Specialization, Homophily, and Gender in a Social Curation Site: Findings from Pinterest

Shuo Chang[1], **Vikas Kumar**[1], **Eric Gilbert**[2], **Loren Terveen**[1]

[1]GroupLens Research
Dept. of Computer Science and Engineering
University of Minnesota
{schang,vikas,terveen}@cs.umn.edu

[2]School of Interactive Computing
Georgia Institute of Technology
gilbert@cc.gatech.edu

ABSTRACT
Pinterest is a popular *social curation* site where people collect, organize, and share pictures of items. We studied a fundamental issue for such sites: what patterns of activity attract attention (audience and content reposting)? We organized our studies around two key factors: the extent to which users *specialize* in particular topics, and *homophily* among users. We also considered the existence of differences between female and male users. We found: (a) women and men differed in the types of content they collected and the degree to which they specialized; male Pinterest users were not particularly interested in stereotypically male topics; (b) sharing diverse types of content increases your following, but only up to a certain point; (c) homophily drives repinning: people repin content from other users who share their interests; homophily also affects following, but to a lesser extent. Our findings suggest strategies both for users (e.g., strategies to attract an audience) and maintainers (e.g., content recommendation methods) of social curation sites.

Author Keywords
Social Network; Data Analysis; Topic Detection; User Profiling

ACM Classification Keywords
H.1.2. User/Machine Systems

General Terms
Human Factors; Algorithms

INTRODUCTION
Social network sites have become central to people's lives. They let people connect and stay in touch with family and friends (e.g., Facebook), find and share information on topics in which they are interested (e.g., Twitter), ask and answer specialized questions (e.g., StackOverflow).

Social curation sites are an important type of social network site. They let users collect, organize, and share collections of items. The focus is not creating new content (unlike, say, YouTube), but rather something akin to, "Here are things I found on the web that I think are interesting." Well-known social curation sites include del.icio.us (for web bookmarks), Digg and Reddit (for news items and other types of web content), and newer sites like Storify, Scoop.it!, and Wanelo.

We have been studying Pinterest as a representative social curation site. Pinterest is organized around the metaphor of a pin board. Users (called "pinners") pin pictures of items they find on the web and organize them into boards representing interests like recipes, crafts, children's toys, etc. Like many other sites, users can follow each other: if you follow someone, their new pins show up on your home page's feed. Pinterest has drawn recent research interest [13, 39, 27] for reasons such as its fast growth, the central role found images play, its strong link to e-commerce, and the gendered nature of its use, with women making up around 80% of US users.

We are interested in a key issue for any social curation or social network site: **What attracts attention?** We study two main ways attention manifests: attracting followers and getting one's content reposted.

We consider two main factors in our study of this issue:

- *Specialization*: the extent to which users specialize in a particular type of content (e.g., a user may pin only recipes) vs. diversifying over a range of topics.
- *Homophily*: similarity of topical interests among people who follow one another.

As we detail below, these factors are well-known to drive behavior both online and offline. For example, large bodies of work have shown the role of homophily in the diffusion of information and behaviors through social networks, online and off (e.g., [14, 25]). We also examine how *gender* may mediate these two factors since a growing body of research has uncovered gender differences in behavior on social network sites. Taken together, these factors allow us to study attention on a social curation site through the lenses of individual characteristics (gender), individual behavior (specialization) and community structure (homophily).

Within this frame, we defined three specific research questions to guide our research:

RQ1-Topics. What is the overall topical structure of con-

CSCW'14, Feb 15–19, 2014, Baltimore, Maryland, USA.

tent on Pinterest? What topics are most popular? How are different topics related? Do men and women differ in the types of content they pin?

RQ2-Specialization. To what extent do users specialize in particular topics? Do women and men differ in their degree of specialization? What factors, including degree of specialization, attract more followers?

RQ3-Homophily. Are users more similar to users whom they follow than to random users? Are users more likely to repin from users who are similar to them than from random users?

We used quantitative methods to study these questions. We wrote a web crawler to obtain data from Pinterest and then performed statistical analyses on the data. Our results both illuminate the nature of activity on Pinterest and have implications for social network and social curation sites in general. For example, we found that users attract more followers when they share content on a range of topics – but there is a limit beyond which increased diversity is not helpful. Our results also suggest that Pinterest users may be more interested in content on topics they care about than in content from their social connections; this can inform the design of content recommendation algorithms for social curation sites.

In the remainder of the paper, we describe the theoretical background for our research and survey related work, describe the data we obtained and how we obtained it, describe how we produced a representation of topics to use in our analyses, present our results, and then close by discussing the implications of our results for research and design.

THEORETICAL FOUNDATIONS AND RELATED WORK

There is much research on social network sites and their underlying social phenomena. Most relevant to our concerns is work that: (i) examines the types of behaviors that attract attention and an audience, and (ii) studies of the role of gender in social network sites. We also examine several related studies conducted on Pinterest.

Attracting audience and attention

What makes users pay attention to content or fellow users of a social network site? Numerous studies have investigated this fundamental issue in different contexts.

What types of questions attract answers? In a series of studies of Usenet newsgroups, researchers from CMU investigated properties of the post, poster, and the group itself that influenced likelihood of reply. For example, Arguello et al. [5] found that writing explicit requests, including personal "testimonials" relating one's connection to the group, and staying on-topic increased the odds of receiving a reply.

What types of posts attract reposting? Researchers have studied Twitter, using *retweeting – reposting* someone else's tweet – as a measure of interest, and have investigated features of tweets and users that predict retweeting. Suh [33] found that the presence of URLs and hashtags in tweets predicted more retweeting, as did richer connections with other Twitter users, both in followers (a larger audience) and followees (indicating access to more diverse information).

What attracts an audience? Recent research has sought to identify structural and content factors that predict formation of ties in social network sites. Several researchers found correlations between the content of users' tweets – such as the expression of emotions like joy or sadness [18] or positive vs. negative sentiments [11] – and the number of followers Twitter users have. More generally, preferential attachment is a foundational phenomenon in the formation of social ties: people in a social network tend to connect to others in the network who already are popular. This is a common property of real-life social networks [6] and has been useful in predicting the formation of new social ties [21].

Homophily and Specialization

The principle of *homophily* states that the more similar people are, the more likely they are to form social connections [25]. Shared interests are one type of similarity. Prior work has explored the role of homophily in online social interaction: for example, Ren et al. tested theories about community attachment by forming groups on the MovieLens film recommendation site based on similarity of movie tastes [28].

Various researchers have studied topical specialization in online communities. For example, Demartini observed topic specialization in Wikipedia [12], and Priedhorsky et al. [2] and Masli et al. [24] observed a geographic analog (region specialization) in a geographic wiki for bicyclists. Ziegler et al. [40] did pioneering work on the utility of diversity in recommender systems. They developed an algorithm to diversify lists of recommendations—i.e., to include items that reduce the overall inter-item similarity in the list—and showed that this increased user satisfaction with the results.

More directly relevant for our purposes, Wang and Kraut [37] studied how the intial topical focus of a Twitter user's tweets affected the number of followers they attained. Interestingly, they found theoretical justification that both high focus (specialization) and low focus (diversity) could lead to more followers. The high focus argument is based on homophily, and the low focus argument is based on network externalities, i.e., more diverse content will appeal to a broader audience. Wang and Kraut found that more focused initial tweets led to more followers. Hutto et al. [16] also built a theory-based prediction model of following behavior that included a broad range of prediction variables, including topical focus of tweets. In contrast to Wang and Kraut, they did not find that greater topical focus predicted more followers. This indicates that more work is needed to understand the role of focused vs. diverse content in attracting an audience.

This work. We explore some of the same issues as prior work, but in a different context and with different types of data; we study what factors – including types of interests and degree of specialization – result in Pinterest users gaining more followers.

Gender

In "A theoretical agenda for feminist HCI" [29], Rode identified three paradigms for thinking about gender within HCI and introduced several outside perspectives that she argues enable a richer and more nuanced approach to gender. The paradigm relevant to our research is (in our words) *gender* *as a variable*. This presumes that gender differences *may* be relevant for analysis and design; thus, quantitative studies include participant gender as an independent variable. For example, Beckwith et al. [7] investigated whether there were differences in "tinkering" behavior and self-efficacy among men and women learning to program; if so, these might inform the design of learning support tools. In another example, Tan et al. [34] found through a controlled experiment that women benefit from a wider field of view when navigating on-screen, 3D displays.

Social media studies. Much research has been done on the role of gender in online social interaction; Herring gives an overview of the early literature [1]. More recent work has studied social network sites. Caverlee and Webb found that female MySpace users were more likely than male users to keep their profiles private, and that men and women used very different terms in their profiles [9]. Thelwall also analyzed MySpace data. His findings included gender differences in social connection: women had more friends than men, and both men and women had more female than male friends [35]. Cunha found that male and female Twitter users used different hashtags for common topics [10]. Lam et al. investigated gender disparities on Wikipedia [19]. They found that men comprised about 84% of Wikipedia editors and made over 90% of edits. They also identified various effects possibly related to this disparity, notably that topics of most interest to women received lower quality coverage.

Potential confounds. Rode identified three potential problems surrounding gender: (1) ignoring the social context in which gender issues occur; (2) treating gender as an essential, immutable characteristic, rather than something socially constructed and performed by individuals; (3) a tendency to design "female versions" of technologies, thus risking ghettoizing women and girls.

This work. We take the *gender as a variable* approach, considering gender in several analyses, notably, whether women and men focus on different topics, and whether women and men differ in the degree to which they are specialists or generalists. Here, Pinterest is the social context in which gender is performed, albeit through the interface affordances provided by Pinterest's designers. Rode's other two confounds are largely problems with the interpretation and application of study results, rather than with the studies per se. This is particularly true for an experiment or quantitative study: this sort of study identifies the *what*, but not the *why* of the results. Therefore, we clearly distinguish our emprical findings from any interpretation or drawing of implications, and keep Rode's concerns in mind when we do offer interpretations and implications. Note that since Pinterest is dominated by female users, it serves as a usefully contrasting context to previous work on sites that are male dominated (e.g., Wikipedia) or have relatively equal gender balance (e.g., MySpace, Twitter).

Aside: popular discourse about Pinterest is gendered. The popular impression of Pinterest is highly gendered: it is *seen as* a site for women [23], and this has engendered much dismissive sexist reaction. A vivid illustration is the emergence of a number of "male Pinterest" sites [32]. A lively debate about the gendered view of Pinterest is occurring online, e.g. [38, 26, 36]. Of course, there is nothing "inherently female" about Pinterest: anyone can pin any sort of content. However, the popular impression of Pinterest as a "women's site" has real effects. Recent research identified stereotypes about Wikipedia editors [4]. Since people are less likely to participate in a group when they don't identify with its members [30], these stereotypes may discourage many people from editing Wikipedia. As we discuss below, stereotypical views concerning Pinterest may have similar affects.

Studies of social curation on Pinterest

Several studies of Pinterest have been published recently. Gilbert et al. [13] did a quantitative study. Their central findings were: female users were repinned more often than males, but had a lower mean number of followers than males; and users employed very different language on Pinterest than on Twitter, with Pinterest language characterized by words of consumption and desire: "use", "look", "want", and "need". Ottoni et al. also did a quantitative study [27] . They found that women on Pinterest used social interaction mechanisms more than men, that men and women concentrated on different topics, and that women tended to be content generalists, while men specialized more. They also reported that women had more followers than men[1].

In contrast to the previous two studies, Zarro et al. [39] used qualitative methods to sketch the nature of social curation on Pinterest. Among their many interesting observations, one particularly relevant to our work is that their participants reported that Pinterest was about *what they enjoyed* (i.e., content), rather than social interaction.

We took a similar approach as Gilbert et al., but studied mostly different issues, although we also modeled the factors that predict the number of followers a user will have. Our model extended theirs by including factors for the type and diversity of content pinned by users. We studied some of the same issues as Ottoni et al., but (as detailed below) we used a richer representation of content and applied more rigorous methods. We studied different questions and applied different methods than Zarro et al. However, our analysis of Pinterest topics and the effect of homophily on repinning and following behavior provide quantitative evidence for the rich observations they drew from qualitative data.

DATA

We describe the data we gathered, how we gathered it, and the properties of our sample. To answer our research questions, we needed data about *pins*, *boards*, and *pinners*. For pinners, we were interested in data that indicated their content and social behavior, including:

- *boards* the user has created;
- *pins* the user has added to his or her boards;
- a list of the users this user is *following*;
- the total number of users this user is *following*;

[1]This result may appear to conflict with that of Gilbert et al. However, Ottoni et al. only present this result via a graph with no accompanying statistical analysis. Therefore, it is not clear whether there is a real inconsistency here.

	Q_1	Median	Mean	Q_3
Following	55.0	104.0	293.6	199
Followers	54	115	22,410	233

Table 1. Following and followers: basic statistics in our sample

- the total *number of followers* this user has.

Many Pinterest users also include a link to their Facebook profile; when this was available, we followed the link to obtain the user's self-reported *gender*.

For pins, we collected data that let us trace how users categorized pins and the flow of information via repinning:

- the *board* to which a specific user added this pin;
- (if this was a repin) the *original pinner*, i.e., the user who first pinned this item and thus started the "tree" of repinning events to which this particular pinning act belonged;
- (if this pin is subsequently repinned by other users) the *repinning user and board to which this pin is repinned*.

For boards, we were interested in data we could use to understand topical structure on Pinterest, specifically:

- the pre-defined Pinterest *category* the board's owner associated with this board (if any).

Gathering the data

Our goal was to obtain a random sample of Pinterest data. However, Pinterest does not provide any publically available way to do so. We therefore wrote a web crawler to approximate a random sample. We ran the crawler in two stages, first to collect a set of *pinners*, then to collect a set of *pins*.

Gathering Pinners

Since we wanted to do analyses based around content, i.e., pinning, we wanted a sample of active pinners. We experimented with several methods of obtaining a sample; we settled on a random walk based method [20], which was simple and performed well. We ran the crawler from Nov 26, 2012 to Jan 15, 2013, performing a random walk from Pinterest's "Everything" page, a public timeline of the most recent pins.

We randomly picked a set of about 10 users whose pins appeared on "Everything" to serve as a "seed set" for the crawler; this seeding method satisfied our bias toward sampling active pinners. The crawler then expanded the seed set by randomly choosing a user from the followers or following list of the user, and then repeating the process from the chosen user. We also did a random "jump" after every 1000 steps of walking the network; again randomly selecting a user from the "Everything" page. We collected the data described above for 46,365 users.

We were able to obtain gender information from Facebook for 32,229 of the users, of whom 30,029 were female and 2,200 were male. This means that 93.2% percent of the known-gender users were female, higher than the overall proportion of female Pinterest users, which is estimated to be about 80%[27, 13]. (Again, this likely is because we focused our sampling on active pinners.)

The Pinterest UI poses a challenge to collecting the following or followed lists of a user. These lists are presented in "infinite scrolling" pages with a limit of about 500 users [2]. Thus, in some cases we were not able to obtain all the following users for a user. For reference, Table 1 shows basic data concerning the size of the following and followers lists of users in our sample [3].

Users	46,365
Female	30,029 (64.8%)
Male	2,200 (4.7%)
Unknown gender	14,136 (30.5%)
$>$ 90% following obtained	42,369
Pins	3,142,941

Table 2. Summary of the data in our sample

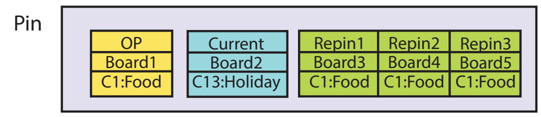

Figure 1. Structure created by different users pinning and repinning the same pin to different boards. OP stands for "Original Pinner".

In practice, the infinite scrolling / 500-user-limit issue was not a serious obstacle: we were able to obtain at least 90% of the following users for 42,369 of the 46.4K users. We consider this proportion of the data to comprise an adequately complete sample, and so we used this set of 42.4K users for analyses based on social network relationships.

Gathering Pins and Boards and Repin Relationships

We next gathered pins (3.1M in total) for the users from the first step. Again, there was a wrinkle in the access Pinterest provides: only the most recent 10% of all pins from a user can be obtained. For each pin, we also retrieved the original pinner of the pin (if it wasn't the current user), and up to 10 random other users who subsequently repinned this pin[4]; we also obtain the boards which the original pinner and the other users had associated this pin. Finally, for each board, we retrieved its Pinterest category.

Table 2 summarizes the data we collected.

REPRESENTING TOPICS

To perform our analyses, we need a set of topics that represent the content that Pinterest users pin. Pinterest provides a set of 33 categories[5], ranging from "Animals" to "Videos". This set is the first and most obvious candidate for us to consider. However, Pinterest data is richly organized, which offered several other opportunities to consider:

- Co-occurrence relationships between different types of data, e.g. between pins and boards;

[2] Javascript embedded in these pages senses when a user has scrolled to the bottom, and then automatically loads more content.

[3] The mean number of followers is so high because a small proprtion of sampled users had huge numbers of followers: the maximum number of followers was over 7 million.

[4] Again, 10 is the limit imposed by Pinterest.

[5] See http://pinterest.com/categories/, downloaded May 2013.

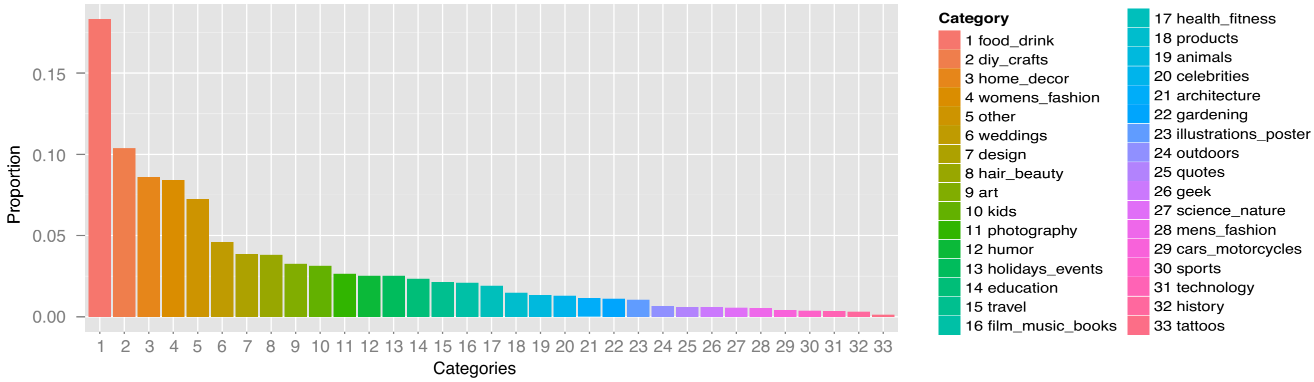

Figure 2. Distribution of content across Pinterest categories. The X axis shows the 33 predefined Pinterest categories, and the Y axis shows the proportion of content in each category.

- Textual comments users associated with pins and boards.

We explored all these approaches, and experimented with different algorithms as well. However using textual comments was not successful, likely because comments on Pinterest are short and scarce. Both other approaches were successful. We now give more detail on how we applied them and how the results compared.

Categorization 1. Users may assign their boards to one of 33 pre-defined Pinterest categories. Further, pins may appear on multiple boards, which in general will belong to different Pinterest categories. These relationships let us:

- represent the topical structure of pins with a vector over the 33 Pinterest categories;
- aggregate the topic vectors of all a user's pins to derive a topic vector for users;
- represent each pre-defined Pinterest category as a binary vector of length 3,142,941 (the number of pins in our sample), where a 1 indicates the pin appeared in the category and a 0 indicates it did not appear;
- compute similarity relationships between pins, between users, and between the pre-defined Pinterest categories.

Categorization 2. Since the same pin may be added to different boards by different users, we can compute interboard relationships based on the number of pin co-occurrences. We then apply a clustering algorithm to these data to derive another topical structure. This method also reveals more detailed topical information, i.e., sub-category relationships, than the first method.

Comparison. We compared the two topical structures. To the extent they are similar, it gives us confidence in the reliability of results we obtain using either of the structures.

Categorization 1 details

Users may categorize their boards using one of the 33 pre-defined Pinterest categories. For example, a user might create a board called "Desserts" and categorize it under "Food & Drink". Users can browse pins by category: pins that were added to a board for a given category may be found under that category. For example, if a user pins a recipe for strawberry shortcake to "Desserts", then the pin for that recipe could be found by browsing the "Food & Drink" category.

Of course, this categorization structure is not perfect: some users don't assign categories to their boards (or use the default "Other" which consists about 7% of pins in the data set), and different users may categorize the same item differently. For example, one might pin a reference to a learning toy to a board categorized as "Education", while another may pin it to a board categorized as "Kids".

Therefore, we take the user defined categorization structure as a starting point to represent pins and users.

1. **Representing pins.** When users repin a pin they find interesting, they are indirectly assigning that pin to multiple Pinterest categories. For example, the pin in Figure 1 has been pinned by five users to five boards: four of these boards were categorized as "Food & Drink", and one was categorized as "Holidays & Events". We define the topic vector of a pin as a normalized count of the categories of the pin:

$$\vec{p_i} = < p_{i,1}, p_{i,2}, ..., p_{i,33} > \quad (1)$$

   where $p_{i,j}$ is the fraction of appearance of jth category on pin p_i. Therefore, the topic vector for the pin from Figure 1 is $< 0.8, 0, .., 0, 0.2, 0, .., 0 >$ where the nonzero entries are the first and the thirteenth.

2. **Representing users.** We represent each user by aggregating and normalizing the topic vectors of all the user's pins:

$$\hat{u} = \frac{\vec{u}}{||\vec{u}||_1}, \quad where \quad \vec{u} = \sum_{\vec{p_i} \in \mathbf{P_u}} \vec{p_i} \quad (2)$$

   and $\mathbf{P_u}$ is the set of pins that belongs to user u.

We emphasize that this representation of topics for pins and users is *community-based*. For example, if one user added a pin to a board categorized as "Geek", and another repinned that pin to a board categorized as "Technology", we would consider that pin to be associated with both topics equally; and from the perspective of the two users, we would say that each had pinned content on "Geek" and "Technology".

This approach is fundamentally different than that of Ottoni et al. [27]. Their approach does not aggregate beyond the individual, simply using whatever categories an individual user had assigned to her boards in analyzing the categories of the user and her pins. In our example, the first user's pin would be categorized only as "Geek", and that user's categories would include only "Geek". Similarly, the second user's (re)pin would be categorized only as "Technology", and her categories would include only "Technology". We believe the community-based categorization approach makes more sense; in any case, this difference means that even when we seem to be studying the same questions as Ottoni et al., our results are not directly comparable.

Categorization 2 details

In the previous approach, boards played only an intermediary role, linking pins to pre-defined categories. However, in the second approach, boards played a primary role, analogous to tags in a folksonomy [22]. That is, adding a pin **P** to a board **B** is analogous to applying the tag **B** to the object **P**. Many board names are used by multiple users, such as "Recipes", "Kids", "Style", and "Sewing". We constructed an undirected weighted graph whose nodes represent (stemmed[6]) board names and whose edges represent a pin appearing on both boards; the edge weight indicates the number of pins that appeared on both boards. Next, we applied the METIS graph partitioning algorithm [17] to the graph, parameterizing it to produce 33 clusters, which enables us to compare its results to the results obtained using the pre-defined Pinterest categories.

We wanted to verify that these clusters were reasonable. A standard way to do this is to measure how distinct the clusters are: the more distinct, the better the clustering. We measured similarity between clusters using cosine similarity. The average cosine similarity of the 33 clusters was 0.25, which indicates that the categories were acceptably distinct.

Comparing the two categorizations

We needed to select a categorization to use for our analysis. Ideally, the two categorizations would be similar, so we could select one for reasons of convenience. To determine similarity, we represent each cluster as a vector over the 33 Pinterest categories. We assigned each stemmed board name in the cluster to one Pinterest category[7]. The entries in the vector then represent the proportion of boards from the cluster in each Pinterest category. The key observation we made is that most of the plots show a single dominant peak: this indicates that the cluster corresponds closely to one pre-defined Pinterest category. For example, one cluster mapped wholly to the category "DIY & Crafts". In a few other cases, a cluster was an interesting mix of intuitively related categories: for example, one cluster had 75% of its content in "Weddings" and 25% of its content in "Holidays & Events".

The automatically derived clusters reveal more fine-grained topical structure. For example, four clusters had a peak for the category "Food & Drink". We examined these clusters and saw that each had a different focus, for example: baking and cakes vs. breakfast vs. salads, soups, and healthy recipes.

Since the comparison between the two topic representations showed that they were quite similar, we decide to use the first one for our analysis. We favored it because (a) it is based directly on pre-defined Pinterest categories, which makes the analyses more comprehensible, and (b) our analyses did not need the more detailed topical structure that the second categorization offers.

[6] Meaning that "bicycles" and "bicycling" are equated, for example.

[7] Since we stem board names, it is possible that a single stemmed board name may belong to multiple categories. In this case, we assign the stemmed board name to the most frequently occurring category. If there is a tie, we pick a tied category at random.

RESULTS AND DISCUSSION

We organize the presentation of our results around our three research questions, *RQ1-Topics*, *RQ2-Specialization*, and *RQ3-Homophily*.

RQ1-Topics

In this section, we give a descriptive overview of the content on Pinterest based on the topic representation we produced. We show how popular the various topics are, illustrate relationships between topics, and investigate whether men and women (collectively) differ in the categories of their pins.

Which topics are most popular?

To reveal the amount of content in each of the 33 Pinterest categories, we summed and normalized the topic vectors of the 3.143M pins in our dataset. Each entry in this vector represents the proportion of content in one category. Figure 2 shows the result.

Content is distributed unequally across the categories. In fact, topic popularity follows a power law ($p < 0.05$, using the Kolmogorov-Smirnov test): "Food & Drink", "DIY & Craft", "Home Decor" and "Women's Fashion" together account for over 45% of the content, while "History", "Tattoos", "Technology", and "Sports" together account for just over 1%. Note also that 7% of content falls into the default "Other" category, meaning it not categorized by users.

How are topics related?

We also wanted to investigate how topics were related. We use the fact that a single pin can appear on boards belonging to different categories as the basis for computing relationships between categories. For example, some users might pin a recipe for the Christmas drink eggnog to boards in the "Food & Drink" category, while others might pin it to "Holiday & Event" boards. We use the frequency of co-appearance of two categories on all pins as a measure of the relatedness of the categories. For each category i, we define C_i as a binary vector of length 3.143M, where each 1 in the vector denotes that the relevant pin was categorized under category i. We compute the similarity of two categories using Jaccard Similarity, where a 1 indicates perfect similarity and a 0 total dissimilarity.

$$sim(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|} \tag{3}$$

Jaccard Similarity is suitable for our purposes since it works well for vectors with many 0 entries, which we have.

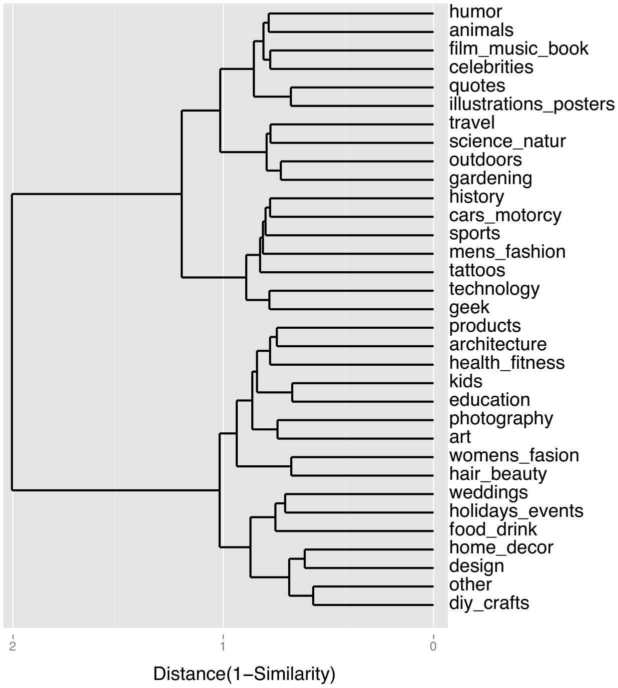

Figure 3. Clustering result of categories

We computed the pairwise similarity of all 33 Pinterest categories. This revealed interesting relationships. (1) "Other" is most similar to the rest of the categories. This makes sense: since "Other" is the default category for boards that users did not categorize, the content of these boards will overlap with content in all the other categories. (2) After "Other", "Art", "Photography", and "Design" are most similar to the remaining categories. This may be due to the fact that most pins on Pinterest indicate visually attractive pictures, which naturally could fit under these broad, almost syntactic categories. (3) "Tattoos", "Sports" and "Cars & Motorcycles" are most distinct from other categories. Two factors might play a role here: first, these categories may be better defined and more narrow; second, these categories may be less likely to co-appear with other categories due to their unpopularity.

To understand inter-topic relationships more deeply, we clustered the topics (using Ward Hierarchical Clustering). The results, shown in Figure 3, help paint a picture of activity on Pinterest. They confirm intuitions about how topics are related: for example "DIY & Crafts", "Home Decor", "Design", "Food & Drink", and "Holiday & Events" cluster together, as do "Outdoors", "Gardening", and "Science & Nature", and "Geek" and "Technology". Some of the topics that cluster closely suggest different purposes of categorization applied by different users. For instance, "Kids" and "Education" may be closely related because some users categorize pins according to the content itself, ("Education"), while others categorize according to the usage of the content, (for "Kids"). In addition, to preview an issue we study in detail next, categories that are proportionally more popular among women such as "Weddings" and "Holidays & Events" tend to cluster together, as do categories proportionally more popular among men, such as "Geek" and "Technology".

While these results may seem intuitive, we think it always is helpful to provide empirical evidence that intuitions are true, since sometimes they aren't.

Do men and women differ in the content they pin?

We based our analysis on the subset of 32.2K users whose gender we knew from their Facebook profiles. As we mentioned, women constitute a larger proportion of our sample than in the general Pinterest population. However, we are not comparing the total amount of content contributed by men and women. Instead, we examine the relative proportion of content men and women pin across the Pinterest categories.

Figure 4 shows the distribution of topics for male and female users. There are noticeable differences in the topics which men (in general) and women (in general) pin. For example, roughly 10% of all pins from men, but only about 3% of pins from women were categorized as "Design". And for men, "Design" was the second most popular category, while for women, it was the 9th most popular category (omitting "Other" in both cases). There are many categories that received differing degrees of interest from men and from women, including "Geek", "History", and "Sports" (more popular among men), and "Kids", "Wedding", and "Holiday & Events" (more popular among women). Indeed, men's and women's pinning activity was significantly different for all categories (at the 0.01 level) except for "Humor", "Tattoos", "Animals", "Quotes", and "Gardening". But "Food & Drink" is the most popular category for both women and men.

In addition, women in general concentrate their pinning activity on fewer categories: their top 5 categories account for just over 56% of all their activity, while the top 5 categories for men account for just under 40% of their activity.

Further, stereotypical men's topics are not even that popular among men! While men devote more attention than women to "Sports", "Technology" and "Cars & Motorcycle", these topics are not even among the 10 most popular categories for men. Instead, male Pinterest users pin more content about "Photography", "Art", "Design", and "Home Decor". We offer several conjectures concerning these results in the *Implications for Research and Design* section below.

RQ2-Specialization

The previous investigations painted a picture of global activity on Pinterest. We now zoom in to the individual level, asking: to what extent do users specialize in particular content topics? Do women and men differ in degree of specialization? What factors, including the degree of specialization, attract more followers?

To what extent do users specialize in particular topics?

Building on the previous research covered above, we investigated the extent to which users concentrate their pinning in a few categories vs. distributing it across many categories.

We measure the diversity of users' interests (the opposite of specialization) by computing the entropy of users' topic vectors. Entropy is the state of the art measure of variety, which is the composition of differences in categories [15]. The entropy of a user topic vector reaches a maximum[8] if a user

[8]Given by $ln(x)$, where $x = 33$ in our case; $ln(33) \sim 3.5$

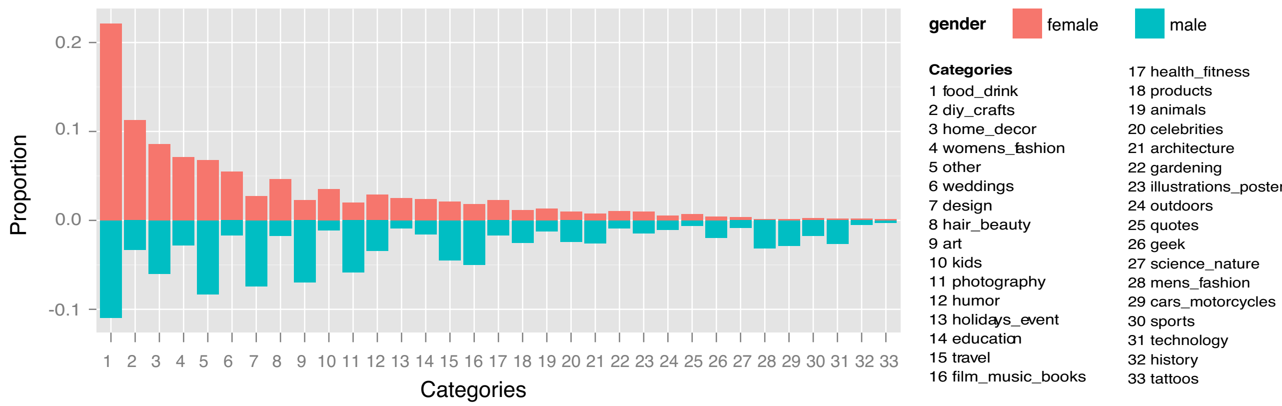

Figure 4. How women's and men's overall pinning activity is distributed across categories. The X axis shows the 33 Pinterest categories, and the Y axis shows the relative proportion of each topic within gender.

pinned the same amount of content from all 33 categories and is equal to zero if all the content is from one category. Entropy is defined as:

$$Entropy(\hat{u}) = -\sum_{j=1}^{33} u_j \log_e u_j \tag{4}$$

where u_j is the jth entry in user vector $\hat{u}$, denoting the fraction of content in jth category.

Intuitively, one would expect users with more pins to cover a wider spectrum of categories; therefore, we group users based on the number of their pins for the topic diversity analysis. Specifically, we divided users into three equal sized groups based on their total number of pins. Figure 5 shows the group intervals and the distributions of the diversity of the groups. Our intuition was borne out: users with larger number of pins do tend to have more diverse interests (differences were significant, $p < .001$). [9] However, there are specialists even among prolific pinners. To take two contrasting examples, one specialist had 1015 pins, with over 90% in "Food & Drink", while a generalist had 1078 pins, with 18% in "Film & Music & Books", 15% in "Food & Drink", 9% in "Other", 8% in "Photography", 6% in "Home Decor", 5% in "Celebrities" etc, spanning 32 categories.

Do women and men differ in degree of specialization?

We next tested whether men and women differed in the diversity of their interests. To handle the greatly differing numbers of men and women in our dataset, we randomly sampled 300 male and 300 female users from each of the three activity-level groups from the previous section.

Table 3 shows the result of comparing the topic diversity of male and female users. In all groups, women pinned significantly more diverse content than men ($p < 0.001$). [10] These results may seem inconsistent with those of Figure 4. However, those results were for the male and the female populations as a whole, while the results of Table 3 aggregate based on individual users. Thus, while men *collectively* have more diverse interests (than women collectively), each *individual* male user is more likely to specialize in specific categories than individual female users.

[9] To ensure that these results were not due to outliers, we did the same analysis with 10 groups of users. The pattern of results was the same, with the exception that the two most active groups were indistinguishable.

[10] We did the same analysis using 10 equal sized groups, and the pattern was the same: women pinned more diverse content for all groups except the least and most active.

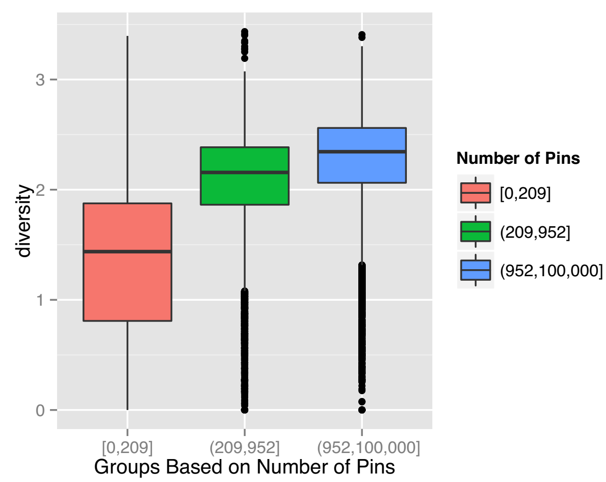

Figure 5. Diversity of 10 equal sized group of users, based on the number of their pins.

Recall that Ottoni et al. [27] found that female Pinterest users were content generalists to a greater extent than males. As we explained above, their results are not directly comparable to ours. However, their analysis was done at the individual level, like our analysis in this section. Thus, in this respect our findings are similar to theirs.

What factors, including the degree of specialization, attract more followers?

The data we gathered and the previous analyses we have done let us investigate what makes a Pinterest user "popular". We use *number of followers* as our popularity metric: since the default home page for a logged in Pinterest user

Group	Group Mean	Gender	Sample Mean
Low	1.327	Male Female	1.040 1.42
Medium	2.078	Male Female	1.82 2.089
High	2.261	Male Female	2.13 2.241

Table 3. Comparison of topic diversity between male and female users. For all three activity levels (Low, Medium, and High), men are significantly less diverse than women.

shows a feed of new pins from the users they follow, the more followers users have, the larger their audience.

We used the following features to model user popularity:

- *gender*: gender of the user;
- *nPins*: number of pins;
- *nBoards*: number of boards;
- *nFollowing*: number of users that the user follows;
- *topics*: the user's normalized topic vector;
- *diversity*: diversity of the user's interest, measured by the entropy of the topic vector.

We use a binary classification model to study the effect of these features on number of followers. We created two classes to predict: the top quartile and the botom quartile of users based on number of followers. We choose a classification model over a regression model because regression models put too much emphasis on predicting the exact number of followers instead of modeling general trends and relationships [3]. Previous work [13] used linear regression to model the effect of several factors on number of followers. However, the model suffers from poor fit (low R^2) and used only basic data from user profiles (like the first four features above) but did not have access to the two features that characterize the content a user pins, *topics* and *diversity*.

In selecting a classification model, our goal was not only high prediction accuracy but also simplicity of model interpretation. We experimented with logistic regression and Random Forest models. The Random Forest model [8] is a widely used and accurate statistical ensemble method that combines multiple simple decision trees, each of which is trained on a bootstrap sample from the entire training data. The Random Forest model performed better and was effective in illuminating the effect of the predictive features. Therefore, we report our findings using this model.

We experimented with models using different number of trees, measuring the error each time. With 25 trees, the error rate was about 0.15, with 100 trees it approached 0.14, and by the time we reached 500 trees, it had converged to 0.136[11]. These low error rates give us confidence in our model.

The Random Forest model provides a robust and accurate way to measure the importance of each predictive feature: the algorithm drops each feature, and then computes the reduction in prediction accuracy using the rest of the features. Figure 7 illustrates the importance of the features. The most

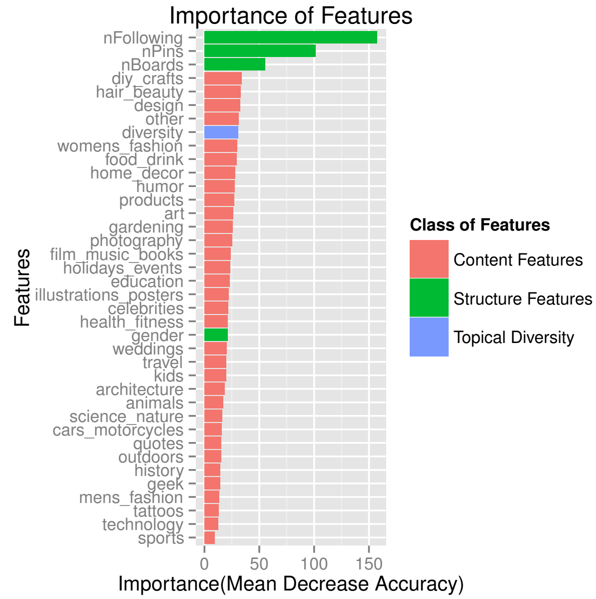

Figure 7. Importance of predictive features

important feature is the number of users a user follows: that is, users who follow a lot of users also get followed a lot. The next two features, number of pins and number of boards, measure quantity of contribution: the more you pin, the more followers you earn. After this, the model mostly ranks the relative importance of pinning to various categories: pinning to "Food & Drink" and "Design" will attract the most followers, and more generally, pinning to popular categories attracts more followers.

Note that gender did not play a critical role in popularity. *Diversity*, on the other hand, was about as important as pinning to the most popular categories. To sum up, the model suggests that to attract lots of followers, you should: follow lots of other pinners, create lots of boards and pin a lot, post on popular topics, and don't concentrate on too few topics.

We also can find the *marginal effect* of predictive features on popularity; this is the effect one feature has on the prediction result when all other features have fixed values. For example, the marginal effect of number of pins is shown in Figure 6: the popularity of a user increases steeply as the number of pins increases to 4000, staying steady thereafter.

The marginal effect of diversity shows a more interesting pattern. When the diversity (measured by entropy) is around 2.4, the probability of being popular reaches a peak. In other words, increasing the diversity of the content you pin is good up to a certain point and thereafter, it becomes a disadvantage to be too diverse. Moreover, when we look at the marginal effect graphs for specific content categories, we see a similar pattern: as the third part of Figure 6 shows, concentrating on "Food & Drink" will gain you more followers until you're devoting about 35% of your pins to that topic; further specialization does no good.

RQ3-Homophily

[11] The error rate is the Out-Of-Bag (OOB) error, which is equivalent to errors of cross validation. [8]

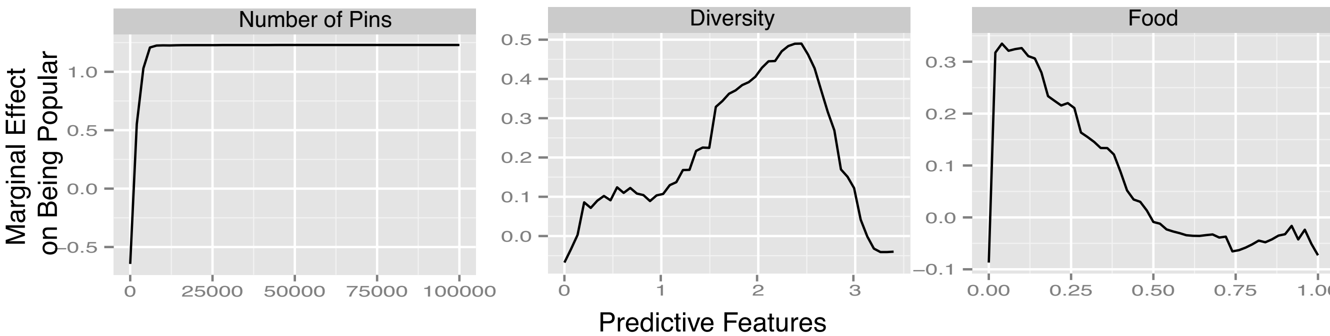

Figure 6. Marginal effect of diversity, food and number of pins on popularity

We were inspired by the observation by Zarro et al. that Pinterest is less of a social network than a *social curation* site [39]; one of their participants reported that "you can like stuff of random people, and you don't have to make a further connection." This suggests that shared interests may be a stronger driver of activity on Pinterest than social connections and interaction.

Based on this suggestion, we studied how homophily of interests influenced the two key Pinterest activities *following* and *repinning*. Specifically, were users more similar in their interests to those they followed than to random Pinterest users? And were users were more likely to repin from users who are similar to them than from random users?

We used the cosine similarity of users' topic vectors as our homophily measure:

$$homophily(\hat{u_1}, \hat{u_2}) = cosine(\hat{u_1}, \hat{u_2}) = \frac{\hat{u_1} \cdot \hat{u_2}}{|\hat{u_1}| \cdot |\hat{u_2}|} \quad (5)$$

We define several terms to facilitate the statement of hypotheses to test our research questions. For each user u, U_{repin} represents the set of users u has repinned, U_{no_repin} represents the set of users u has not repinned, U_{follow} represents the set of users u follows, and U_{no_follow} represents the set of users u does not follow. We then define $S_{repin,u}$ to represent the average cosine similarity between u and U_{repin} and define $S_{no\_repin,u}$, $S_{follow,u}$, and $S_{no\_follow,u}$ correspondingly. Finally, to test our hypotheses, all we need to do is test whether S_{follow} and S_{no_follow} have the same means and whether S_{repin} and S_{no_repin} have the same means, which we can do with a one-sided T-test.

However, while conceptually simple, testing these hypotheses is computationally expensive since S_{no_repin} and S_{no_follow} are very large sets for all the 46.4K users in our sample. We therefore made several simplifications to come up with a feasible computation. *Simplification 1.* We investigated whether we could use a smaller set of users as a "stand-in" for S_{no_repin} and S_{no_follow}. Experimentation showed that sets of size 5000 sufficed. *Simplification 2.* However, a problem remains: we still would have to perform 5000 cosine similarity computations for every user in our sample. To avoid this, we clustered users based on their topic vectors to form groups of users with similar interests. We then used the cluster centroid as the representative of all users in that cluster in the computation of S_{no_repin} and S_{no_follow}, which meant we only had to carry out one statistical test per cluster.

Testing the hypotheses. Different values of the clustering parameter K (the number of clusters the algorithm will create) result in clusters of different size and composition. Therefore, we tested our hypotheses for different values of K. The process we followed to test our hypotheses was:

- For each user, compute the similarity to those users they follow S_{follow} and those users from whom they repin S_{repin} (these are exact values, not approximations).
- For $K = (5, 10, 30, 50, 80, 100)$
  - Obtain a set of K clusters of users.
  - For each cluster C in the set of K clusters:
    - ∗ Approximate U_{no_follow} / U_{no_repin} with 5000 randomly selected users whom no user in C follows / repinned.
    - ∗ Approximate S_{no_follow} / S_{no_repin} as the similarity of the centroid of C to each of the users in U_{no_follow} / U_{no_repin}.
    - ∗ $S_{follows}$ / S_{repin} is the set of similarities of all the users in C to all the users whom they follow / repinned.
    - ∗ Run a t-test to compare S_{repin} and S_{no_repin} / S_{follow} and S_{no_follow}.

Table 4 shows the results. The results for repinning behavior were unequivocal: for every case we tested (a total of 275 clusters), users were more similar to those from whom they repinned than they were to random Pinterest users. This is consistent with the conjecture, based on Zarro et al., that content plays a bigger role than social interaction in driving Pinterest activity. The trend was the same for following behavior, but there were many exceptions: in 63 of the 275 clusters, users were not significantly more similar to those they followed than they were to random users. One factor that could contribute to this is if people follow those whom they know – family and friends – but they do not necessarily share many interests with these people. Indeed, we found that people often repin from users whom they do not follow, over 50% of the repin events in our sample.

IMPLICATIONS FOR RESEARCH AND DESIGN

K	Repin	Follow
5	5	3
10	10	7
30	30	26
50	50	41
80	80	64
100	100	71

Table 4. Result of homophily test. Repin: # of clusters where users are more similar to users from whom they repin than random users. Follow: # of clusters where users are more similar to users whom they follow than random users.

Using Pinterest as a research context, we set out to investigate how two factors, *specialization* and *homophily*, influence the attraction of audience. We also considered how gender might mediate the effects of these factors. We articulated specific research questions and carried out a quantitative study to answer these questions.

We built on a set of content categories provided by Pinterest to create a community-based representation of Pinterest content. This let us characterize the content on the site, including quantifying relationships between categories. We then used our representation of topics as the basis for a number of statistical analyses. Our key findings included:

- Women and men differed both in the types of content they collected, and the degree to which they specialized in certain types of content; interestingly, despite these gender differences, men were not particularly interested in stereotypically male topics.
- Sharing diverse types of content is correlated with a larger number of followers, but only up to a certain point; after that, more diversity is not helpful. This result helps refine researchers' understanding of the role of content diversity in attracting an audience, particularly in light of the differing findings of prior work (compare [16] to [37]).
- Homophily of interests is a major driver of repinning: people repin from other users who share their interests; people also share the interests of those they follow, but there are more exceptions to this rule.

Our results raise a number of issues deserving additional research and design attention.

Most generally, our quantitative results illuminate *what* Pinterest users do, but not *why* they behave this way or how they understand their own behaviors. Followup work, particularly using qualitative methods, is needed to answer the latter questions. Some of the most compelling questions concern gender; for example, why are male Pinterest users not particularly interested in stereotypically male content?

Before attempting to answer this question, we should be careful in making assumptions about "stereotypical male" content. A useful guide here is work by Lam et al. [19], which investigated gender differences among Wikipedia editors. Rather than taking it for granted which topics were of interest to men or women, they found an external data source that allowed them to compute this objectively within a specific domain (movies). We recommend following this approach in future studies of Pinterest. Once this is done, and assuming our results hold, there are several possible conjectures worth investigating. (i) As we discussed in *Theoretical Foundations and Related Work*, the dominant popular portrayal of Pinterest is as a site for women's interests. Therefore, perhaps men who choose to join Pinterestare disproportionately more interested in steroetypically female topics than most men. This is a *selection effect* explanation; some men may perceive Pinterest as "not my kind of place" [4]. (ii) Women make up a large majority of Pinterest users and pin content they find interesting. This content is visually prominent for all users, female and male, and thus influences repinning behavior by all users. This *social influence* has been observed in tagging systems: the tags users apply are strongly influenced by the community tags that are visible to them [31]. Notice that neither of these explanations presumes that the differences between male and female Pinterest users' behavior are due to *inherent* gender differences; instead, they are based on different types of social context.

Second, more research is needed on the utility of topic specialization vs. diversity. For example, are users aware of being topic specialists or generalists? If so, how do they think about this dichotomy (or is it a spectrum?): do they consciously seek a particular level of specialization, and are their self perceptions of their degree of specialization consistent with quantitative measures? Interviews would an appropriate way to study these questions.

Third, we explored two different ways to represent topics on Pinterest, one based on pre-defined Pinterest categories and one consisting of algorithmically derived clusters. While we did not use the second representation of topics for this research, it has the intriquing property that it reveals more fine grained topical structure. In-depth explorations with this representation would illuminate Pinterest users' own emergent categorizations of content, particularly distinctions (of meaning and use) within the pre-defined categories.

Finally, nearly all social network sites provide mechanisms to recommend or filter content and potential social connections. Our results highlight the importance of effective methods for early personalization of site content; if new users perceive a site as "not my kind of place", they are likely to abandon it. While improved personalization methods might make Pinterest "more friendly" to (some) men, the effect is more general and gender neutral: if done well, all users would experience content that matches their interests. We suggest several specific personalization techniques worth exploring:

- Focus more on recommending content from users with similar interests, rather than those whom the user is socially close to.
- Maintain an appropriate amount of diversity in content recommendations.
- Promote the ability of users to follow single *boards*, rather than users.
- Use the category of items as a more prominent visual organizing principle. Even better, try using an automatically derived, more detailed categorization like we described in the *Representing Topics* section: for example, woodworking and jewelry projects both may be included under "DIY and Crafts", yet may appeal to different audiences.

REFERENCES

1. *Gender and Power in On-line Communication*. Blackwell Publishing Ltd, 2008.
2. Eliciting and Focusing Geographic Volunteer Work. *CSCW* (2010).
3. Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. Discovering value from community activity on focused question answering sites. In *KDD*, ACM (2012).
4. Antin, J. My kind of people?: perceptions about wikipedia contributors and their motivations. In *CHI*, ACM (2011).
5. Arguello, J., Butler, B. S., Joyce, E., Kraut, R., Ling, K. S., Rosé, C., and Wang, X. Talk to me: foundations for successful individual-group interactions in online communities. ACM (2006).
6. Barabási, A. Emergence of scaling in random networks. *Science* (1999).
7. Beckwith, L., Kissinger, C., Burnett, M., Wiedenbeck, S., Lawrance, J., Blackwell, A., and Cook, C. Tinkering and gender in end-user programmers' debugging. In *CHI*, ACM (2006).
8. Breiman, L. Random Forests. *Machine Learning 45*, 1 (Oct. 2001), 5–32.
9. Caverlee, J., and Webb, S. The Social Honeypot Project: Protecting Online Communities from Spammers. In *WWW* (2010).
10. Cunha, E., Magno, G., Almeida, V., Gonçalves, M. A., and Benevenuto, F. A gender based study of tagging behavior in twitter. In *HT*, ACM (2012).
11. Daniele Quercia, Jonathan Ellis, Licia Capra, J. C. In the mood for being influential on twitter, 2011.
12. Demartini, G. Finding experts using wikipedia. *ISWC/ASWC2007* (2007).
13. Gilbert, E., Bakhshi, S., Chang, S., and Terveen, L. I Need to Try This!: A Statistical Overview of Pinterest. *CHI* (2013).
14. Goel, S., and Mason, W. Real and perceived attitude agreement in social networks. *Journal of Personality and Social Psychology 99*, 4 (2010), 611–621.
15. Harrison, D., and Klein, K. What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *Academy of Management Review* (2007).
16. Hutto, C., Yardi, S., and Gilbert, E. A longitudinal study of follow predictors on twitter. In *CHI* (2013).
17. Karypis, G., and Kumar, V. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing* (1998).
18. Kivran-Swaine, F., and Naaman, M. Network properties and social sharing of emotions in social awareness streams. In *CSCW* (2011).
19. Lam, S. T. K., Uduwage, A., Dong, Z., Sen, S., Musicant, D. R., Terveen, L., and Riedl, J. WP:clubhouse?: an exploration of Wikipedia's gender imbalance. In *WikiSym*, ACM (2011).
20. Leskovec, J., and Faloutsos, C. Sampling from large graphs. *KDD* (2006).
21. Liben-Nowell, D., and Kleinberg, J. The link prediction problem for social networks. In *CIKM* (2003).
22. Macgregor, G., and McCulloch, E. Collaborative tagging as a knowledge organisation and resource discovery tool. *Library review* (2006).
23. Marcotte, A. A social network of one's own. *The American Prospect* (March 2 2012). `http://prospect.org/article/social-network-ones-own`.
24. Masli, M., Priedhorsky, R., and Terveen, L. Task specialization in social production communities: The case of geographic volunteer work. *ICWSM* (2011).
25. McPherson, M., L. S.-L., and Cook, J. Birds of a feather: Homophily in social networks. *Annual review of sociology* (2001), 415–444.
26. nathanjurgenson. Pinterest and feminism. *thesocietypages.org* (March 5 2012). `http://thesocietypages.org/cyborgology/2012/03/05/pinterest-and-feminism/`.
27. Ottoni, R., Pesce, J., and Casas, D. L. Ladies First: Analyzing Gender Roles and Behaviors in Pinterest. *ICWSM* (2013).
28. Ren, Y., Harper, F. M., Drenner, S., Terveen, L. G., Kiesler, S. B., Riedl, J., and Kraut, R. E. Building member attachment in online communities: Applying theories of group identity and interpersonal bonds. *MIS Quarterly* (2012).
29. Rode, J. A. A theoretical agenda for feminist hci. *Interacting with Computers* (2011).
30. Selwyn, N. Whats in the box? exploring learners rejection of educational computing. *Educational Research and Evaluation* (1998).
31. Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F. M., and Riedl, J. Tagging, communities, vocabulary, evolution. *CSCW* (2006).
32. Stevenson, S. Pinterest for dudes. *slate.com* (April 1 2013). `http://www.slate.com/articles/technology/technology/2013/04/gentlemint_manteresting_tapiture_is_it_sexist_to_create_a_pinterest_for.single.html`.
33. Suh, B., Hong, L., Pirolli, P., and Chi, E. H. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *SocialCom* (2010).
34. Tan, D. S., Czerwinski, M., and Robertson, G. Women go with the (optical) flow. In *CHI*, ACM (2003).
35. Thelwall, M. Social networks, gender, and friending: An analysis of MySpace member profiles. *Journal of the American Society for Information Science and Technology 59*, 8 (2008), 1321–1330.

36. Wanenchak, S. Pinterest and playing gender. *thesocietypages.org* (April 5 2013). `http://thesocietypages.org/cyborgology/2013/04/05/pinterest-and-playing-gender/`.

37. Wang, Y.-C., and Kraut, R. Twitter and the development of an audience. In *CHI* (2012).

38. Williams, M. E. Pinterests gender trouble. *salon.com* (May 2 2012). `http://www.salon.com/2012/05/02/pinterests_gender_trouble/`.

39. Zarro, M., Hall, C., and Forte, A. Wedding dresses and wanted criminals: Pinterest. com as an infrastructure for repository building. *ICWSM* (2013).

40. Ziegler, C.-N., McNee, S. M., Konstan, J. A., and Lausen, G. Improving recommendation lists through topic diversification. In *WWW* (2005).