Sharing Topics in Pinterest: Understanding Content Creation and Diffusion Behaviors

Jinyoung Han[1], Daejin Choi[2], A-Young Choi[3], Jiwon Choi[4], Taejoong Chung[2]
Ted "Taekyoung" Kwon[2], Jong-Youn Rha[4], Chen-Nee Chuah[1]
Dept. of Electrical & Computer Engineering, University of California, Davis[1]
Dept. of Computer Science & Engineering, Seoul National University[2]
Business School, Sungkyunkwan University[3], Dept. of Consumer Science, Seoul National University[4]
rghan@ucdavis.edu, djchoi@mmlab.snu.ac.kr, {cay0908, jiwon}@snu.ac.kr,
tjchung@mmlab.snu.ac.kr, {tkkwon, jrha}@snu.ac.kr, chuah@ucdavis.edu

ABSTRACT

Pinterest provides a social curation service where people can collect, organize, and share content (pins in Pinterest) that reflect their interests. This paper investigates (1) the differences in pinning (i.e., the act of posting a pin) and repinning (i.e., the act of sharing other user's pin) behaviors by topics and user gender, and (2) the relations among topics in Pinterest. We conduct a measurement study using a large-scale dataset (1.6 M pins shared by 1.1 M users) in Pinterest. We show that there is a notable discrepancy between pinning and repinning behaviors on different topics. We also show that male and female users show different behaviors on different topics in terms of dedication, responsiveness, and sentiment. By introducing the notion of a *Topic Network (TN)* whose nodes are topics and are linked if they share common users, we analyze how topics are related to one another, which can give a valuable implication on topic demand forecasting or cross-topic advertisement. Lastly, we explore the implications of our findings for predicting a user's interests and behavioral patterns in Pinterest.

1. INTRODUCTION

Over the past decade, online social networks (OSNs) have become platforms to create and maintain social relationships, disseminate content, exchange opinions, share news or images, and conduct political campaigns. This in turn has led researchers to examine how interests are shared/propagated among users [5, 10], revealing valuable insights into understanding users and their interests. Such studies form a fertile ground for industry to develop new online services or to grow their businesses by identifying potential consumers who may be willing to use their services, or by recommending useful goods/services [5, 10, 12, 23]. A recent New York Times article reported that traditional retailers like Target and Walmart have started to recognize the importance of such endeavors and sought their partners among OSN companies [11].

Recently, *Pinterest*, an emerging OSN, has been reported to become the fastest growing web site to reach 10 million visitors [6] and the third most popular OSN in the United States, behind Facebook and Twitter [16]. Pinterest provides a social curation service where people can collect, organize, and share content that reflect their tastes or interests [9, 10, 25]. Each content in Pinterest is called a pin, and a pinboard is a collection of pins organized by a user, each of which belongs to one of the 33 categories (or topics[1]) defined by Pinterest, varying from "food & drink" to "travel".

The great upsurge and popularity of Pinterest has spurred research into its usage patterns [4, 9, 10, 13, 15, 25], revealing valuable insights into the characteristics of Pinterest. One of the unique properties of Pinterest is that interests (or topics) drive user activities and connectivities (among users or their pins) in Pinterest [10, 22], in contrast to other OSNs such as Facebook. However, relatively little attention was paid to how different *"topics"* in Pinterest are shared by (i) *pinning* (i.e., the act of posting a pin) and *repinning* (i.e,. the act of sharing other user's pin) behaviors, and (ii) gender differences in such behaviors, which might be the key to understanding what attracts the users to post and to interact with one another in Pinterest in the first place.

We believe that investigating 'which' users post/share 'what' content, and 'with whom' those content are shared can provide valuable information for online retailers to enhance their marketing strategies. From consumer perspectives, content posting/sharing behaviors of users can be interpreted as behaviors that reflect the latent factors such as their needs, interests, and desires. In this sense, understanding users' posting/sharing behaviors can shed light on users' interests beyond their words, which in turn can be used to better satisfy them. By using Pinterest as a research context, we set out to investigate pinning (or posting) and repinning (or sharing) behaviors that can be construed as information creation and diffusion behaviors, respectively [18]. It has

[1]In this paper, we regard a Pinterest category (e.g., "history" or "travel") as a particular topic (see Table 1).

been reported that information creation and diffusion behaviors may happen due to the different motivations of users, the former requiring more efforts and dedication than the latter [18].

In this paper, we strive to shed light on such issues by performing a large-scale trace-driven analysis on topics shared by users in Pinterest and users' pinning/repinning behaviors. In particular, we seek to answer the following three questions:

- **Q1 - Topic Curated/Shared:** How are different topics shared by pinning/repinning behaviors and by male/female users? Are there any similarities or differences in user behaviors across different topics?
- **Q2 - Topic Relation:** How are topics related to each other? Can those topics be clustered into groups, and how?
- **Q3 - Application:** What would be the potential applications of the observed user behaviors on different topics? Can we forecast the topic usage/curation pattern of each user in Pinterest?

To address the above questions, we first build a bipartite network consisting of two types of nodes: (i) topics and (ii) users interested in those topics (Figure 1). If a user has a pin in a particular topic, there is a link between the user and its corresponding topic in the bipartite network. Projecting [26] a topics-users bipartite network into the topic space results in a *Topic Network (TN)*, whose nodes are topics and are linked if they share at least one common user (Figure 1). The rationale behind the proposed method is that two topics linked in the TN are likely to be shared by common users who are interested in both topics. This TN model has great utility in resource allocation of online retailers, e.g., via topic demand forecasting or cross-topic advertisement.

We conduct a measurement study using a dataset (1.6 M pins shared by 1.1 M users) that we collected by crawling web pages from Pinterest. We fetched the web information of all the newly-posted (i.e., pinning) and shared (i.e., repinning) pins in each category (topic) of Pinterest from June 5 to July 18, 2013. Using the dataset, we analyze: (1) the differences in the pinning and repinning behaviors by topic characteristics and user characteristics (gender), and (2) the relations among topics in Pinterest. In addition, we explore the implications of our findings to predict a user's interests and behavioral patterns in Pinterest.

We highlight the main contributions and key findings of our work as follows:

- **Topic Curated/Shared:** We comprehensively investigate how different topics in Pinterest are curated (and shared) from the perspectives of (i) pinning and repinning behaviors and (ii) gender differences in such behaviors, which shows completely different patterns in terms of dedication, responsiveness, and sentiment. We find female users play more roles in repinning (disseminating the existing content) than pinning (creating a new content) whereas male users play more roles in pinning than repinning. We observe the notable differences in the pinning and repinning behaviors: (1) users' efforts on pinning are likely to be more skewed in some topics than repinning, and (2) repinning users (or *repinners*[2]) tend to show more positive sentiments than pinning users (or *pinners*), implying that users who engage in diffusing content tend to be more amiable. We believe this analysis can provide valuable implication on what attracts users to post/share content and to interact with one another in Pinterest-like social curation service.
- **Topic Relation:** To model the relations among topics, we apply a network-theoretic approach and propose the notion of a *Topic Network (TN)*. Based on the TN model, we analyze (i) how topics are related to one another, and (ii) how topics are clustered into groups (or communities), which can be used in identifying hidden (but important) links among the topics (or interests) towards topic demand forecasting or cross-topic advertisement. We find some topics (e.g., "animals", "film, music & books", or "travel") have more links (to other topics) than the others, each of which plays a role as a hub (or a portal) among the topics in the TN. We also identify which topics belong to which communities (e.g., "food & drink", "health & fitness", or "hair & beauty" belong to the same community), which can give valuable implications for online retailers to develop targeted-advertisement or cross-selling services.
- **Application:** We explore the implications of our findings for predicting which topics a user will be interested in the future. Our trace-driven study for predicting topic consumption patterns in Pinterest demonstrates that the proposed TN model (that reflects the collective opinions of other like-minded users) is useful in accurately predicting a user's interest and behavioral pattern.

The rest of this paper is organized as follows. After reviewing the related work in Section 2, we describe our measurement methodology in Section 3. We then present our results on how different topics are shared by pinning/repinning behaviors and by male/female users, and how topics are related to one another in Sections 4 and 5, respectively. We finally suggest prediction models to forecast how topics are shared in Pinterest in Section 6.

[2]In this paper, pinning and repinning users are referred to as *pinners* and *repinners*, respectively.

2. RELATED WORK

Despite its young age, Pinterest has attracted great attention [6, 16]. The huge popularity of Pinterest is attributed to its unique properties [6]. First, it is reported that over 80% of Pinterest users are female [4, 9, 10], which exhibits a different demographic distribution compared to other OSNs such as Facebook or Twitter. This allows researchers to investigate the gender differences in Pinterest [4, 9, 10, 15]. Ottoni *et al.* observed that female users are more active and invest more efforts in bi-directional social links than male users in Pinterest [15]. Gilbert *et al.* showed that female users share more pins but have fewer followers than male users [9]. Han *et al.* observed different preferences of male and female users on different topics (or categories) in Pinterest; the portions of male and female users are significantly different across different topics [10]. Chang *et al.* investigated which topics are popular to male and female users in Pinterest, and showed that male and female users differ in collecting content across different topics [4]. We go one step further; while previous work (e.g., [10] or [4]) showed general preferences of male and female users on different topics, we perform an in-depth analysis on topics shared by male and female users with *different motivations* (i.e., (i) pinning; content creation, and (ii) repinning; content diffusion). This analysis may shed light on what attracts users to post/share content and to interact with one another in Pinterest-like social curation services.

Second, Pinterest supports a social curation service that allows users to collect, organize, and share content that reflect their tastes or interests [4, 10, 24, 25]. A growing number of popular Internet services have started to support *"curation"*, which is a process of searching/collecting information, organizing the collected information in meaningful or personal ways, and creating values beyond the sum of assets. A variety of valuable information on the Internet has made curation one of central elements for innovation and creativity [13, 19]. Pinterest provides a curation platform with social functionalities which is called *"social curation"*: users can follow other curators whose content they find interesting [4, 10, 13, 25]. Linder, Snodgrass, and Kerne conducted an interview with twenty Pinterest users and found that social curation allows users to engage in the process of everyday ideation; they use collected digital objects as creative resources to develop ideas for shaping their lives [13]. Han *et al.* investigated how Pinterest curators collect and curate pins in terms of number of pins, boards, categories, and followings/followers, and showed that the curators' efforts on pinning are skewed in a few categories [10]. Zhong *et al.* showed that curators with consistent activities and diverse interests attract more followers [25]. Chang *et al.* also found that sharing diverse types of content increases the number of followers up to a certain point [4]. Zhong *et al.* investigated how Pinterest (as a new social curation platform) can benefit from social bootstrapping by copying links from already established OSNs like Facebook [24]. Ottoni *et al.* analyzed cross-OSN user behaviors between Pinterest and Twitter, and showed that uses likely to generate new content in Pinterest, and then spread it in Twitter [14]. We analyze how Pinterest curators show different behaviors in content creation (i.e., pinning) and diffusion (i.e., repinning) on different topics. In addition, we apply the insights learned from our study to develop models for predicting which topics an individual Pinterest curator will be interested in.

Based on the curator behaviors and content properties, some studies have tried to predict users' future activities in Pinterest [10, 12, 23]. Kamath *et al.* proposed a supervised model for board recommendation in Pinterest, and showed that using social signals such as 'likes' can achieve a higher recommendation quality [12]. Han *et al.* showed that the properties of pins (e.g., category or source) are more important factors than those of users (e.g., number of followers a user has) in predicting which pins an individual user will be interested in the future [10]. Zhong *et al.* proposed models to predict whether a user will be interested in repinning the given pin [23]. Given a user and an image repinned by her, they also suggested models for predicting which category she will repin the image into [23]. We introduce a new notion — a topic network — that represents the relations among topics in forecasting 'which topics' an individual user will be interested in, which might have an important implication on topic demand forecasting or cross-topic advertisement.

Lastly, another interesting property of Pinterest is that interests drive user activities or connectivities (among pins or users) in Pinterest [10, 22], in contrast to other popular OSNs such as Facebook. Zarro, Hall, and Forte reported that one participant in the interview mentioned that Pinterest was about *what* they enjoyed, not about *who* they were [22]. Han *et al.* showed that sharing pins in Pinterest is mostly driven by pin's properties like its topic, not by users' characteristics such as the number of followers [10]. This was confirmed by Gelley and John [8], who showed that following is not significantly utilized in content sharing in Pinterest. We focus on 'what' topics (or interests) are shared in Pinterest, and characterizes user behaviors on different topics in terms of dedication, responsiveness, and sentiment. We further investigate the relations among topics (or interests) established by (i) different levels of dedication and (ii) gender differences, which can be used in capturing common interests of users in Pinterest.

3. METHODOLOGY

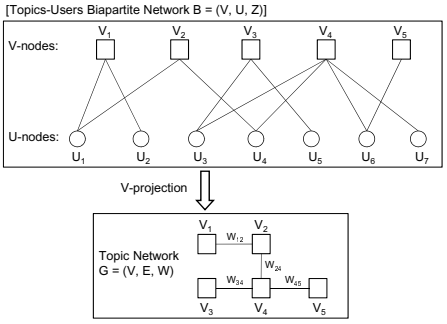


Figure 1: An example of a topics-users bipartite network B, as well as its V projection, is illustrated. V represents the set of topics and U is the set of users associated in those topics. An undirected weighted graph $G = (V, E, W)$ represents a notion of topic network (TN) where V is the set of topics, and E is the set of (undirected) edges between two topics.

In this section, we first illustrate how to model topics and their associated users. We also explain our measurement methodology for data collection, and describe the dataset used in this paper.

3.1 The Model: Topics & Users

To describe topics and users in Pinterest, we first consider a topics-users bipartite network $B = (V, U, Z)$ whose nodes are divided into two disjoint sets V and U, such that every edge in Z connects a node in V to one in U [26] (Figure 1). Here, V and U represent the sets of topics and users, respectively. If a user has pin(s) in a particular topic, there is a link in Z between the user (in U) and its corresponding topic (in V) in the topics-users bipartite network. Prior work reported male and female users show different characteristics in terms of user activities or connectivities in Pinterest [4, 9, 10, 15]. We conjecture that male and female users may also have different tastes for different topics, and hence we consider two divided user sets in the model for male and female users: U_{male} and U_{female}. We also conjecture that pinning and repinning may happen due to different motivations of users; i.e., pinning is likely to create content while repinning is likely to share or distribute content. Rha [18] reported that two information-related activities (i.e., (i) information creation; pinning in our case, and (ii) information diffusion; repinning in our case) show different characteristics, e.g., level of dedication. Thus, we consider two ways to connect a node in V to another in U: Z_{pin} and Z_{repin}. Finally, we build four topics-users bipartite networks: (i) $B_{male,pin} = (V, U_{male}, Z_{pin})$ to model pinning of male users, (ii) $B_{female,pin} = (V, U_{female}, Z_{pin})$ for female users' pinning, (iii) $B_{male,repin} = (V, U_{male}, Z_{repin})$ for male users' repinning, and (iv) $B_{female,repin} = (V, U_{female}, Z_{repin})$ for female users' repinning.

To show the relations in a particular set of nodes (i.e., V or U), bipartite networks can be compressed by one-mode projection [26]. That is, the one-mode projection onto V (V projection for short) results in a network that consists of nodes only in V where the nodes are connected if they have at least one common node (i.e., user) in U. See Figure 1 as an illustrative example of a bipartite network B and its V projection. We assume that an undirected weighted graph $G = (V, E, W)$ resulted from the V projection represents a notion of a topic network (TN) where V is the set of topics and E is the set of (undirected) edges between two topics. An edge $E_{i,j}$ in a TN exists between two topics V_i and V_j if there is at least one user who has pins both in V_i and in V_j. The rationale behind the proposed method of abstraction is that, if two topics are related to each other, they will have common users (who are interested in both topics). To understand the (statistically) meaningful relations among topics, we only consider the edges (in the TN) that have more users than the ones in the uniform case where each edge has the same number users (i.e., users are uniformly distributed in the edges of the TN). In transforming from a bipartite network into a one-mode projection, there can be a loss of information; *weighted projection* is one way to remedy this problem [26]. To this end, we define the weight $W_{i,j}$ of a given edge $E_{i,j}$ as the Jaccard coefficient between two topics V_i and V_j, $\frac{|U(V_i) \cap U(V_j)|}{|U(V_i) \cup U(V_j)|}$, where $U(V_i)$ and $U(V_i)$ are the sets of users which are associated with topics V_i and V_j, respectively. Projecting [26] our four topics-users bipartite networks $B_{male,pin}$, $B_{female,pin}$, $B_{male,repin}$, and $B_{female,repin}$, into the topic space, we finally obtain four topic networks: (i) $TN_{male,pin}$, (ii) $TN_{female,pin}$, (iii) $TN_{male,repin}$, and (iv) $TN_{female,repin}$, respectively.

3.2 Data Collection and Dataset

Data Collection. Since Pinterest does not provide an official API for data collection, we developed web crawling software. Our crawling software fetched web pages in Pinterest, from which the relevant information is extracted; for instance, the data about each pin can be extracted. At the moment of our data collection, we found that Pinterest shows all the recent activities including pinning, repinning, and commenting in the menu of each category in the chronological order. To capture all the pinning activities, we fetched 10 recent web pages periodically (every five minutes) from the menu of each category not to miss any newly posted pins. If user B shares an original pin from user A, Pinterest provides a link of user B's pinboard to the original pin page (of user A); hence we could find and fetch the corresponding (shared) pin pages by user B. Whenever a pin of a user is shared by another user, we obtained the corresponding shared pin information (i.e., its category, description, and comments) as well as the corresponding user information (i.e., his/her description, gender,

1	diy & crafts	2	food & drink	3	education	4	animals	5	health & fitness
6	design	7	architecture	8	products	9	art	10	home decor
11	film, music & books	12	women's fashion	13	humor	14	quotes	15	men's fashion
16	gardening	17	hair & beauty	18	science & nature	19	technology	20	travel
21	cars & motorcycles	22	geek	23	shop	24	weddings	25	outdoors
26	celebrities	27	tattoos	28	photography	29	illustrations & posters	30	kids
31	history	32	sports	33	holidays & events				

Table 1: Pinterest categories (topics) with indexes are summarized.

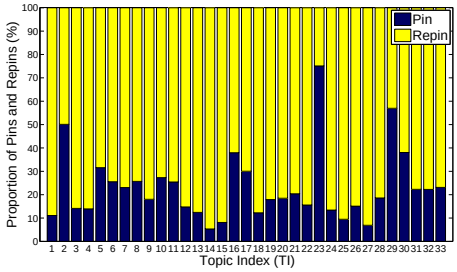


Figure 2: Proportion of pins and repins across different topics are plotted.

number of followers, number of pins, etc). In this way, we could capture nearly all the pinning and repinning activities of users for each topic during our measurement period. To identify the gender of users, we used external links to Facebook and Twitter, which can be found in the profile pages of users. By querying Facebook and Twitter through their APIs, we could obtain the gender information of Pinterest users if available.

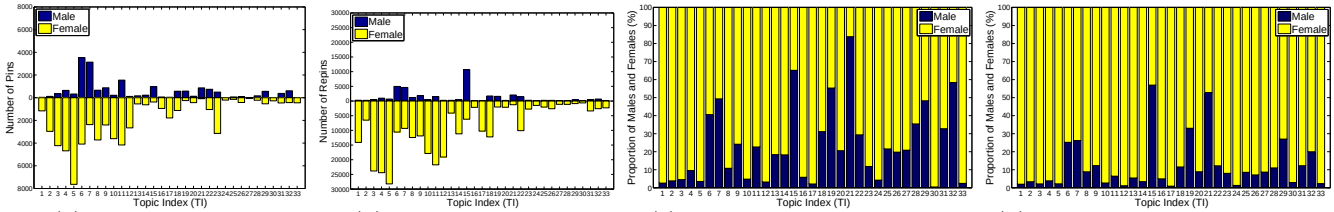
Dataset. We had collected the dataset for 44 days from June 5 to July 18, 2013, which contains 346,305 (original) pins and their 1,215,054 repins, shared by 1,051,054 users. The user dataset collectively contains 2,908,107,606 pins, 595,489,616 followers, 182,381,056 followings, 708,657 Facebook links, and 104,308 Twitter links. We obtained the gender information of 225,382 users. The numbers of male and female users are 23,215 and 202,167, respectively. Table 1 summarizes the Pinterest topics with their indexes, which are sorted in terms of the number of corresponding pins/repins. In addition to the 32 Pinterest topics, we investigate one more special menu in Pinterest, *"shop"*, which is directly related to online stores. Note that pins in the shop menu contain the prices of the corresponding items as well as links to the websites of the online stores.

4. TOPIC SHARING PATTERN

In this section, we discuss our observations on the differences in pinning and repinning behaviors in terms of topics and gender, to answer the first question, *Q1 - Topic Curated/Shared*. We first investigate the ratio of the number of pins (or repins) to the sum of numbers of pins and repins in each topic in Figure 2. In most cases, the portion of repins is higher than that of pins; the average portion of repins across the 33 topics is 77%. However, the portions of pins are higher than those of repins in "food & drink" (Topic Index (TI) 2), "shop" (TI 23), and "illustration & posters" (TI 29); users in those topics tend to upload new content more than share them. For example, designers who are interested in "illustration & posters" (TI 29) may want to upload their own paintings; online retailers may be just interested in uploading their products in "shop" (TI 23).

We next examine the numbers of pins and repins by male/female users across different topics in Figures 3(a) and 3(b), respectively. We also plot the portions of male and female users in pinning and repinning in each topic in Figures 3(c) and 3(d), respectively. Overall, female users tend to post and share more pins than male users as shown in Figures 3(a) and 3(b). Note that the sample size for "kids" (TI 30) by male pinner is not significant, hence we exclude it in the following analyses. When we look at Figures 3(c) and 3(d), we find that male and female users play different roles in pinning (creating a new content) and repinning (disseminating the existing content) across different topics in Pinterest. We observe that the portions of male users in pinning are higher than those in repinning, which implies that male users play more roles in pinning than repinning. The portions of male users in "design" (TI 6), "architecture" (TI 7), "men's fashion" (TI 15), "technology" (TI 19), "cars & motorcycles" (TI 21), "illustration & posters" (TI 29), and "sports" (TI 32) are higher than the ones in the other topics both in pinning and repinning. In particular, the portions of male users in "men's fashion" (TI 15) and "cars & motorcycles" (TI 21) are even higher than 50% both in pinning and repinning, meaning that those two topics are male-dominant. On the other hand, "sports" (TI 32) shows an interesting pattern; while the portion of male users in pinning is higher than 50%, the portion of male users in repinning is less than 20%. This indicates that content in "sports" (TI 32) are likely to be uploaded by male users but (mostly) shared by female users. Some topics show female-dominant characteristics; for example, ~~the portion of female users in "kids" (TI 30) is 99% in pinning;~~ the portions of female users in "women's fashion" (TI 12) and "weddings" (TI 24) are 99% in repinning; the portions of female users in "hair & beauty" (TI 17) are around 99% both in pinning and repinning. Interestingly, ~~"kids" (TI 30) and~~ "weddings" (TI 24) is the topic that might be generally relevant to both male and female users, but male users are not interested in uploading or sharing content in the topic.

4.1 Pinning: Creating a New Content



(a) Number of pins (b) Number of repins (c) Male and female pinners (d) Male and female repinners

Figure 3: Numbers and Proportions of pins/repins and male/female users across different topics are plotted, respectively.

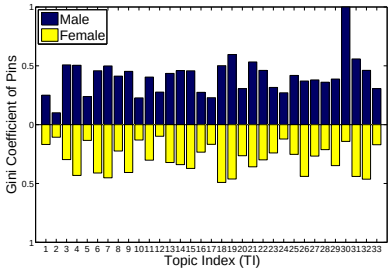
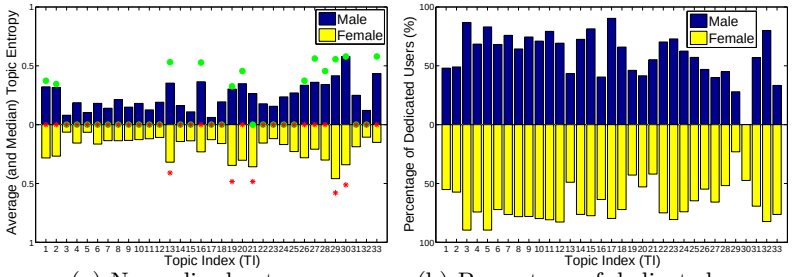


Figure 4: Gini coefficients of pins are plotted on each topic.



(a) Normalized entropy (b) Percentage of dedicated users

Figure 5: Topic concentration in pinning is explored in terms of topic entropy and percentage of dedicated users for each topic.

We now analyze users' pinning behaviors on different topics. We first investigate the distribution of the numbers of pins among users in each topic by calculating the *Gini coefficient of pins*, a well-known indicator to evaluate the disparity of the income distribution in Economics [7]. The Gini coefficient is within the range of [0, 1], where 0 and 1 indicate a perfect uniform distribution (where all values are the same, e.g., everyone has the same income (or pins)) and an extremely skewed distribution (e.g., only one person has all the income (or pins), and all the others have none), respectively [7]. Figure 4 shows the Gini coefficient of pins in each topic, each of which (for a particular topic) is calculated based on the distribution of the numbers of pins each male or female user has posted on the topic. As shown in Figure 4, the Gini coefficients of pins in many topics are lower than 0.5, which signifies that most of Pinterest users contribute to posting pins without substantial disparity. Note that the "food & drink" (TI 2) shows the lowest Gini coefficient (< 0.1); users tend to evenly contribute to posting pins on that topic. The Gini coefficients of pins posted by female users are mostly lower than those posted by male users, implying that female users tend to contribute more evenly in pinning. Interestingly, some topics contributed by male users (e.g., "technology" (TI 19) and "history" (TI 32)) show relatively skewed distributions; a small portion of male users on those topics may be specialists on such topics, and are likely to post pins much more than the others. Note that the Gini coefficient of pins posted by male users in "kids" (TI 30) is 1.0 since there is only one male pinner in "kids" (TI 30). Likewise, a small portion of female users on particular topics (e.g., "science & nature" (TI 18), "technology" (TI 19), "history" (TI 31), and "sports" (TI 32)) are likely to post most pins; although those topics are not female-dominant ones, there may exist female specialists on such topics.

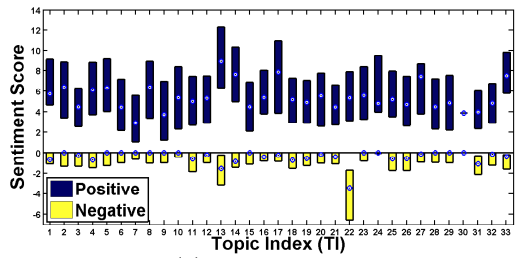
We also examine whether users' pinning efforts are skewed in some topics or evenly distributed across many topics. To this end, we adopt the Shannon's entropy [21], a well-known measure of variety. We calculate the normalized version of entropy for user u as follows:

$$TopicEntropy(u) = -\sum_{i=1}^{T_u} \frac{p_i^u \ln p_i^u}{\ln T_u} \quad (1)$$

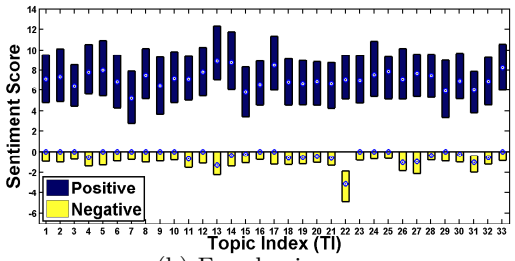
where T_u is the number of topics that user u has, and p_i^u is the relative portion of the pins in the i^{th} topic of u. The topic entropy of a user is one if she has pinned the equal share of content across the topics that she has, and is zero if all of her pins are pinned in a single topic.

The bar plots in Figure 5(a) show the average topic entropy of male and female users for each topic. Note that a circle and a star indicate the median values for male and female users in each topic, respectively. In our dataset, we observe that users are interested in pinning up to five topics. As shown in Figure 5(a), the topics which users are interested in are skewed since all the average and median topic entropy values are below 0.6. Average and median topic entropy values in "education" (TI 3), "health & fitness" (TI 5), "hair & beauty" (TI 17), and "sports" (TI 32) are even lower than 0.2, implying that users pinning on those topics are highly likely to focus on the particular topics. Interestingly, female users show lower entropy values than male users in pinning, which indicates that pinning efforts by female users tend to be skewed in less topics.

We further plot the percentage of dedicated users who post pins only on a single topic in Figure 5(b). For in-

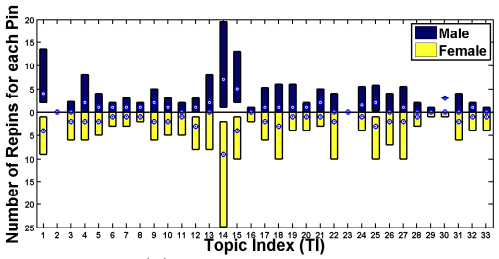


(a) Male pinners

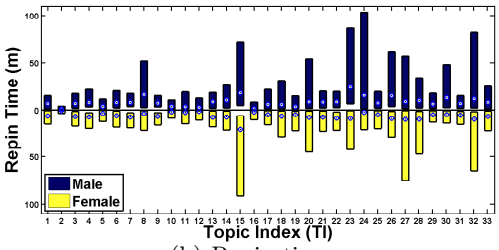


(b) Female pinners

Figure 6: Positive and negative sentiment scores of pinners on each topic are plotted.



(a) Number of repins



(b) Repin time

Figure 7: Repinning behaviors are analyzed in terms of the number of repins (for each pin) and repin time (in minute) on each topic.

stance, 95% of male users in "hair & beauty" (TI 17) post pins only on that topic. Overall, female users tend to have higher portions of dedicated users than male users in pinning, which indicates pinning efforts by female users are likely to be dedicated to a particular topic. The percentages of dedicated users in "education" (TI 3), "health & fitness" (TI 5), and "hair & beauty" (TI 17) are over 90% both for male and female pinners, which signifies a high topic concentration. On the other hand, the percentages of dedicated users in "illustrations & posters" (TI 29) are lower than those in the other topics, meaning that users who are interested in the topic tend to have interests in other topics as well.

We then investigate how strong emotions are exhibited depending on topics. We perform a sentiment analysis by using LIWC (Linguistic Inquiry and Word Count), a transparent text analysis program that counts words into psychologically meaningful categories [17]. For a given text, the LIWC tool provides a positive and a negative emotion scores, each of which is calculated as the relative frequency of the words in the given sentiment category (i.e., positive and negative emotions) on a percentile scale, out of all the words in the text. For example, the words "love" and "sweet" belong to the positive emotion category while "hurt" and "nasty" belong to the negative emotion category.

We collect all the texts a user has written including his/her descriptions and titles/texts/comments for his/her pins/boards, and then calculate the positive and negative sentiment scores for each user using LIWC. Figure 6 shows the distributions of sentiment scores for (a) male and (b) female users, who have been pinned in each topic, respectively. Note that the bottom and top of the box in Figure 6 are the first (25th) and third (75th) quartiles, respectively, and the circle inside the box is the second quartile (the median). Overall, positive emotions are much stronger than negative ones in most cases, revealing that Pinterest users generally exhibit positive emotions, which is in line with the Pollyanna hypothesis that suggests a universal positivity bias in human communications [3]. Also, female users show higher positive scores and lower negative scores than male users on average. The "humor" (TI 13) shows the most positive score while the "geek" (TI 22) shows the most negative score, which is interesting since those two topics are related in the sense that funny things are shared, but the users who are interested in those topics show the opposite emotions. The positive emotions in "humor" (TI 13), "quotes" (TI 14), and "hair & beauty" (TI 17) are stronger than the ones in other topics both for male and female pinners. On the other hand, the negative emotions for male pinners in "weddings" (TI 24) and "kids" (TI 30) are almost zero; male pinners are likely to have negligible negative emotions on family-related topics like weddings or kids. Interestingly, the absolute values of both positive and negative scores in "architecture" (TI 7) are relatively low, meaning that users interested in that topic tend to be calm.

4.2 Repinning: Disseminating the Existing Content

We now turn our attention to users' repinning behaviors on different topics. We first investigate how many pins are repinned and how fast users share pins in each topic. Figures 7(a) and 7(b) plot the distributions of number of repins for each pin and repin times

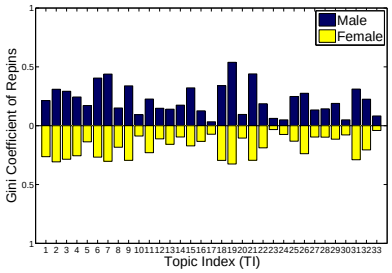
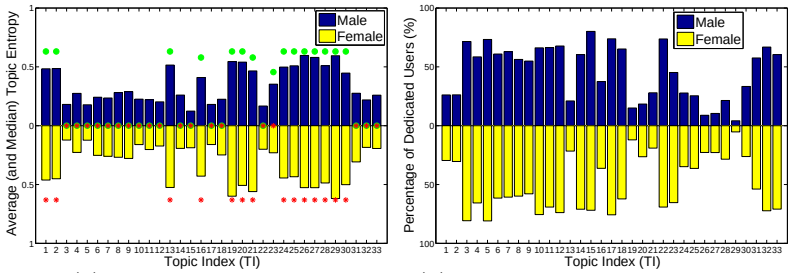
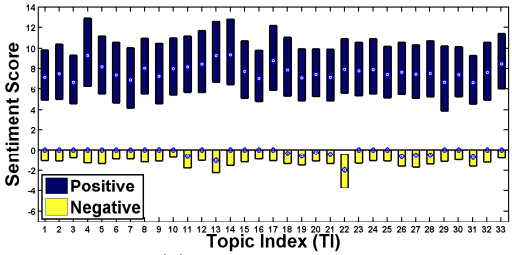


Figure 8: Gini coefficients of repins are plotted on each topic.

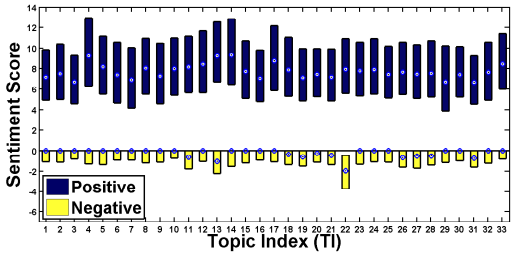


(a) Normalized entropy (b) Percentage of dedicated users

Figure 9: Topic concentration in repinning is explored in terms of topic entropy and percentage of dedicated users for each topic.



(a) Male repinners



(b) Female repinners

Figure 10: Positive and negative sentiment scores of repinners are plotted on each topic.

(in minutes) in each topic, respectively. As shown in Figure 7(a), the number of repins in "quotes" (TI 14) is higher than those of the other topics; popular quotations are usually spread more widely both by male and female repinners. However, some topics show different patterns depending on male and female repinners. For instance, pins in "education" (TI 3), "women's fashion" (TI 12), "science & nature" (TI 18), "geek" (TI 22), "outdoors" (TI 25), or "tattoos" (TI 27) are likely to be repinned many times by female users but a few times by male users, while the number of repins by male users is much higher than that of female users in "diy & crafts" (TI 1) or "animals" (TI 4).

Figure 7(b) shows the distributions of repin times of content across the topics; note that a repin time is an interval between pinning and the 1st repinning or two consecutive repinning moments (of the same pin). The repin time of male users is mostly higher than that of female users, meaning that female users tend to spread content more quickly; note that the numbers of repins by gender are not so different mostly as shown in Figure 7(a). The repin times of "food & drink" (TI 2), "health & fitness" (TI 5), "home decor" (TI 10), "gardening" (TI 16), "illustrations & posters" (TI 29), and "history" (TI 31) are much lower than those of the other topics, which implies that users interested in those topics tend to react quickly or check frequently. Note that content in "food & drink" (TI 2) are shared in four minutes on average. On the other hand, the repin times of "men's fashion" (TI 15), "travel" (TI 20), "shop" (TI 23), "weddings" (TI 24), "tattoos" (TI 27), and "sports" (TI 32) are much higher than those of other topics. Interestingly, a significantly different pattern is observed between male- and female-dominant topics. For example, in the male-dominant topic, "men's fashion" (TI 15), the average repin time of a female user is higher than that of a male user. A similar pattern can be found in the female-dominant topic, "weddings" (TI 24); the repin time of a male user is much higher than that of a female user.

Figure 8 shows the distribution of the numbers of repins among users in each topic by calculating the *Gini coefficient of repins*. The Gini coefficients of repins of female users in most cases are lower than those of male users, which indicates that female users tend to contribute more evenly in repinning. The "hair & beauty" (TI 17), "weddings" (TI 24), and "kids" (TI 30) show the lowest Gini coefficients (< 0.05) of repins for male users while the "outdoors" (TI 23) and "holidays & events" (TI 33) show the lowest Gini coefficients (< 0.05) of repins for female users. Note that the "technology" (TI 19) shows the relatively skewed distributions both for male and female repinners; a small portion of users who are interested in that topic are more actively repinning than the others. From Figures 4 and 8, the Gini coefficients of repins are lower than those of pins in many cases, meaning that repinners tend to more evenly contribute than pinners.

We also examine whether users' efforts on repinning are skewed in some topics or evenly distributed across multiple topics. Figures 9(a) and 9(b) show the average (and median) topic entropy of repinners and percentage of dedicated users who only share (or repin) pins of a single topic, respectively. Note that users are interested in repinning up to twelve topics in our dataset. Pinning the less topics (i.e., up to five) than repinning (i.e., up to twelve) may be due to the fact that information cre-

ation behaviors (i.e., pinning) require more efforts and dedication than information diffusion behaviors (i.e., repinning) [18]. Figures 5(a) and 9(a) reveal that most of topic entropy values in pinning are lower than those in repinning, which indicates that users tend to concentrate on less topics in pinning compared to repinning. Similarly, the percentage of dedicated users in repinning is lower than the one in pinning as shown in Figures 5(b) and 9(b), meaning that pinners tend to focus more on a particular topic than repinners. The topic entropy values in "education" (TI 3) and "health & fitness" (TI 5) are lower than 0.2, implying that users who are interested in those topics are highly likely to focus only on a single topic. On the other hand, users' efforts on repinning in "illustrations & posters" (TI 29) are likely to be evenly distributed across multiple topics.

We finally investigate whether users show distinct emotions in sharing different topics. Figure 10 shows the distributions of positive/negative sentiment scores for (a) male and (b) female users, who have been repinned in each topic, respectively. As shown in Figure 10, like the pinning case in Figure 6, the "geek" (TI 22) shows the most negative score. From Figures 6 and 10, repinners tend to show more positive scores than pinners, implying that users who engage in diffusing content tend to be more amiable.

5. TOPIC NETWORK

In this section, we seek answers for the second question, *Q2 - Topic Relation*, by analyzing the four topic networks defined in Section 3: (i) $TN_{male,pin}$, (ii) $TN_{female,pin}$, (iii) $TN_{male,repin}$, and (iv) $TN_{female,repin}$, respectively.

5.1 How are topics related?

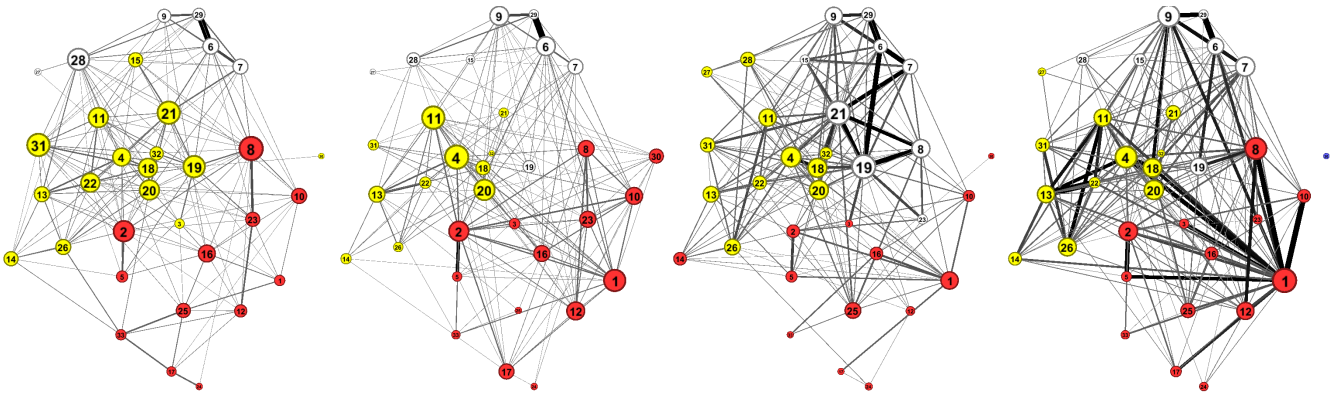
To investigate relations among topics in the TNs, we plot the graph models, whose nodes and edges represent topics and common users in two topics, respectively, in Figure 11. For illustration purposes, the thickness of an edge indicates the weight (defined in the methodology section). A larger circle indicates a node with a higher degree, and the same color of nodes indicates the same community, which will be explained later. As shown in Figure 11, relations among topics are significantly different across the TNs. For example, while "diy & crafts" (TI 1) and "products" (TI 8) are strongly tied together (i.e., having a relation with high weight) in $TN_{female,repin}$, there is no direct link between them in $TN_{male,pin}$. The top 3 related topics (in terms of weight) of "shop" (TI 23) in $TN_{male,repin}$ are "technology" (TI 19), "products" (TI 8), and "food & drink" (TI 2) while the ones in $TN_{female,repin}$ are "products" (TI 8), "diy & crafts" (TI 1), and "design" (TI 6), which might provide important implications for online retailers to develop their marketing strategies. As to the "hair & beauty" (TI 17), which is a female-dominant topic as shown in Figure 3, the related topics of "hair & beauty" (TI 17) by male users are "weddings" (TI 24) and "holidays & events" (TI 33), implying that male users interested in "hair & beauty" may have a particular motivation, e.g., preparing their weddings.

We next examine the degree of each topic in the four TNs. The degree of a topic indicates how many topics have relations with the given topic. If a particular topic has a large degree, the topic may play a role as a *hub* among topics. The average degrees of the nodes in $TN_{male,pin}$, $TN_{female,pin}$, $TN_{male,repin}$, and $TN_{female,repin}$ are 11.03, 10.48, 11.21, and 12.25, respectively. We observe that the degree of each topic is significantly different across the TNs. For example, "celebrities" (TI 26) is connected to many other topics in $TN_{female,repin}$, but it is connected to much less topics (i.e., small degree) in $TN_{female,pin}$. The topics of "products" (TI 8), "cars & motorcycles" (TI 21), and "history" (TI 31) in $TN_{male,pin}$, "animals" (TI 4) and "film, music & books" (TI 11) in $TN_{female,pin}$, "technology" (TI 19) and "cars & motorcycles" (TI 21) in $TN_{male,repin}$, and "animals" (TI 4), "products" (TI 8), and "art" (TI 9) in $TN_{female,repin}$ show the highest degrees, meaning that those topics have relations with many other topics and play roles as hubs, respectively. The degrees of "animals" (TI 4), "film, music & books" (TI 11), and "travel" (TI 20) are substantially high across the four TNs, which implies that those topics generally contain common interests with other topics. On the other hand, the degrees of "weddings" (TI 24), "tattoos" (TI 27), "kids" (TI 30), and "holidays & events" (TI 33) are generally lower than those of others, which means users interested in those topics tend to focus on the topics. From business perspectives, it may be more effective to identify the users who mostly focus on a particular set of topics for resource allocation in targeted marketing.

5.2 How are topics clustered into groups?

We now examine how topics in the TN are clustered into groups (or communities). Here, a community is a group of topics, within which edges are denser, but between which edges are sparser. We identify communities of the four TNs using the Louvain method [2], a well-known fast community detection algorithm that maximizes the ratio of the number of edges within communities to that of edges between communities. We use the weighted version of Louvain method. Recall that the same color of nodes in Figure 11 indicates the same community.

Table 2 lists the topics in each of the identified communities in the four TNs. In $TN_{male,pin}$, there are three communities, and the member topics in the second community are related to fine arts or design. $TN_{male,repin}$ also has three communities but their members are some-



(a) $TN_{male,pin}$ (b) $TN_{female,pin}$ (c) $TN_{male,repin}$ (d) $TN_{female,repin}$

Figure 11: Graph models of four TNs are illustrated. The thickness of an edge indicates the weight (defined in Section 3). A larger circle indicates a node with a higher degree, and the same color of nodes indicates the same community.

TN	C	Member Topics
$TN_{m,p}$	1	education(3), animals(4), film,music&books(11), humor(13), quotes(14), men's fashion(15), science&nature(18), technology(19), travel(20), cars&motorcycles(21), geek(22), celebrities(26), kids(30), history(31), sports(32)
	2	design(6), architecture(7), art(9), tattoos(27), photography(28), illustrations&posters(29)
	3	diy&crafts(1), food&drink(2), health&fitness(5), products(8), home decor(10), women's fashion(12), gardening(16), hair&beauty(17), shop(23), weddings(24), outdoors(25), holidays&events(33)
$TN_{f,p}$	1	animals(4), film,music&books(11), humor(13), quotes(14), science&nature(18), travel(20), cars&motorcycles(21), geek(22), celebrities(26), history(31), sports(32)
	2	design(6), architecture(7), art(9), men's fashion(15), technology(19), tattoos(27), photography(28), illustrations&posters(29)
	3	diy&crafts(1), food&drink(2), education(3), health&fitness(5), products(8), home decor(10), women's fashion(12), gardening(16), hair&beauty(17), shop(23), weddings(24), outdoors(25), kids(30), holidays&events(33)
$TN_{m,r}$	1	animals(4), film,music&books(11), humor(13), science&nature(18), travel(20), geek(22), celebrities(26), tattoos(27), photography(28), history(31), sports(32)
	2	design(6), architecture(7), products(8), art(9), men's fashion(15), technology(19), cars&motorcycles(21), shop(23), illustrations&posters(29)
	3	diy&crafts(1), food&drink(2), education(3), health&fitness(5), home decor(10), women's fashion(12), quotes(14), gardening(16), hair&beauty(17), weddings(24), outdoors(25), kids(30), holidays&events(33)
$TN_{f,r}$	1	animals(4), film,music&books(11), humor(13), quotes(14), science&nature(18), travel(20), cars&motorcycles(21), geek(22), celebrities(26), tattoos(27), history(31), sports(32)
	2	design(6), architecture(7), art(9), men's fashion(15), technology(19), photography(28), illustrations&posters(29)
	3	diy&crafts(1), food&drink(2), education(3), health&fitness(5), products(8), home decor(10), women's fashion(12), gardening(16), hair&beauty(17), shop(23), weddings(24), outdoors(25), holidays&events(33)
	4	kids(30)

Table 2: The topics in each community (C) are identified in each TN: $TN_{male,pin}$, $TN_{female,pin}$, $TN_{male,repin}$, and $TN_{female,repin}$.

what different from $TN_{male,pin}$. For example, in addition to topics related to fine arts or design, "men's fashion" (TI 15) and "technology" (TI 19) are also members in the second community of $TN_{male,repin}$, which implies that "men's fashion" (TI 15) and "technology" (TI 19) are somewhat linked to the topics related to fine arts or design for male repinners. While "tattoos" (TI 27) is the member of the second community in both $TN_{male,pin}$ and $TN_{female,pin}$, it belongs to the first community in $TN_{male,repin}$ and $TN_{female,repin}$, respectively. This indicates that "tattoos" (TI 27) is connected to design-related topics in pinning, but is more closely linked to light-hearted topics such as "humor" (TI 13) or "geek" (TI 22) in repinning. In $TN_{female,repin}$, there are four communities. The "kids" topic (TI 30) is the only member in the fourth community in $TN_{female,repin}$, meaning that female repinners are interested in the "kids" solely.

Overall, some particular topics generally form the same community; for instance, "diy & crafts" (TI 1), "food & drink" (TI 2), "health & fitness" (TI 5), "women's fashion" (TI 12), "gardening" (TI 16), and "hair & beauty" (TI 17) belong to the third community across the TNs. This implies that (users of) those topics share common interests regardless of genders or pinning/repinning behaviors. On the other hand, some topics (e.g., "shop" (TI 23) or "kids" (TI 30)) belong to the different communities depending on the TNs, which may give valuable implications for online retainers to develop targeted-advertisement or cross-selling services.

We believe our analysis on the four TNs can be used in identifying hidden (but important) links among the topics (or interests). One of well known examples of the hidden links is the association between beer and diapers [1]. The identification of such links can be applied to promotion strategies such as cross-product advertisement, bundling, or product-pairing. For example, in $TN_{male,pinner}$, we observe that there are close relations among "technology", "product", "sports", and "men's fashion". If products related to these topics are

displayed together (in a department store or an online store), *relevant* products are exposed together to consumers, which might help to increase sales.

6. PREDICTING WHICH TOPIC A USER WILL BE INTERESTED IN

In this section, we strive to answer the third question, *Q3 - Application*: can we predict which topic a user will be interested in? To answer the question, we propose and evaluate the following prediction methods:

- $Popularity-based$ selects the most popular topic, among the topics which a user has not been interested in. This method is suggested for comparison purposes.
- $CF-based$ adopts the *(item-to-item) collaborative filtering (CF)* technique [20], a well-known recommendation algorithm, whose basic idea is to find the most similar topic that other users tend to consume together. For this, we define a topic vector whose elements are users who pin/repin the pins of the topic; the dimension of the vector is the number of entire users. The similarity between two topics is calculated by the cosine similarity of two topic vectors. If the selected topic has already been consumed by the target user, we select the next most similar topic.
- $TN-neighbor-based$ uses the relation information of topics in the TN which the target user belongs to (based on her gender information and pinning/repinning preferences). This method first finds topic A, which the target user has been most interested in, and selects the topic that has the strongest relation (i.e., the highest weight in the TN) with A. If the selected topic has already been consumed by the target user, we choose the topic which has the next strongest relation with A.
- $TN-community-based$ further utilizes the community information of the TN which a user belongs to (based on her gender information and pinning/repinning preferences). We first find a corresponding community that contains the largest number of the topics (which the target user has been interested in), and select a topic (which the target user has not been interested in) from the same community. If there are multiple candidate topics in the community, we choose the topic that has the strongest relation with the one which the target user has been most interested in. The basic idea of this method is to find a similar topic based on the collective opinions of other like-minded users.

To evaluate the proposed prediction methods, we first select 1,913 target users satisfying two criteria: (i) her gender information is available, and (ii) she has at least 10 pins. Based on the proposed methods, as of Jul. 18th, 2013, we choose a candidate topic that may be consumed by each target user in the future. To validate whether the predicted topic is actually shared by each target user after 125 days, we collected another dataset (for 10 days, from Nov. 20th to 30th, 2013) that contains the target users' pins and their corresponding topic information. For the purpose of evaluation, we measure the hit ratio of each method, by calculating the ratio of the number of users who actually consume the predicted topic to the total number of target users.

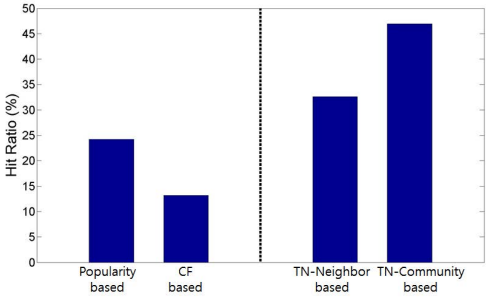


Figure 12: Hit ratios of the proposed prediction methods are plotted; $TN-community-based$ prediction performs the best.

Figure 12 shows the hit ratios of the four proposed prediction methods: $popularity-based$, $CF-based$, $TN-neighbor-based$, and $TN-community-based$. As shown in Figure 12, the prediction methods utilizing the TN information ($TN-neighbor-based$ and $TN-community-based$) perform better than $popularity-based$ and $CF-based$, which indicates the relations among topics is useful in predicting a user's interested topic in the future. $TN-community-based$ outperforms the others (i.e., the hit ratio is close to 50%), implying that the community information (of the TN) that reflects the collective opinions of other like-minded users is an important predictor in predicting which topic a user will be interested in. We believe this can give important implications on topic demand forecasting or cross-topic advertisement in Pinterest-like social curation services.

7. CONCLUSION

This paper analyzed (1) the differences in pinning/repinning behaviors by topics and user gender, and (2) the relations among topics in Pinterest. We summarize three main contributions as follows. First, we investigated how different topics are shared from the perspectives of (i) pinning/repinning behaviors and (ii) gender differences in such behaviors, which shows significantly different patterns in terms of dedication, responsiveness, and sentiment. Second, by introducing the notion of topic networks, we analyzed (i) how topics are related to one another, and (ii) what topics are clustered into the same community, which can provide a valuable implication on topic demand forecasting or cross-topic ad-

vertisement. Lastly, we explored the implications of our findings for predicting which topics a user will be interested in later. We demonstrated that the notion of topic networks (that reflect the collective opinions of other like-minded users) is useful in accurately predicting a user's interest and behavioral pattern in Pinterest.

8. REFERENCES

[1] M. J. Berry and G. S. Linoff. *Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons, Inc., 1997.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), 2008.

[3] J. Boucher and C. E. Osgood. The pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behavior*, 8(1):1–8, 1969.

[4] S. Chang, V. Kumar, E. Gilbert, and L. G. Terveen. Specialization, homophily, and gender in a social curation site: Findings from pinterest. In *ACM CSCW*, 2014.

[5] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: Experiments on recommending content from information streams. In *ACM CHI*, 2010.

[6] J. Constine. Pinterest hits 10 million u.s. monthly uniques faster than any standalone site ever -comscore. `http://goo.gl/EZFftf`, 2012.

[7] C. Dagum. The generation and distribution of income, the Lorenz curve and the Gini ratio. *Economie Appliquée*, 33(2), 1980.

[8] B. Gelley and A. John. Do i need to follow you?: Examining the utility of the pinterest follow mechanism. In *ACM CSCW*, 2015.

[9] E. Gilbert, S. Bakhshi, S. Chang, and L. Terveen. "i need to try this!": A statistical overview of pinterest. In *ACM CHI*, 2013.

[10] J. Han, D. Choi, B.-G. Chun, T. T. Kwon, H.-c. Kim, and Y. Choi. Collecting, organizing, and sharing pins in pinterest: Interest-driven or social-driven? In *ACM SIGMETRICS*, 2014.

[11] E. A. Harris. Retailers seek partners in social networks. `http://goo.gl/gzaKng`, 2013.

[12] K. Y. Kamath, A.-M. Popescu, and J. Caverlee. Board recommendation in pinterest. In *Conference on User Modeling, Adaptation and Personalization*, 2013.

[13] R. Linder, C. Snodgrass, and A. Kerne. Everyday ideation: All of my ideas are on pinterest. In *ACM CHI*, 2014.

[14] R. Ottoni, D. L. Casas, J. P. Pesce, W. Meira Jr., C. Wilson, A. Mislove, and V. Almeida. Of Pins and Tweets: Investigating how users behave across image- and text-based social networks. In *ICWSM*, 2014.

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

[16] C. Palis. Pinterest traffic growth soars to new heights: Experian report. `http://goo.gl/yMJCiG`, 2012.

[17] J. W. Pennebaker, M. R. Mehl, and K. Niederhoffer. Psychological Aspects of Natural Language Use: Our Words, Ourselves. *Annual Review of Psychology*, 54:547–577, 2003.

[18] J.-Y. Rha. Consumers' usage of online social networks: application of use-diffusion model. *Journal of Consumer Studies*, 21(2):443–470, 2010.

[19] S. Rosenbaum. *Curation Nation: How to Win in a World Where Consumers are Creators*. McGraw-Hill, 2011.

[20] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *ACM WWW*, 2001.

[21] C. E. Shannon. Prediction and entropy of printed english. *Bell System Technical Journal*, 30(1):50–64, 1951.

[22] M. Zarro, C. Hall, and A. Forte. Wedding dresses and wanted criminals: Pinterest.com as an infrastructure for repository building. In *ICWSM*, 2013.

[23] C. Zhong, D. Karamshuk, and N. Sastry. Predicting pinterest: Automating a distributed human computation. In *ACM WWW*, 2015.

[24] C. Zhong, M. Salehi, S. Shah, M. Cobzarenco, N. Sastry, and M. Cha. Social bootstrapping: How pinterest and last.fm social communities benefit by borrowing links from facebook. In *ACM WWW*, 2014.

[25] C. Zhong, S. Shah, K. Sundaravadivelan, and N. Sastry. Sharing the loves: Understanding the how and why of online content curation. In *ICWSM*, 2013.

[26] T. Zhou, J. Ren, M. Medo, and Y. C. Zhang. Bipartite network projection and personal recommendation. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 76(4):046115+, 2007.