their self-reported gender and location, if provided, as noted earlier. In addition, from a pinner who specified their public Twitter handle, we selected a random set of 2,616 users. We then obtained all tweets from these pinners from the last year, as well as all as the text from these pinners' pins. In total, we collected datasets of 217,174 pins and 737,491 tweets.

Preparing the data for analysis

Once we obtained the basic data described above, we had to do some additional work to prepare it for analysis.

Location. The location data we obtained from Pinterest and Facebook typically were not clean, normalized geo-location data. Therefore, we used the Google Maps API[5] to conform user-entered location data to the level of a country. For the purpose of our models, we operationalize this as 45 binary

[2]Our dataset: `https://github.com/compsocial`

[3]`http://pinterest.com/popular`

[4]`https://www.planet-lab.org`

[5]`https://developers.google.com/maps`

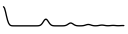
Variable	med	mean	max	distribution
pin's repins	0	0.96	5K	
pin's likes	0	0.213	980	
pin's comments	0	0.012	46	
pinner's followers	86	288.5	195K	
pinner's follows	86	125.1	29K	
pinner's boards	16	21.2	390	
pinner's pins	1K	1.7K	95K	
(categorical)				
pinner's gender	52.9K female		13.3K male	
country	49.8K USA		27.7K Great Britain	
	3.1K Brazil		2.9K France	
	2.5K Spain		1.1K Netherlands	
	6.9K elsewhere			

Table 1. Basic statistical descriptions of the variables collected in this paper. All quantitative variables have zero as their minimum. All distribution sparklines begin at zero and end at their variable's maximum value, on log-scale. The distributions themselves were induced with Gaussian kernel density estimates, a smoothing procedure.

Repins predictor	β	std. err	z	p
intercept	0.423	0.007	60.3	$< 10^{-15}$
pin's likes	0.527	0.003	190	$< 10^{-15}$
pin's comments	0.22	0.009	25.1	$< 10^{-15}$
pinner is female	0.0801	0.005	15	$< 10^{-15}$
pinner's followers	0.000733	$< 10^{-5}$	178	$< 10^{-15}$
pinner's follows	0.0000398	$< 10^{-5}$	9.45	$< 10^{-15}$
pinner's boards	-0.000226	$< 10^{-4}$	-9.1	$< 10^{-15}$
pinner's pins	-0.0000124	$< 10^{-6}$	-70.7	$< 10^{-15}$
from united states	0.104	0.005	20.6	$< 10^{-15}$
from great britain	0.0773	0.006	12.3	$< 10^{-15}$
from canada	0.102	0.0258	3.95	0.00008
Summary	**null dev.**	**res. dev.**	χ^2	**p**
	864K	476K	388K	$< 10^{-15}$

Table 2. The results of a negative binomial regression with number of repins as the dependent variable. We include only a selection of the geo-location variables in this model, as only these three countries corresponded to more than 1,000 data points. In the model summary, *Dev.* refers to deviance, a measure of the goodness of fit similar to the better-known R^2. The β coefficients here are not standardized; they remain on their original scales.

country variables, such as *from united states* and *from china*. We experimented with a more refined latitude and longitude operationalization, but this led (as might be expected) to too many confounds in the models.

Text. As noted above, we collected text for over 200K pins—their descriptions and any comments from other users—and over 700K tweets. This text serves as the basis for our investigation of whether users behave differently on Pinterest than on other social networking sites: specifically, do they talk about different things than they do on Twitter? We think this gives a richer characterization of behavior than simply modeling count variables such as *likes* and *repins*. Observe that this is a within-subjects sample, so we can more confidently attribute differences in language use to differences between Pinterest and Twitter, rather than to differences between the user populations of the sites.

We transformed all the pins and tweet text into a sparse matrix representation. The columns of the matrix represent all the words used in the entire corpus, and a "1" in a column indicates that that column's word appeared in the corresponding pin or tweet. We explain in detail below how our statistical methods use this matrix.

Statistical Methods

We used two statistical techniques to examine activity on Pinterest. For our first research question, we used negative binomial regression to model Pinterest activity as a function of the predictive variables listed above. We use negative binomial regression because our dependent variable is a count. For our second research question, we also used negative binomial regression. For our third question, we used penalized logistic regression (PLR) [15] to study the differences between Twitter and Pinterest text. PLR allows us to determine the relative membership of each phrase in Twitter and Pinterest.

Implemented in the R package `glmnet`[6], PLR predicts a binary response variable while guarding against collinearity among predictors, something not done by traditional logistic regression. This is important in this context since language often exhibits high degrees of collinearity: i.e., the word *transportation* will follow the word *public* much more often than *mouse*. The implementation of PLR we employ in this paper handles correlated predictors using the lasso technique, shifting most of a group of correlated predictors into the mass of the most predictive feature. Consequently, the vast majority get left out of the model, with PLR assigning them zero coefficients. The implementation also handles sparsity, which is important for natural language processing since a message contains only a small percentage of the globally observed phrases. While neither negative binomial regression or PLR likely offers the most powerful purely predictive model (i.e, SVMs would likely provide better predictive accuracy), these statistical approaches permit close inspection of the importance of the predictive variables. This is important since we care most about illustrating the forces driving Pinterest use.

RESULTS

Descriptive Statistics. Table 1 reports basic statistical descriptions of the variables (predictors) collected in this paper. Table 1 does not report minimums, since they are zero for all the quantitative variables. As is typical for social datasets, the means and medians of these variables often differ considerably. The last column of Table 1 shows distributions that illustrate the shape of the variable. We induced these distributions by computing Gaussian kernel density estimates over the log-transformed variable. Gaussian kernel density estimates smooth datasets in an attempt to expose a more

[6] `http://cran.r-project.org/web/packages/glmnet`

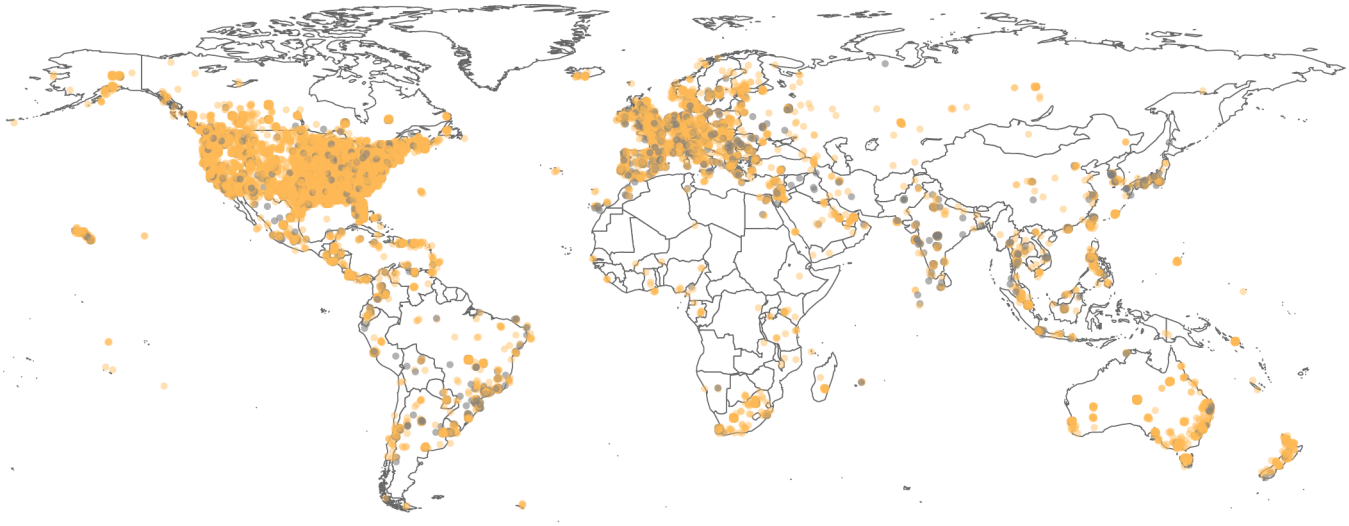


Figure 2. A heatmap of the distribution of Pinterest users by gender across the world. Females are represented as orange and males as gray on the map. The U.S. and Europe (particularly the U.K.) dominate Pinterest membership.

continuous, realistic picture of the underlying distribution [8]. We used log-transformation because otherwise a handful of data points would dominate the charts. We also summarize the distributions of the categorical variables, ordering country membership in descending order. For example, roughly 53% of the pinners in our sample come from the United States, while 29.4% live in Great Britain.

Figure 2 plots the location of the pinners in our sample on a map of the world, using color coding to distinguish men and women. This illustrates both where Pinterest users come from, and how gender of Pinterest users differs by region: for example, we see a higher concentration of men (represented by gray dots) in Europe than in the United States.

We next present the results of the analyses we did to investigate our research questions.

Followers predictor	β	**std. err**	**z**	**p**
intercept	5.89	0.0103	571.72	$< 10^{-15}$
pinner is female	-1.77	0.007	-234	$< 10^{-15}$
pinner's follows	0.00294	$< 10^{-5}$	422	$< 10^{-15}$
pinner's boards	-0.00391	$< 10^{-4}$	-47.8	$< 10^{-15}$
pinner's pins	0.000339	$< 10^{-6}$	456	$< 10^{-15}$
from united states	0.0917	0.00819	11.19	$< 10^{-15}$
from great britain	0.19	0.0101	18.84	$< 10^{-15}$
from canada	-0.891	0.0665	-13.39	$< 10^{-15}$
Summary	**null dev.**	**res. dev.**	χ^2	**p**
	1.3M	789K	523K	$< 10^{-15}$

Table 3. The results of a negative binomial regression with number of followers as the dependent variable. Again, we include only a selection of the geo-location variables in this model, as only these three countries corresponded to more than 1,000 data points. The β coefficients here are not standardized; they remain on their original scales.

R1-Activity: What drives activity on Pinterest?

We used standard techniques to measure the effectiveness of our models that investigated what drove Pinterest activity; specifically, we compared them to intercept-only (null) models and examined the reduction in deviance. Taking *pin's repins* as the dependent variable and using all others as predictor variables in a negative binomial regression provides considerable explanatory power, with an improvement in deviance of 864K - 476K = 388K. Deviance is related to a model's log-likelihood, and is an analog of the R^2 statistic for linear models A difference in deviances approximately follows an χ^2 distribution, so we can test for the overall likelihood of this model explaining our data, $\chi^2(10, N{=}588\text{K}^7)$ = 864K - 476K = 388K, $p < 10^{-15}$. Table 2 summarizes the predictors and overall fit of the *pin's repins* model.

Table 2 only presents three country-level variables because most countries are only sparsely represented and this caused problems for fitting a negative binomial model. All the variables possess non-random coefficients, with *pin's likes* perhaps unsurprisingly driving *pin's repins* the most, $\beta = 0.527$, $p < 10^{-15}$ (even on its original scale). Note that in this paper's models, we have opted to not standardize β weights, so they remain on their original scales. This is one reason they vary by orders of magnitude, for instance.

R2-Connection: What structures connections?

Table 3 presents the results of a negative binomial regression, with follower count as the dependent variable. The overall model explains a considerable amount of overall deviance, $\chi^2(7, N{=}672\text{K})$ = 1.3M - 789K = 523K, $p < 10^{-15}$. Perhaps surprisingly, given the large proportion of women on Pinterest and the gendered rhetoric often used to describe the site, being female suggests fewer followers β = -1.77, $p < 10^{-15}$.

[7] N=588K because only 588K of the 2.9M pins we obtained from Pinterest had gender and geo-location information.

Pinterest words	β	Pinterest words	β
DIY	3.003	[unicode heart]	2.776
cup	2.427	cute	1.746
recipe	1.745	dress	1.718
idea	1.71	color	1.52
hair	1.344	Love	1.072
use	0.969	Great	0.919
pretty	0.721	old	0.671
baby	0.641	made	0.624
love	0.537	book	0.506
great	0.429	photo	0.363
look	0.31	awesome	0.263
before	0.2	want	0.173
way	0.168	fun	0.156
one	0.118	over	0.112
need	0.0916	kid	0.085
little	0.0685	home	0.0483
girl	0	none	0

Table 4. The 34 phrases predicting a post belongs to Pinterest, not Twitter. The table flows left to right, then top to bottom. All phrases are significant at the 0.001 level.

Figure 3, on the other hand, shows follower count distributions by gender. We provide this figure because something interesting is happening: while the median male follower count is lower than the female median follower count (67 vs. 86, respectively), the mean male follower count is substantially higher than the mean female follower count (1,063 vs. 270, respectively). You can see this fact visually in Figure 3 by the fatter tail in the gray distribution. We revisit what might be driving this difference later in the *Discussion*.

R3-Comparison: How do Pinterest and Twitter compare?

Tables 4 and 5 present the results of our PLR model, showing words most effective for distinguishing Pinterest from Twitter discourse. As mentioned earlier, our implementation of PLR will assign most variables zero coefficients in order to discover a parsimonious model. While we have separated our results into two tables, all the coefficients derive from the same model, with dependent variable *post from pinterest*. The overall model explains a significant, if modest, amount of the variance distinguishing Pinterest from Twitter, $\chi^2(98, N{=}955\text{K}) = 1.03\text{M} - 816\text{K} = 219\text{K}, p < 10^{-15}$. All of the words in Tables 4 and 5 are significant at the 0.001 level, the default cutoff for inclusion in a `glmnet` model. This model, based solely on words contained in tweets and pins, accounts for approximately 21% of the deviance in the dependent variable. A model that incorporates other features important to Pinterest and Twitter (e.g., image features) presumably would explain more.

Figure 4 provides a deeper view into the structure around the predictive words in Tables 4 and 5. It shows Word Tree visualizations [32] (created using the Many Eyes site [30]) of searches for "want to" and "need to" in only the text from Pinterest. For example, Figure 4 shows that the phrases "want to do" and "want to make" appear many times across multiple pins. We think it offers an intriguing, deeper look into the structure of the text behind our statistical techniques.

Twitter words	β	Twitter words	β
"	-4.946	que	-3.7
Thanks	-2.859	thanks	-2.849
watching	-2.736	tonight	-2.403
today	-2.213	Thank	-2.134
la	-1.694	Happy	-1.46
going	-1.273	lol	-1.193
friend	-1.006	see	-0.989
feel	-0.969	new	-0.96
:-D	-0.959	last	-0.948
still	-0.948	getting	-0.929
here	-0.92	haha	-0.906
think	-0.903	people	-0.9
ca	-0.885	sure	-0.88
go	-0.747	now	-0.717
wait	-0.703	know	-0.689
night	-0.683	week	-0.681
come	-0.653	Photo	-0.614
really	-0.603	right	-0.602
thing	-0.597	well	-0.56
time	-0.545	much	-0.54
day	-0.512	work	-0.508
next	-0.5	New	-0.482
always	-0.473	oh	-0.441
Oh	-0.408	good	-0.404
something	-0.371	find	-0.317
back	-0.311	first	-0.31
better	-0.31	even	-0.31
blog	-0.241	Good	-0.189
more	-0.169	never	-0.155
take	-0.105	Day	-0.089
year	-0.053	very	-0.017
down	-0.012	life	-0.011

Table 5. The 64 phrases that predict a post belongs to Twitter and not Pinterest. The table flows left to right, then top to bottom. All phrases are significant at the 0.001 level.

DISCUSSION

We have presented a number of statistical models investigating behavior on Pinterest. We now reflect on these findings as framed by our research questions.

R1-Activity: What drives activity on Pinterest?

Table 1 shows a few surprising and telling features of Pinterest as compared to other online social networks. First, while repins, likes and comments are all relatively rare, repins happen considerably more often than either likes or comments. This shows that reposting content from others—here, repinning—is a first-class activity, much like retweeting on Twitter and re-blogging on Tumblr. Furthermore, we see that on average pinners have 86 other pinners following them (median). Perhaps most surprisingly, the average pinner in our sample has created over 1,000 pins. This probably is due to our sampling strategy (i.e., sampling pinners based

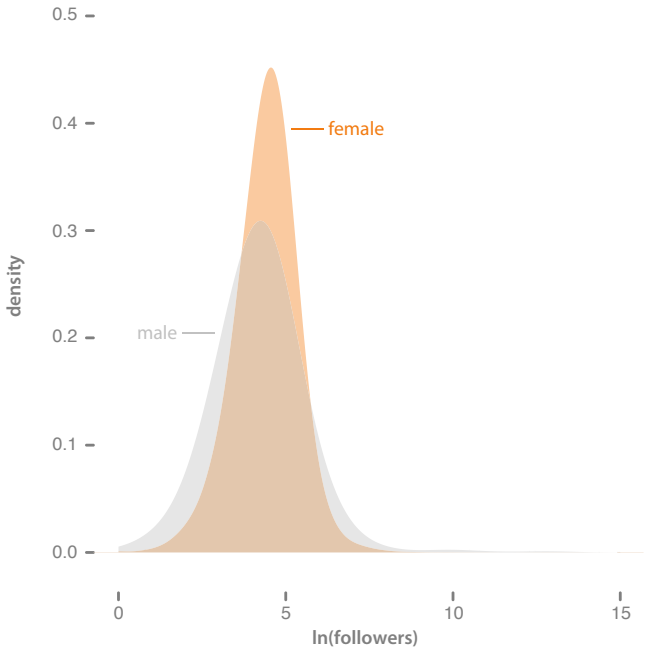


Figure 3. A comparison of follower count by gender, with orange representing females and gray representing males. While males have a lower median, there is also more mass in the tail of the male follower distribution. We induced the distributions with Gaussian kernel density estimates, a smoothing procedure.

on recently created pins); however, if our sample is biased in this way, it does not invalidate our findings. At most, it limits them to apply to the active sub-population of Pinterest responsible for much of the content and behavior on the site. Finally, our data support the popular and trade press view suggesting that Pinterest has a female supermajority: 80% of the pinners in our data are women.

Likes, comments, followers all drive repins. When we look at the factors that drive pinners to repin each other's content, we see that likes, comments, gender, followers, following count and total pins all contribute. Our model (shown in Table 2) captures 44.9% of the variance around repinning; while this is a fairly good explanatory model, including other features in the model would presumably increase explanatory power. We see some predictors going in expected directions. For example, the more followers a pinner has, the more repins they can expect, $\beta = 0.000733$, $p < 10^{-15}$. The more pins a pinner creates, the less likely any one of them will be repinned, $\beta = -0.0000124$, $p < 10^{-15}$. And as is to be expected, more likes and more comments are strongly correlated with the likelihood of repinning, $\beta = 0.527$, $p < 10^{-15}$ and $\beta = 0.22$, $p < 10^{-15}$, respectively.

Being female means more repins. However, the result that stands out most in Table 2 is that pins from females receive dramatically more repins, $\beta = 0.0801$, $p < 10^{-15}$. While smaller in effect size than likes and comment count, gender ranks as the third most powerful predictor and is two orders of magnitude bigger than the next largest, a pinner's follower count. Do men pin content that is less interesting to the broader Pinterest community? Or could assortativity [23] be at work, a pattern often found in social networks whereby people tend to connect with those similar to them (in this case, of the same gender)? If this is in fact the case, then men would immediately find themselves at a disadvantage for repins since they would have smaller audiences. We feel this deserves more work and a deeper analysis in the future.

Being American or British earns repins. While we were unable to analyze every country in our dataset because of sparsity, we do find that pinners from the United States and the United Kingdom attract more repins than the rest of the world, $\beta = 0.104$, $p < 10^{-15}$ and $\beta = 0.0773$, $p < 10^{-15}$, respectively. We see this as a rich area for future research, investigating the linguistic and cultural reasons behind the attraction of content by American and British pinners.

R2-Connection: What structures connections?

In this paper, we examine the accumulation of followers as a function of other activity metrics. Again, we see similar patterns as we saw with repins: pins, following count and gender all impact follower count, and in roughly the same order (in terms of magnitude).

Being female means fewer followers. Most surprisingly, while female Pinterest users get more repins, they get fewer followers: $\beta = -1.77$, $p < 10^{-15}$. This finding seems perplexing. Why would the signs of the coefficient reverse in the two regressions? Figure 3 provides some insight on this point. First, as we would have suspected given our conjecture about assortativity above, men have a lower median follower count than women. However, the mean is substantially higher, reflected in Figure 3 in the tail down the right side of the axis. We see two possible explanations for this finding. First, simple statistical artifacts could be at work: because four times as many women as men use Pinterest, the variance of the male distribution is bumpier, wider and noisier. If this is true, the simple passage of time (and more men joining the site) will make the distributions even out. However, possibly the difference we see is due to other factors and will persist. For example, maybe the genders are at different places along the technology adoption curve, with proportionally more early adopters among men (who disproportionately attract attention)? Or, in a parallel argument, perhaps the men on the site disproportionately attract followers simply because they are scarce. Further research is needed to produce a more conclusive answer.

R3-Comparison: How do Pinterest and Twitter compare?

We selected Twitter as a foil for Pinterest for multiple reasons. First, we have seen quite a bit of scholarly work on Twitter in recent years, and this can act as a backdrop for the findings we present here. Second, the two sites share many of the same structural affordances, notably they both adopt a following & follower social network model to structure interactions, users can repost (retweet or repin) other users' contributions, and references to external content are either frequent (URLs on Twitter) or mandatory (on Pinterest, each pin refers to an image and its containing web page).

Pinterest's verbs: use, look, want, need. We performed a textual analysis comparing what users say on Pinterest to what they say on Twitter. The differences are illuminating,

Figure 4. Word Tree visualizations of searches in the Pinterest text dataset. At top, a visualization depicting a search for "want to". The visualization shows phrases that branch off from "want to" across all the text in the Pinterest dataset. A larger font-size means that the word occurs more often. "want to" appeared 2,614 times. At bottom, a Word Tree visualization of a search for "need to", a phrase that appeared 2,878 times. Both "want" and "need" are among the top textual features for distinguishing Pinterest from Twitter. The visualizations come from the site Many Eyes.

providing a concise set of terms that serve to identify Pinterest. Some describe topics that enjoy considerable attention on the site, like *DIY* (β = 3.003), *recipe* (β = 1.745), *book* (β = 0.506), *photo* (β = 0.363) and *cup* (β = 2.427), another manifestation of *recipe*. Still others depict qualities that set the site apart: *old* (β = 0.671), *fun* (β =0.156) and *pretty* (β = 0.721). Of the 34 words listed in Table 4, only four are verbs in their primary sense, however. They are *use* (β = 0.969), *look* (β = 0.31), *want* (β = 0.173) and *need* (β = 0.0916). Many popular press articles have focused on Pinterest's commercial potential, and here we see verbs illustrating that consumption truly lies at the heart of the site.

Twitter: lol, watching now. Turning to the language of Twitter, we something altogether different. First and foremost, words jump out proclaiming the importance of *now* (β = 0.0916), such as *today* (β = -2.213), *tonight* (β = -2.403) and *time* (β = -0.545). This perhaps simply reflects Twitter's "What's happening?" prompt above the box for entering tweets. We also see words that proxy for underlying demographic differences between the sites, such as the Spanish words *que* (β = -3.7) and *la* (β = -1.694). Given previous research on Twitter's importance surrounding currently unfolding events [25], we unsurprisingly see words like *watching* (β = -2.736), *see* (β = -0.989) and *going* (β = -1.273) acting as strong predictors for Twitter over Pinterest. We also see some very social words in Twitter's lexicon, like *thanks* (β = -2.849), *lol* (β = -1.193), :-D (β = -0.959) and *haha* (β = -0.906), all mimicking processing we would normally associate with face-to-face social life.

Limitations and Future Work

Our work has several limitations, which in turn suggest avenues for future research.

Obtaining data. The way we obtained a sample of Pinterest data was fairly labor-intensive and doesn't offer a guarantee of randomness. For example, the fact that the average pinner in our sample had 1K pins suggests that we were sampling from the high end of the activity distribution. While we believe our results still stand, we obviously would prefer to obtain a random sample. Clearly the best way for researchers to be able to obtain appropriate data samples would be for Pinterest to publish an API.

Data analyzed. We did not sample the main type of content that drives Pinterest activity: pictures. The count and text data we did sample is more analytically tractable: there are scalable, efficient and validated methods for such data, but no analogous methods for analyzing image data. For example, recognizing objects in images (which certainly would be useful for understanding Pinterest behavior) without an extensive training set remains an unsolved computer vision problem. In fact, simply generating the training set has served as the motivating problem for influential work in human computation [31]. We see analysis of pictures—whether through automated or human-computation techniques—as an important area for future research. Another way to get at some of the same issues would be to analyze the external web pages that contain images from pins; for example, it would be interesting simply to identify the proportion of pins that link to e-commerce sites and which e-commerce sites are most popular.

Methods used. Our approach is wholly quantitative and statistical. Such methods have built-in blind spots: most significantly, while they permit making claims about broad, large-scale practices on the site, they cannot uncover users' motivations and goals for using Pinterest. Qualitative research can be used to enrich the picture we painted in this paper, providing thick descriptions of motivations and goals for using Pinterest. Further, the statistical techniques we used examine only a small segment of the possible behaviors on Pinterest. As we mentioned above, for instance, our methods ignore photo-sharing practices themselves. With current technologies, to take one example, you could examine the distribution of color palettes as they vary with popularity indicators. We look forward to both quantitative and qualitative work that follows up on the findings we obtain in this early study of Pinterest.

Additional analysis of location. While we represented location in terms of countries, it may be productive to explore a more detailed representation in the future, for example, one that enables reasoning based on geographic proximity. For example, we would like to investigate locality of connection: say, to what extent do pinners from the UK follow other UK pinners, repin their pins, etc.?

Additional analysis of gender. It would be interesting to do further analysis of the effects of gender on Pinterest interaction; indeed, we plan to do so. This would strengthen the contrast to findings from male-dominated sites like Wikipedia. For example, we would like to know whether men and women connect more to other users of the same gender or another gender, whether women and men are more likely to repin pins from users of the same gender, whether men and women favor different vocabulary terms in their descriptions and comments, and whether men and women focus on different topics in their pins and comments.

CONCLUSION

In this paper, we provide a statistical overview of Pinterest, a new social network site surging into prominence. We arrive at three main findings. First, being female on the site leads to more repins, while geography seems to play no role. Second, women have fewer followers. Finally, comparing the language used on Pinterest to language used on Twitter, we come to a concise set of terms defining the two sites. Notably, the four verbs uniquely describing Pinterest are "use," "look," "want," and "need," reflecting the "things" at the heart of Pinterest. We hope these early findings act as springboard for future research on Pinterest.

ACKNOWLEDGEMENTS

We would like to thank our reviewers and our respective research groups for making valuable comments that improved this work. We also thank the NSF for supporting this research with grant IIS-1212338.

REFERENCES

1. J. Arguello, B. S. Butler, E. Joyce, R. Kraut, K. S. Ling, C. Rosé, and X. Wang. Talk to me: foundations for successful individual-group interactions in online communities. In *Proc. CHI*, pages 959–968, 2006.
2. E. Barnett. Barack Obama signs up to Pinterest. http://www.telegraph.co.uk/technology/social-media/9170718/Barack-Obama-signs-up-to-Pinterest.html.
3. P. Biernacki and D. Waldorf. Snowball Sampling: Problems and Techniques of Chain Referral Sampling. *Sociological Methods & Research*, 10(2):141–163, 1981.
4. d. boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proc. HICSS*, HICSS '10, pages 1–10, 2010.
5. M. Burke, E. Joyce, T. Kim, V. Anand, and R. Kraut. Introductions and requests: Rhetorical strategies that elicit response in online communities. In *Communities and Technologies 2007*, pages 21–39. 2007.
6. M. Burke and R. Kraut. Mind your ps and qs: the impact of politeness and rudeness in online communities. In *Proc. CSCW*, CSCW '08, pages 281–284, 2008.
7. M. Burke, C. Marlow, and T. Lento. Social network activity and social well-being. In *Proc. CHI*, pages 1909–1912, 2010.
8. R. Cao, A. Cuevas, and W. Gonzalez Manteiga. A Comparative Study of Several Smoothing Methods in Density Estimation. *Computational Statistics & Data Analysis*, 17(2):153–176, 1994.
9. J. Caverlee and S. Webb. A large-scale study of myspace: Observations and implications for online social networks. In *Proc. ICWSM*, 2008.
10. M. Cha, H. Haddadi, F. Benevenuto, and K. Gummad. Measuring user influence on twitter: The million follower fallacy. In *Proc. ICWSM*, 2010.
11. J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy. Make new friends, but keep the old: recommending people on social networking sites. In *Proc. CHI*, pages 201–210, 2009.
12. J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In *Proc. CHI*, pages 1185–1194, 2010.
13. E. Cunha, G. Magno, V. Almeida, M. A. Gonçalves, and F. Benevenuto. A gender based study of tagging behavior in twitter. In *Proc. HT*, pages 323–324, 2012.
14. N. B. Ellison, C. Steinfield, and C. Lampe. The benefits of facebook "friends:" social capital and college students¡¯ use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, 2007.
15. J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.
16. E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proc. CHI.*, pages 211–220, 2009.
17. F. M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan. Predictors of answer quality in online q&a sites. In *Proc. CHI*, pages 865–874, 2008.
18. S. C. Herring. *Gender and Power in On-line Communication*, pages 202–228. Blackwell Publishing Ltd, 2008.
19. S. T. K. Lam, A. Uduwage, Z. Dong, S. Sen, D. R. Musicant, L. Terveen, and J. Riedl. Wp:clubhouse?: an exploration of wikipedia's gender imbalance. In *Proc. WikiSym*, WikiSym '11, pages 1–10, New York, NY, USA, 2011. ACM.
20. C. A. Lampe, N. Ellison, and C. Steinfield. A familiar face(book): profile elements as signals in an online social network. In *Proc. CHI*, pages 435–444, 2007.
21. K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Proc. ICWSM*, 2010.
22. M. Naaman, J. Boase, and C.-H. Lai. Is it really about me?: message content in social awareness streams. In *Proc. CSCW*, CSCW '10, pages 189–192, 2010.
23. M. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.
24. R. Relevance. Richrelevance shopping insights study reveals how social networks stack up for retailers. http://www.richrelevance.com/blog/2012/09/social-infographic.
25. D. Shamma, L. Kennedy, and E. Churchill. Peaks and persistence: Modeling the shape of microblog conversations. In *Proc. CSCW*, pages 355–358, 2011.
26. P. Sloan. Pinterest: Crazy growth lands it as top 10 social site. http://news.cnet.com/8301-1023_3-57347187-93/pinterest-crazy-growth-lands-it-as-top-10-social-site.
27. B. Suh, L. Hong, P. Pirolli, and E. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *IEEE SocialCom*, pages 177 –184, aug. 2010.
28. C. Taylor. Women Win Facebook, Twitter, Zynga; Men Get LinkedIn, Reddit. http://on.mash.to/NwOEcR.
29. M. Thelwall. Social networks, gender, and friending: An analysis of myspace member profiles. *JASIST*, 59(8):1321–1330, 2008.
30. F. Viegas, M. Wattenberg, F. Van Ham, J. Kriss, and M. McKeon. Many Eyes: a site for visualization at internet scale. In *InfoVis*, pages 1121–1128. Published by the IEEE Computer Society, 2007.
31. L. von Ahn, R. Liu, and M. Blum. Peekaboom: A Game for Locating Objects in Images. In *Proc. CHI*, pages 55–64, 2006.
32. M. Wattenberg and F. Viégas. The Word Tree, an interactive visual concordance. In *InfoVis*, pages 1221–1228, 2008.