Collecting, Organizing, and Sharing Pins in Pinterest: Interest-driven or Social-driven?

Jinyoung Han[*]
Seoul National University
Seoul, Korea
jyhan@mmlab.snu.ac.kr

Daejin Choi
Seoul National University
Seoul, Korea
djchoi@mmlab.snu.ac.kr

Byung-Gon Chun
Seoul National University
Seoul, Korea
bgchun@snu.ac.kr

Ted “Taekyoung” Kwon
Seoul National University
Seoul, Korea
tkkwon@snu.ac.kr

Hyun-chul Kim
Sangmyung University
Cheonan, Korea
hkim@smu.ac.kr

Yanghee Choi
Seoul National University
Seoul, Korea
yhchoi@snu.ac.kr

ABSTRACT

Pinterest, a popular *social curating service* where people collect, organize, and share content (pins in Pinterest), has gained great attention in recent years. Despite the increasing interest in Pinterest, little research has paid attention to how people collect, manage, and share pins in Pinterest. In this paper, to shed insight on such issues, we study the following questions. How do people collect and manage pins by their tastes in Pinterest? What factors do mainly drive people to share their pins in Pinterest? How do the characteristics of users (e.g., gender, popularity, country) or properties of pins (e.g., category, topic) play roles in propagating pins in Pinterest? To answer these questions, we have conducted a measurement study on patterns of pin curating and sharing in Pinterest. By keeping track of all the newly posted and shared pins in each category (e.g., animal, kids, women’s fashion) from June 5 to July 18, 2013, we built 350 K pin propagation trees for 3 M users. With the dataset, we investigate: (1) how users collect and curate pins, (2) how users share their pins and why, and (3) how users are related by shared pins of interest. Our key finding is that pin propagation in Pinterest is mostly driven by pin’s properties like its topic, not by user’s characteristics like her number of followers. We further show that users in the same community in the interest graph (i.e., representing the relations among users) of Pinterest share pins (i) in the same category with 94% probability and (ii) of the same URL where pins come from with 89% probability. Finally, we explore the implications of our findings for predicting how pins are shared in Pinterest.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Web-based services; J.4 [**Computer Applications**]: Social and behavioral sciences

General Terms

Human Factors, Measurement

Keywords

Pinterest; Online Social Network; Social Curating; Content Propagation; Repin

1. INTRODUCTION

Pinterest[1] provides a social curating service where people can collect, organize, and share content[2]. Pew Research Center reported that 15% of online adults use Pinterest as of Dec. 2012 [3]. Pinterest has also been marked as the fastest growing web site to reach 10 million unique visitors [5]. According to the marketing service Experian, Pinterest has become the third most popular social network in the United States (as of Mar. 2012), behind Facebook and Twitter [6]. Currently, Pinterest is the 26th and 12th most popular web site in the world and Unite States (as of Nov. 2013), respectively [1].

The huge upsurge and popularity of Pinterest are attributed to its unique and attractive properties. First, Pinterest allows users to collect, organize, and share content (i.e., pins in Pinterest) by their tastes or interests [2, 4, 7, 15]. Second, about 80% of Pinterest’s users are female [2, 7, 15, 25], which contrasts with male-centric services like *Quora.com* that mostly relies on early adopters of technology.

The popularity of Pinterest has attracted the research community to investigate user behaviors. Recent studies have revealed valuable insights into Pinterest internals [8, 15, 16, 19, 25, 32, 33]. However, most of these studies paid little attention to how pins are collected, organized, and propagated over Pinterest, which is the key to understanding Pinterest-like social networks.

[*] Jinyoung Han is currently a post-doctoral researcher at University of California-Davis.

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

[2] Pinterest serves two types of content: image and video. Since image content is dominant in Pinterest, we focus on analyzing image content in this paper.

SIGMETRICS’14, June 16–20, 2014, Austin, Texas, USA.
Copyright 2014 ACM 978-1-4503-2789-3/14/06 ...$15.00.
http://dx.doi.org/10.1145/2591971.2591996.

In this paper, we seek to demystify Pinterest by answering the following questions. How do people collect and manage pins by their tastes? How many interests do people usually have? How do pins propagate in Pinterest? What factors drive people to share their pins in Pinterest? How do the characteristics of users (e.g., gender, popularity, country) or properties of pins (e.g., topic, content) play roles in propagating pins in Pinterest? Are there any differences in the way of sharing pins between Pinterest and other social networks?

We answer the questions from the perspective of pin curating and sharing with the dataset (350 K pins and 3 M users) we collected by crawling pages from Pinterest. That is, we investigate pin (or content) propagation patterns in Pinterest by keeping track of a propagation path of each pin. For example, if user A shares user B's pin, we can learn that B's pin is shared by A, which indicates the pin is propagated (or *repinned*) from B to A. Note that a pin can be propagated over multiple hops (e.g., from user A to user B to user C) by *repinning* in Pinterest. By keeping track of all the newly-posted and shared pins in each category (e.g., animal, kids, women's fashion) of Pinterest from June 5 to July 18, 2013, we construct pin propagation trees, each of which is built for a pin. With the dataset, we have analyzed: (1) how users collect and curate pins, (2) how users share their pins and why, and (3) how users are related to one another by shared interests.

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

- **Measurement:** To our knowledge, this is the first measurement study to comprehensively investigate how people collect, manage, and share pins in Pinterest. We make our anonymized dataset online at: `http://mmlab.snu.ac.kr/traces/pinterest`.

- **Key Findings:** We reveal that pin propagation in Pinterest is mostly driven by its properties like its topic or content, not by user's characteristics like her number of followers. We also find that users in the same community in the interest graph (i.e., representing the relations among users) of Pinterest share pins (i) in the same category with 94% probability and (ii) of the same URL where pins come from with 89% probability, which implies that Pinterest users are highly related if they share interests.

- **Implications - Predictions on Pin Consumptions:** We explore the implications of our findings for predicting how pins are shared in Pinterest. We demonstrate that our proposed predicting method considering properties of pins exhibits higher precision (about 4.5 times) than other methods. Our trace-driven study for predicting pin consumption patterns in Pinterest suggests that properties of pins are much more important factors than those of users.

The rest of this paper is organized as follows. We introduce Pinterest and review related work in Section 2. We present the measurement methodology in Section 3. We start our analysis by investigating how users collect and organize their pins in Section 4. In Section 5, we analyze how and why pins propagate in Pinterest. We then analyze how users are related to one another by their shared interests in Section 6. Finally, we explore how to apply our findings to methods for predicting how pins are shared among users in Section 7.

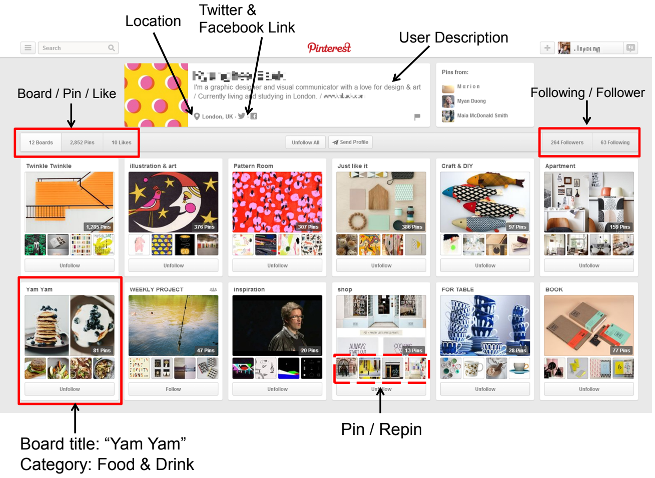


Figure 1: An illustration of a user's profile page in Pinterest.

2. BACKGROUND

2.1 Pinterest Overview

Pinterest [2, 7, 15, 25] is a pinboard-style content sharing platform that allows users to exhibit collections of images or videos. The main idea of Pinterest is to collect, organize, and share content (mostly images since image content is dominant in Pinterest) that users find interesting; Pinterest focuses on collecting and sharing content (i.e., pins in Pinterest). That is, Pinterest's basic function is to let users collect, organize, and share pins by their tastes or interests. Direct communications (e.g., private messages in Facebook or Twitter) between users are not available in Pinterest. Instead, user interactions mostly occur at the time they write feedbacks on or share pins (e.g., a user can *like* or *comment* on someone's pin). Figure 1 illustrates a user's profile page in Pinterest. We describe key terminologies in Pinterest below.

- Pin/Repin: Each image/video is called as a pin, and the act of posting a pin is referred to as pinning. If a posted pin is shared by someone, the shared pin is called as a repin, which is similar to *retweet* in Twitter. The act of sharing other user's pin is called repinning. Users who posts and shares (i.e., repins) a pin are the (original) pinner and repinner, respectively.

- Source: A user can directly upload a pin or fetch a pin from other websites like *Tumblr.com*. In the latter case, the URL of the pin is referred to as a source.

- Like/Comment: Similar to Facebook, a user can push a like button for a pin that she likes and leave a comment on a pin.

- Board/Category: A board is a collection of pins organized by a user. Each board belongs to one of the categories in Pinterest. At the moment there are 32 categories in Pinterest, varying from "animals" to "history" to "women's fashion".

- Following/Follower: Like Twitter, the relation between two users in Pinterest is not symmetric. The fact that user A follows user B does not necessarily mean B follows A. If A follows B, A can see the updated news (e.g., the act of posting a new pin) of B.

2.2 Related Work

Pinterest: Despite its young age (only 3 years old), Pinterest has attracted much attention since its launching. This in turn has spurred research into its gender differences [15, 25], repository perspectives [16,32], and applications [19,33]. Ottoni *et al.* showed that females are more active, make more use of lightweight interactions, and invest more effort in reciprocating social links than males in Pinterest [25]. Gilbert *et al.* found that being females means doing more repins but having fewer followers [15]. Some studies considered Pinterest as a repository like a digital library [16,32]. Zarro *et al.* found that Pinterest serves as an infrastructure for a repository that supports the following activities: discovery, collecting, collaborating, and publishing [16]. Zarro and Hall further discussed about how digital libraries could take advantage of Pinterest, allowing users to create personalized collections incorporating their content [32]. As an application on Pinterest, Kamath *et al.* introduced a supervised model for board recommendation [19]. They found that using social signals (e.g., repins, likes) can achieve higher recommendation quality. Zoghbi *et al.* suggested and evaluated information retrieval models for linking the texts of pins to web pages in Amazon [33]. While these studies mostly focus on user behaviors in Pinterest, we focus on pin propagation patterns with empirically-grounded evidences.

Information/Content Propagation in Online Social Networks: As online social networks are becoming the norm to spread information of interest, there have been studies to investigate patterns of information propagation in online social networks such as Twitter [11,12,20,26,27,31], Flickr [13], Facebook [9], and Digg [21]. Kwak *et al.* observed that retweets in Twitter reach a large audience and spread fast. Rodrigues *et al.* analyzed the word-of-mouth exchange of URLs among Twitter users and showed that users who are geographically close together are more likely to share the same URL [27]. Park *et al.* investigated how bad news about a company spreads in Twitter [26]. Ye and Wu analyzed the propagation patterns of messages in Twitter and showed how a breaking news (i.e., Michael Jackson's death) spreads in Twitter [31].

Some of the studies have focused on what drives information propagation in Twitter. Cha *et al.* analyzed how users play different roles in spreading popular and unpopular news in Twitter [11]. Cha *et al.* also showed that the most influential users (in terms of spawning retweets) are not necessarily the most followed ones [12]. Wang *et al.* found that information spreading is dependent on social context and the individuals' characteristics [30].

Flickr, Facebook, and Digg have also been analyzed as media to spread content or news. Cha *et al.* analyzed image-content propagation patterns in the Flickr social network; they showed that even popular photos do not spread widely and quickly in Flickr [13]. Bakshy *et al.* examined the patterns of information diffusion in Facebook, and found that weak ties play a more important role in dissemination of information in Facebook [9]. Lermn and Ghosh *et al.* tracked how interest in news stories spread on Digg, a popular social news service, and showed that social networks play a crucial role in the spread of interests [21].

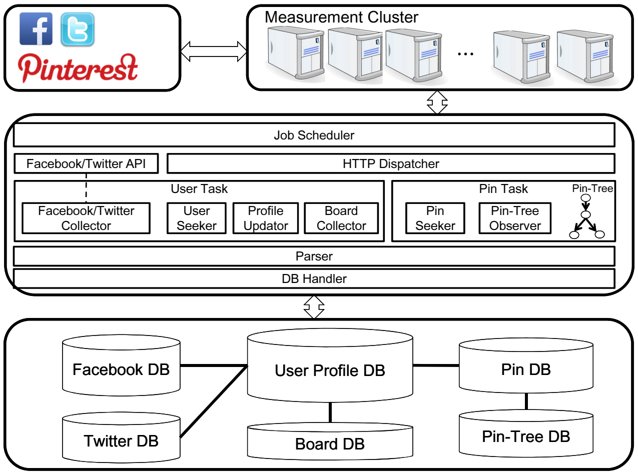


Figure 2: The architecture of our Pinterest crawling and analysis system.

While these studies have revealed valuable insights into the information propagation in Twitter, Flickr, Facebook, and Digg, there has been little research on how content is collected, organized, and propagated in Pinterest, which is a unique social curating service. We believe Pinterest is an ideal platform to investigate content propagation patterns since we can easily keep track of the propagation path of each pin in Pinterest. To our knowledge, this is the first comprehensive measurement study on how people collect, manage, and share pins in a social curating service.

3. METHODOLOGY

In this section, we illustrate our measurement methodology for data collection, and describe the dataset that is used in this paper.

3.1 Data Collection

Since Pinterest does not provide an official API for data collection, we developed our measurement system by crawling Pinterest pages as shown in Figure 2. We fetch web pages in Pinterest, from which the relevant information is extracted; the data about each pin or board can be extracted from a web page. This is challenging since we need to crawl a large number of web pages from Pinterest. For example, if a user has 1,000 boards, we need make 1,000 HTTP requests to collect the data about her board. To address this problem, we designed a distributed crawling system. Our measurement cluster consists of 25 PCs, which continuously send HTTP requests assigned by the job scheduler. The HTTP dispatcher processes the HTTP requests and responses according to the tasks explained below.

There are two main tasks in our system: *pin task* and *user task*. Unlike prior measurement studies (e.g., [15, 25]), we focus on pin propagation patterns. To this end, we periodically (every five minutes) monitor all the newly-posted pins in the menu of each category (e.g., animal, kids, women's fashion). Since Pinterest shows all the recent activities including posting a pin, repinning, and leaving a comment in the menu of each category in the chronological order, our pin seeker fetches 10 recent web pages not to miss newly-posted pins. The pin-tree observer keeps track of each pin and its associated repins to build a pin propagation tree, which is called a *pin-tree*. If a user repins the original pin, Pinterest provides a link to the board that includes the repinned one; we can find and fetch the associated web page of the repinned one among other pin pages in the board, so that we can keep track of the chain of the pin-tree. The collected information of each pin-tree are stored in the pin-tree database. The pin (and repin) dataset consists of the number of likes, number of comments, its category, its source, and its description, which is stored in the pin database.

In the user task, we collect the information (e.g., number of pins, number of followers, number of boards, gender, country, etc.) of each user. In addition to the 1 M users found in pin-trees, the user seeker additionally finds 2 M users using a breath first search (BFS) in Pinterest. For the discovered 3 M users, we collect the information of each user, including her name, her description, gender, number of followers, number of followings, number of boards, number of pins, number of likes, her external website, location, and Facebook/Twitter links, which are stored in the user profile database. Along with the user profiles, the board collector collects the information of each board including its category, and number of pins, which are stored in the board database. To identify the gender and country of users, we use external links to Facebook and Twitter, which can be found in the profile pages of users. The Facebook/Twitter collector sends queries to Facebook and Twitter through their APIs and fetches the gender and country information of each user if available. We finally decide the gender and country of each user by collectively combining information from Pinterest, Facebook, and Twitter.

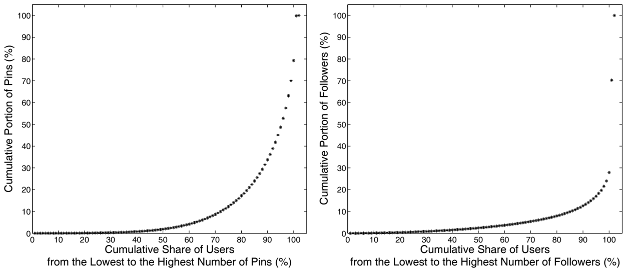
3.2 Dataset

Our dataset had been collected from June 5 to July 18, 2013. We kept track of 346,329 pin-trees, which contain 346,329 (original) pins and their 1,215,045 repins, which are shared by 1,561,374 users. In addition to the users found in pin-tress, we further discovered 1,412,754 users using a breath first search (BFS) through Pinterest. Finally, our dataset includes 2,974,128 users (i.e., 1,561,374 users found in pin-trees + 1,412,754 users discovered through BFS). The dataset collectively contains 40,800,940 boards, 3,362,100,884 pins, 656,123,740 followers, 302,363,300 followings, 1,392,394 Facebook links, and 183,900 Twitter links. We also obtained the country and gender information of 1,354,132 and 1,392,394 users, respectively. We found that 85% of Pinterest users are female in our dataset, which is congruent with prior reports (e.g., [2, 7, 15, 25]). Top five countries in terms of the number of users in our dataset are United States (85%), United Kingdom (6%), Brazil (1%), Netherlands (0.6%), and Spain (0.6%). The ratios of female to male users in these five countries are 87:13, 80:20, 68:32, 75:25, and 67:33, respectively.

4. SOCIAL CURATING

In this section, we investigate social curating behaviors in terms of the number of pins, boards, categories, and followings/followers. Here, social curating indicates the act of collecting and organizing content by tastes or interests, with social functionalities such as liking, commenting, and following [15, 16, 25].

4.1 Pin and Follower Distributions



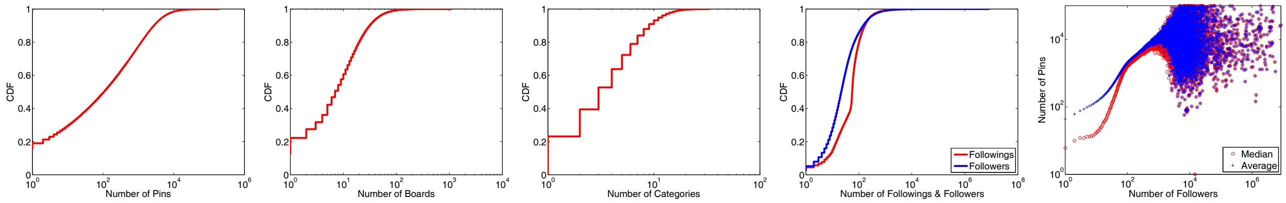
(a) Pin distribution (b) Follower distribution

Figure 3: Gini Triangles of (a) pins and (b) followers. A small portion of Pinterest users own a large portion of pins and followers.

We first analyze the distribution of pins and followers in Pinterest by calculating the *Gini coefficient*, a well-known estimator to evaluate the disparity of a distribution in Economics [14]. The Gini coefficient is always within the range of [0, 1], where 0 and 1 indicate a perfect uniform distribution and an extremely skewed distribution, respectively [14]. Figure 3(a) shows the Lorenz curve [23] of the pins where the x-axis is the cumulative share of users from the lowest to the highest number of pins, while the y-axis is the cumulative portion of the number of pins. Figure 3(b) is the Lorenz curve of the followers. Obviously, a small portion of users own a large portion of pins and followers. For example, top 20% of users in terms of the number of pins own 82% pins of Pinterest and top 20% of users in terms of number of followers have 90% of followers. Note that the Gini coefficients of pins and followers are 0.78 and 0.90, respectively, which exhibits high skewness. Interestingly, only 8% of top 1% users in terms of the number of pins are males, while 18% of top 1% users in terms of the number of followers are males. This implies that males are more interested in social networking than pinning in Pinterest.

4.2 Curating behavior

We next investigate the curating behaviors of users in terms of the numbers of pins, boards, categories, and followings/followers in Figure 4. As shown in Figure 4(a), 45% of users have fewer than 100 pins while top 20% of users have more than 1000 pins; the average, median, and maximum numbers of pins are 1130, 106, and 194,515, respectively. 32 categories are Pinterest-defined topics while a board can be interpreted as a user-defined topic. For example, if a user has a basketball board and a baseball board, both of which may belong to *sports* category, but she may have two topics personally. Figure 4(b) shows around 55% of users have fewer than 10 boards, while top 1% of users have more than 100 boards. The average and median numbers of boards are 14 and 7, respectively. Figure 4(c) shows the number of categories on which the user has posted pins. While 23% of users have only one category, top 10% of users are interested in more than 10 categories. The average and median numbers of categories are 4 and 3, respectively, meaning that an average user of Pinterest manages a few Pinterest-defined topics. We next examine the number of followings and followers in Figure 4(d). Interestingly, most of users have more followings than followers, but a small portion of users have a



(a) Pins (b) Boards (c) Categories (d) Social Ties (e) Pin vs. Followers

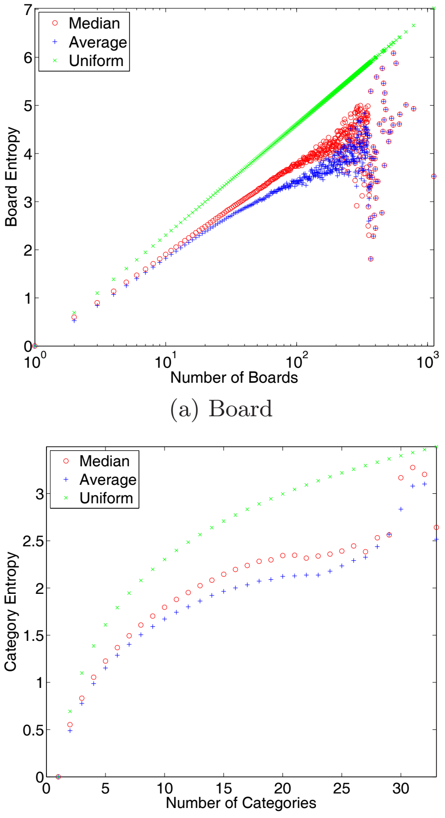
Figure 4: Distributions of the number of (a) pins, (b) boards, (c) categories, and (d) followings/followers. The numbers of followers and pins for each user are further illustrated in (e) to investigate their correlation.

very large number of followers, hence the average number of followers (221) is about two times greater than that of followings (102). As to the number of followers, 25% of users have fewer than 10 followers while top 15% of users have more than 100 followers. Note that the maximum number of followers is 8,430,910. Finally, to investigate the correlation between the number of followers and that of pins for each user, we plot the number of followers (x-axis) against the number of posted pins (y-axis) in Figure 4(e). The average is mostly above the median until the number of followers reaches 1,000, indicating that there are outliers who pin far more than ordinary users with the same number of followers. There is a positive correlation between the number of followers and that of pins up to $x = 1{,}000$. However, there is a weak correlation beyond $x = 1{,}000$, which signifies that a user who has a large number of followers does not necessarily post a large number of pins.

We next investigate whether and how user's efforts on pinning are evenly distributed across different boards by calculating the Shannon's entropy [29] defined by:

$$H_{board} = -\sum_{i=1}^{B} p_i \ln p_i \qquad (1)$$

where B is the number of boards and p_i is the relative portion of the pins in the i^{th} board of a user. We can easily calculate $H_{category}$ similarly. Figure 5 shows the correlations between the number of interested topics (boards and categories) and its corresponding entropy. Users' interested topics are skewed since all the median and average values are notably below the values of the uniform case in which a user posts pins/repins evenly across different boards or categories. As the numbers of boards and categories increase, the gaps between the values of the uniform cases and the average values becomes wider until the numbers of boards and categories are 200 and 27, respectively, which implies that users who have pins in more boards or categories focus on a small number of interests. However, $H_{category}$ values are increasing somewhat sharply after the number of categories is 28, which needs further investigation.



(a) Board

(b) Category

Figure 5: The correlation between the number of interested topics (boards and categories) and its corresponding entropy.

5. PIN TREE ANALYSIS

In this section, we first seek to model how pins propagate in Pinterest. We then characterize the patterns of pin propagation from a graph-theoretical perspective, and investigate what factors affect pin propagation in Pinterest.

5.1 Definition of Pin-Tree

We define a pin-tree as a directed graph, $T = (V, E)$, where V is the set of users and E is the set of repins (i.e., pin propagations), based on the Krackhardt's hierarchical tree model [17]. That is, if user j repins a pin from user i, there exists an edge $E(V_i, V_j)$ from user V_i to user V_j. Note that all nodes in a hierarchical tree (except for the root) have the same ancestor (which is the root). Figure 6 shows an example of a pin-tree. A pinner is a user who posts an original pin and a repinner is a user who fetches the pin from her parent. We define the max depth in a pin-tree as the maximum hop count from the root to any of the leaf nodes. The max width refers to the maximum number of child nodes of a parent node in a given tree. The max depth and max width of a pin-tree in Figure 6 are 2 and 4, respectively.

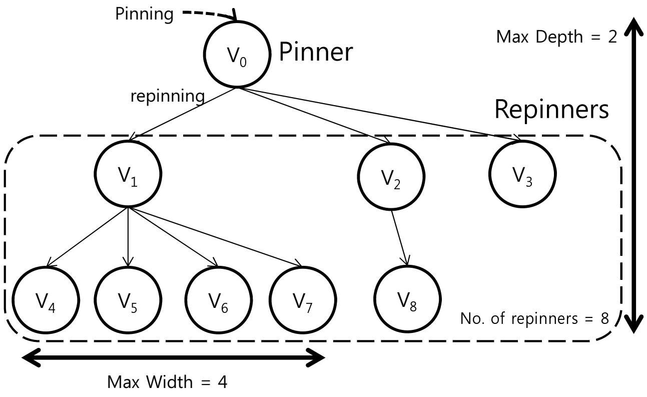


Figure 6: An example of a pin-tree.

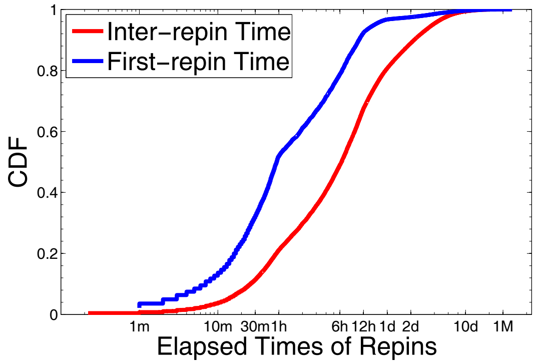
5.2 Structure of Pin-Tree

5.2.1 *Structural analysis*

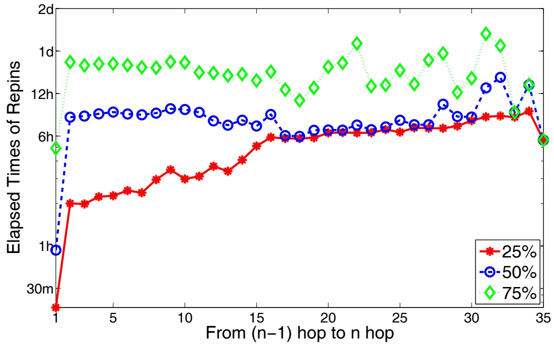
We first examine the structural patterns of pin propagations (i.e., pin-trees). In this analysis, we consider 144,080 pin-trees which have at least one repinner; 57% of (original) pins are not propagated in our dataset. Figure 7 shows the distributions of the number of repinners, max depth, max width, the number of likes, and the number of comments of a pin-tree, respectively. As shown in Figure 7(a), 77% of pin-trees have fewer than 10 repinners while top 0.5% of pin-trees have more than 100 repinners. The average, median, maximum number of repinners of a pin-tree are 8.26, 4, and 3,400, respectively. When we look at the number of likes and comments in Figures 7(d) and 7(e), a small portion of pins have received great attention; e.g., while the average number of likes of pin-trees is 2.95, the maximum number of likes of pin-trees is 1,303. We also find that 67% of pin-trees have the max depth of 1 in Figure 7(b), which means that the propagation of pins tends to be bounded in one hop for the majority of pin-trees. The median max depth and max width are 1 and 3, which suggests that pin-trees in Pinterest are usually wider than deeper. Note that the maximum max depth and max width are 35 and 1,066; they are two-orders of magnitude different.

5.2.2 *Temporal analysis of repinning*

We next investigate how fast repins spread in Pinterest. Figure 8(a) plots the distributions of the first-repin times (i.e. time from the original pinning to the first repinning) and the average inter-repin times of pin-trees, respectively. As shown in Figure 8(a), 52%, 90%, and 97% of first-repinnig occur within an hour, 6 hours, and a day, respectively, which means a pin in Pinterest tends to spread rapidly. However, first-repinnings of a few pin-trees occur a month later; the maximum first-repin time in our dataset is about 40 days. The average and median inter-repin times are 1,187 and 377 minutes, respectively, which suggests that repinning usually occurs within 20 hours. In Figure 8(b), we also show the elapsed times of repins from $(n-1)$ hop to n hop on a pin-tree. As shown in Figure 8(b), first-repinnings (i.e., the first hop) only take less than 10 minutes in the 25st percentile, an hour with the median, and 6 hours in the 75st percentile, which signifies the fast interest diffusion in Pinterest since most of pin-trees have a depth of one. Interestingly, the median inter-hop times are around 8 hours from the second hop.



(a) Elapsed times of repins



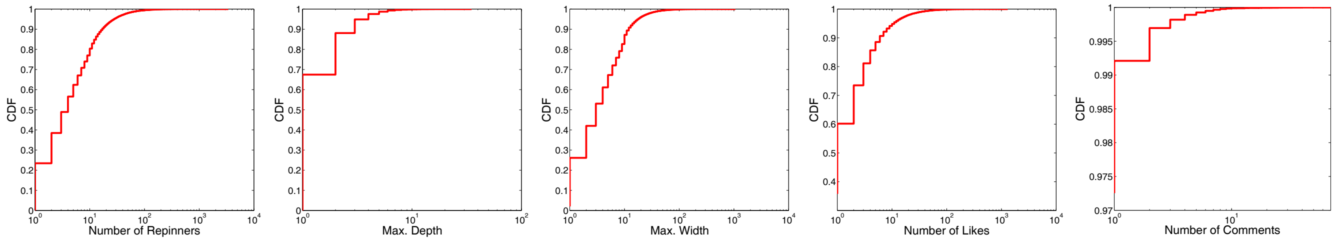
(b) Elapsed times of repins from $(n-1)$ hop to n hop

Figure 8: Elapsed times of repins.

5.3 Pin Propagation Factor Analysis

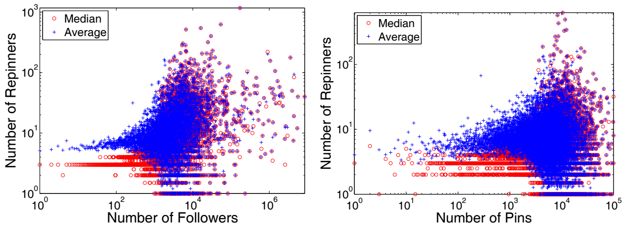
We now analyze which factors affect the pin propagation in Pinterest: (i) pinner's influence and (ii) pin's influence itself. For the first factor, we investigate two influence metrics of pinners: (a) the number of followers and (b) the number of pins the pinner has. Figure 9 shows the average and median numbers of repinners for each pin-tree against the numbers of followers and pins of the pinner, respectively. As shown in Figure 9, the average number of repinners is not shown to be affected by the number of followers or pins of the pinner. To quantify the correlation between the number of repinners of the pin-tree and the pinner's influence (i.e., number of followers and pins), we calculate the Pearson's correlation coefficient [18], denoted by ρ. The ρ values of the number of followers and the number of pins the pinner has are 0.046 and 0.007, respectively, which indicates that there is little correlation between the number of repinners and the pinner's influence.

To examine the second factor, we examine (i) the number of likes of an original pin and (ii) the total number of likes in a pin-tree against the number of repinners of the pin-tree in Figure 10. As shown in Figure 10, there is a significant positive correlation (Pearson's correlation coefficient $\rho = 0.689$) between the number of likes in a pin-tree and the number of repinners, which means the number of repinners increases as the pin becomes increasingly popular. Likewise, there is also a significant positive correlation ($\rho = 0.423$) between



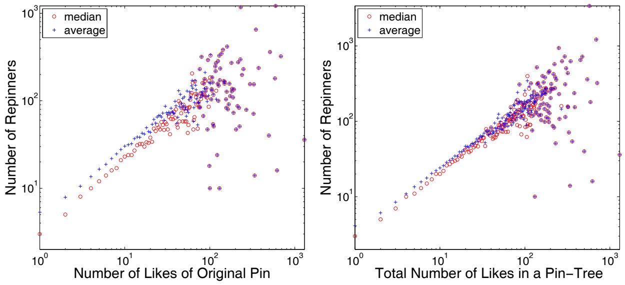
(a) Repinner (b) Max Depth (c) Max Width (d) Like (e) Comment

Figure 7: Distributions of the number of repinners (a), max depth (b), max width (c), the number of likes (d), and the number of comments (e) of a pin-tree.



(a) Followers (b) Pins

Figure 9: There is no strong correlation between the number of repinners of the pin-tree and the pinner's influence (i.e., number of followers and pins).



(a) Number of likes of the original pin (b) Total number of likes in a pin-tree

Figure 10: There is a significant correlation between the number of repinners of the pin-tree and the popularity of the pin.

the number of likes of the original pin and the number of repinners; we can predict whether the number of repinners in a pin-tree will be large or not by looking at the number of likes the original pin has received.

In summary, pin propagations in Pinterest are not due to pinner's influence on Pinterest but due to the popularity of the pin. In the following sections, we will further investigate pin propagation patterns (i.e., number of repinners and inter-repin time of a pin-tree) depending on different properties of pins (e.g., categories, sources).

5.4 Category and Source Analysis

This subsection analyzes pin propagation patterns based on the 32 Pinterest-defined categories and the top 20 sources which are sorted in terms of the number of corresponding pins. Overall, there are notable differences in pin propagation patterns across different categories and sources. We observe that even though there are many pin-trees in particular categories or from particular sources, the number of repinners per pin-tree may not be high, and vice versa. We also find that a large portion (65%) of top 20 sources are in the specific categories where their pins are mostly posted. We describe detailed analysis below.

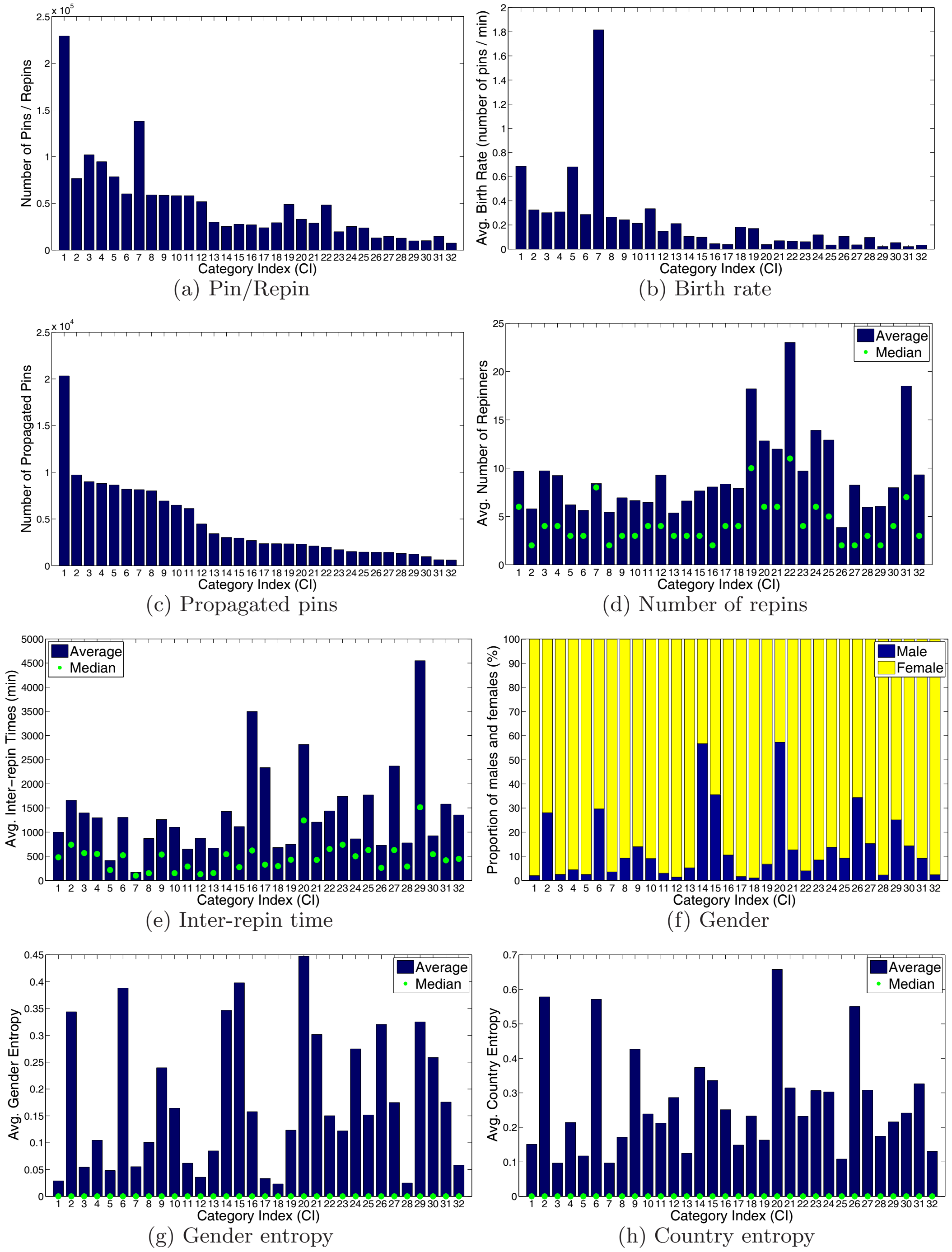
5.4.1 Category analysis

We investigate the pin propagation patterns in Pinterest across different categories as described in Table 1. There were 32 categories in Pinterest during the period of our measurements, which are sorted in terms of the number of pin-trees in Table 1. Figures 11(a), 11(b), and 11(c) show the number of posted pins/repins, the birth rate of pins, and the number of propagated (original) pins in each category, respectively. Here, the birth rate is the number of newly posted pins per minute; the average birth rate of Pinterest during our measurement period is 2.39/min. The number of propagated pins indicates the number of pin-trees, each of which has at least one child. As shown in Figure 11(a), "diy & crafts" (category index (CI) 1) and "food & drink" (CI 7) have the highest and second highest number of pins/repins, respectively. Intuitively, their birth rates should be proportionally high; however, Figure 11(b) reveals that new pins are posted more frequently in the "food & drink" category than the "diy & crafts" category. That is, the pins in "food & drink" tend to be not propagated much. Note that pins are propagated more broadly in the "diy & crafts" than the ones in the "food & drink". Overall, newly-posted pins in the "diy & crafts" tend to be propagated most broadly among all the categories.

Figures 11(d) and 11(e) plot the average number of repinners and the average inter-repin time of each category, respectively. As shown in Figure 11(d), the average numbers of repinners in the categories of "humor" (CI 19), "quotes" (CI 22), "geek" (CI 24), and "tattoos" (CI 31) are higher than the ones in other categories. This is interesting because these categories are usually for particular communities. When we look at the average inter-repin times in each category, we find that the "food & drink" (CI 7) exhibits the shortest inter-repin time, which implies that the "food & drink" category is the most active category since pins are posted and spreading quickly (the birth rate is the highest while the inter-repin time is the shortest). Likewise, the pins in the "health & fitness" (CI 5) are also posted and spreading quickly. The average inter-repin times of the "sports" (CI

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

Table 1: Pinterest categories with indexes.



(a) Pin/Repin

(b) Birth rate

(c) Propagated pins

(d) Number of repins

(e) Inter-repin time

(f) Gender

(g) Gender entropy

(h) Country entropy

Figure 11: Pin propagation patterns across different categories.

Index	Source URL	Type	Alexa rank
1	blogspot.com	blog	13
2	tumblr.com	microblog	25
3	fitsugar.com	fitness	8,127
4	buzzfeed.com	news	178
5	teacherspayteachers.com	education	5,446
6	imdb.com	movie	48
7	saatchionline.com	art	11,745
8	wordpress.com	blog	15
9	akamaihd.net	CDN (image)	72
10	designspiration.net	design	10,838
11	womenshealthmag.com	fitenss	5,624
12	greatist.com	fitness	3,715
13	archdaily.com	architect	3,277
14	blog.com	blog	1,266
15	google.com	search engine	1
16	ebay.com	shopping	20
17	fitnessmagazine.com	fitness	6,088
18	wikipaintings.org	painting	21,912
19	streetartutopia.com	street art	51,788
20	houzz.com	home deco	586

Table 2: A summary of top 20 sources. We fetch ranking information of each source from Alexa [1] as of Nov. 2013.

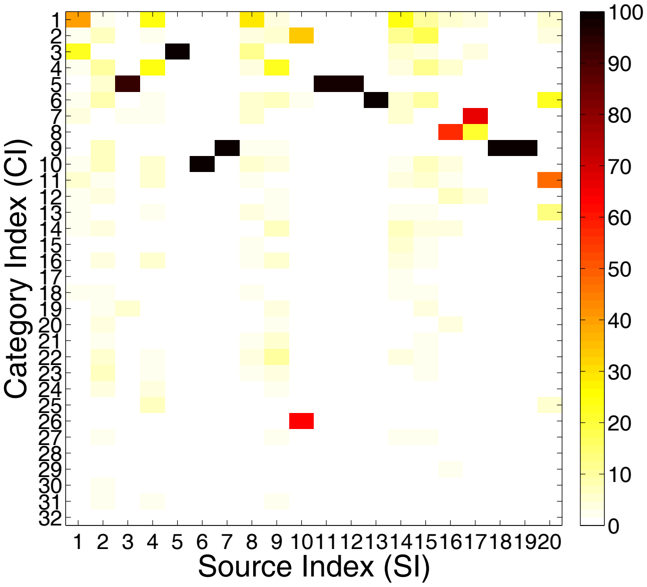


Figure 12: A heatmap of the portion of each source across different categories.

29) and "travel" (CI 16) are longer than those of the other categories, which indicates that pins of these categories tend to spread slowly.

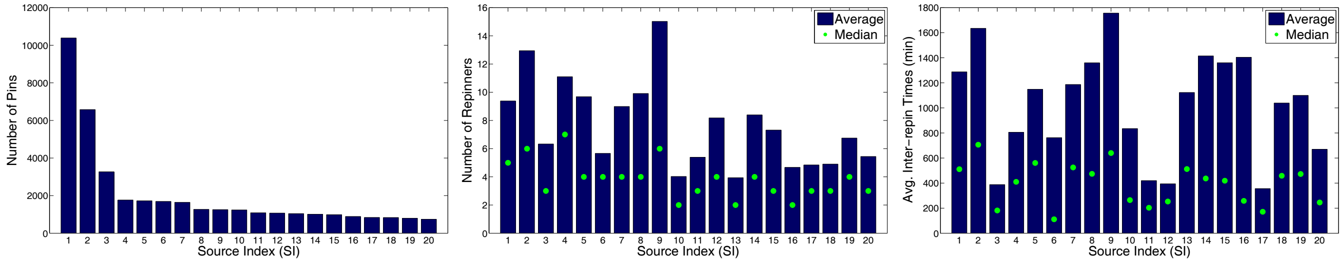
To investigate how the gender and country distributions of pin-trees are skewed in each category, we calculate the entropies for the gender (H_{gender}) and the country ($H_{country}$) as similar to Equation 1 in Section 4.1, respectively. Figure 11(f) shows the ratio of males and females in each category. As shown in Figure 11(f), the ratios of males in the "design" (CI 2), "architecture" (CI 6), "cars & motorcycles" (CI 14), "technology" (CI 15), "men's fashion" (CI 20), and "illustrations & posters" (CI 26) are higher than those in the other categories. Interestingly, the categories of "design" and "illustrations & posters" seem to be not so inclined to females; by the deep analysis, we find that there is a number of male designers in Pinterest. Figure 11(g) shows that the gender entropies of the "architecture" (CI 6), "technology" (CI 15), and "men's fashion" (CI 20) are higher than those of the other categories, which reflects the relatively high interest of males. However, the gender entropies of the most of female-oriented categories like "hair & beauty" (CI 18) and "kids" (CI 28) are low, which means that pin-trees in those categories show a highly skewed distribution in terms of gender. The country entropies of "design" (CI 2), "architecture" (CI 6), "men's fashion" (CI 20), and "illustrations & poster" (CI 26) are higher than the other categories, implying a relatively even distribution of countries. This is interesting since those categories with the higher country entropies often also have the higher gender entropies, which means those categories are relatively less dependent on gender or country. Note that the median values of the gender and country entropies for each category are zero, respectively, which indicates that most of pin-trees are skewed in terms of both gender and country.

5.4.2 Source analysis

We next investigate the source information of each pin, which was briefly mentioned in Section 2.1. Here, a source is a URL which the pin comes from. Table 2 summarizes the information of the top 20 sources, which are sorted in terms of the number of corresponding pins. We further manually examine the type and Alexa rank of each source as shown in Table 2. We find that there are various types such as blog, image, news, and fitness; interestingly, the fitness type takes a substantial portion in the top 20 sources (4 of 20). Also, the Alexa ranks of a half of the top 20 sources is greater than 3,000, which means they are not so popular web sites but popular sources in Pinterest.

Figure 12 shows the heatmap of the portion of each source across 32 categories. For example, *imdb.com* (source index (SI) 6) and *saatchionline.com* (SI 7) are mostly consumed (colored black) in the categories of "film, music & books" (CI 10) and "art" (CI 9), respectively. Also, 60% of pins in the category of "illustrations & posters" (CI 26) and 40% of pins in the category of "art" (CI 9) are from the source *designspiration.net*, an image-based website of art galleries. We find that 65% of the top 20 sources are strongly related to specific categories where their pins are mostly posted (over 60% of pins are posted on a specific category). However, pins of some sources like "tumblr.com" (SI 2) or "google.com" (SI 15) are spread across several categories.

Figure 13 analyzes the pin propagation patterns depending on the sources. As shown in Figure 13(a), the numbers of corresponding pins of top 3 sources are about 10,000, 6,000, and 3,000, respectively, which accounts for over 60% of total pins, while those of the other sources are around 1,800. From Figure 13(b), the numbers of repinners are higher in the source indexes 2, 4, 8, and 9; their pins spread across several categories since they are mostly general websites like blogs or news; however, the speeds of spreading of the pins of these sources are mostly slower as shown in Figure 13(c). On the contrary, sources related with "fitness" (SI 3, 11, 12, and 17) have a small number of repinners but their pins spread faster.



(a) Number of corresponding pins (b) Number of repinners (c) Avg. inter-repin time

Figure 13: Pin propagation patterns depending on different sources.

5.4.3 Analysis on Top 1% Pin-Trees

We finally investigate the top 1% of all the pin-trees in terms of the number of repinners. We observe that the averages of the number of repinners, number of likes, max depth, and max width of top 1% pin-trees are 299, 60, 5, and 68, respectively, which are significantly higher (17 times for the number of repinners, 20 times for the number of likes, 3.3 times for the max depth, and 11.3 times for the max width) than those of all the pin-trees. Figure 14 shows the numbers of the top 1% pin-trees across different categories and sources, respectively. We notice that a large number of the top 1% pin-trees belong to the "humor" (CI 19) and "quotes" (CI 22) even though there are relatively fewer pin-trees in those categories. On the contrary, a small number of the top 1% pin-trees belong to the "design" (CI 2), while it ranks second among 32 categories in terms of the total number of pin-trees (See Figure 11(c)). This implies that the top 1% pin-trees exhibit different patterns compared to all of the pin-trees. When we look at the top 1% pin-trees depending on the top 20 sources in Figure 14(b), we find that most of the top 1% pin-trees belong to *blogspot.com* (SI 1) and *tumblr.com* (SI 2). We further observe that the top 1% pin-trees often belong to *akamaihd.net* (SI 9), an image-hosting content delivery network (CDN) for *Facebook.com*.

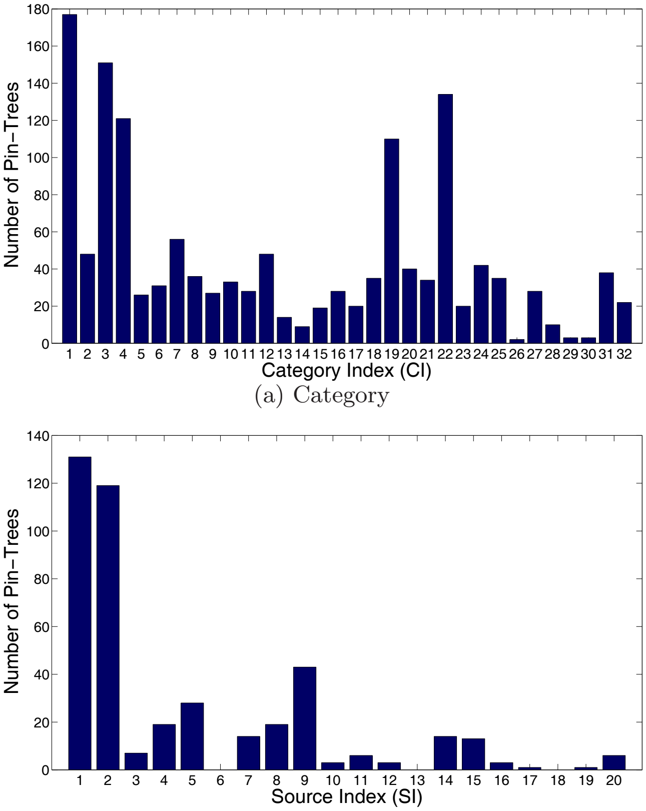
6. INTEREST GRAPH ANALYSIS

In this section, we investigate how users are related to one another in terms of shared interests. To this end, we introduce the notion of *"Interest Graph"* to represent the relations among users.

6.1 Definition of Interest Graph

We assume that a directed weighted graph $I = (V, E, W)$ represents an interest graph where V is the set of users and E is the set of (directed) edges between two users. That is, there exists an edge $E(V_i, V_j)$ from user V_i to user V_j if user V_j repins a pin from user V_i. The weight $W(V_i, V_j)$ of an edge $E(V_i, V_j)$ is the total number of propagated pins from V_i and V_j. Overall, the interest graph I in Pinterest is designed to show how users are related to one another by their shared interests. Figure 15 illustrates an interest graph, where three pins A, B, and C are shared among the users. As shown in Figure 15, the interest graph I is constructed by merging the three pin-trees. The weights $W(V_1, V_3)$ and $W(V_6, V_7)$ are two in this example.

The numbers of nodes and edges in the interest graph of Pinterest are 936,795 and 1,170,192, respectively, which means the interest graph of Pinterest is sparse. In particular, the average degrees and weights of I are 2.50 and 1.29, respectively, meaning that most of users are likely to be connected with two or three people and share one or two pins on average. We also calculate that the clustering coefficient [24] of I, which is defined as the probability that two neighbors of a given node are also neighbors. We find that the clustering coefficient of I is much higher than that of a random network (with the same numbers of nodes and edges), by a factor of 207 or larger, which suggests a significant "small-worldness" of I in Pinterest.



(a) Category

(b) Source

Figure 14: Numbers of the top 1% pin-trees across different categories (a) and sources (b).

6.2 Community Analysis

To investigate how users form a group (or a community) in the interest graph I, we first identify communities of I using

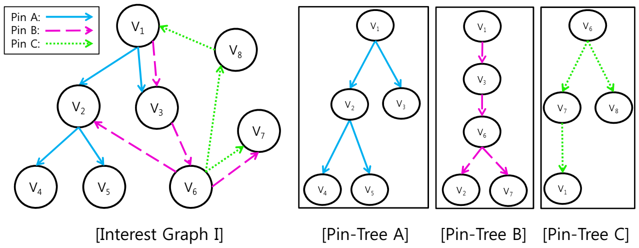


Figure 15: An interest graph with three pins propagated.

Property	Community	Random
Gender	0.81	0.77
Country	0.91	0.82
Major Category	0.21	0.11
Interest Feature	0.34	0.23
Category	**0.94**	**0.03**
Source	**0.89**	**0.01**

Table 3: Uniformities of properties of users or shared interests in the same community.

the Louvain method, a well-known fast community detection algorithm that maximizes the ratio of the number of edges within communities to that of edges across communities [10]. Note that we use the weighted version of Louvain method. The number of identified communities is 15,936 and the average number of members of a community is 58.78. Based on the identified communities, we examine what makes users belong to the same community. To this end, we devise a *uniformity* metric U, which quantifies how much similarity exists among members or among shared interests. That is, U denotes the probability that randomly selected two users in the same community have the same property (e.g., gender or country). The uniformity U is defined as:

$$U = \frac{2}{\sum_{k=1}^{c} n_k \left(n_k - 1\right)} \sum_{k=1}^{c} \sum_{i=1}^{n_k - 1} \sum_{j=i+1}^{n_k} \delta\left(P\left(v_i^k\right), P\left(v_j^k\right)\right) \qquad (2)$$

where c is the number of communities in the given network, n_k is the number of nodes (or users) in the community k, and v_i^k means node i in the community k. Note that $P(v)$ is the property (e.g., gender or country) of user v and δ is the Kronecker's delta ($\delta(i,j) = 1$ if $i = j$, and $\delta(i,j) = 0$ otherwise).

We conjecture that users may belong to the same community if properties of users or shared interests are similar. To validate this conjecture, we first consider four properties of users: (i) gender, (ii) country, (iii) major category to which the majority of (the user's) pins belong, and (iv) interest feature that characterizes how user's interests are distributed across the 32 categories. The interest feature is defined as a vector with 32 entries [C_1, C_2, ..., C_{32}], each of which is a portion of pins that a user has posted to the corresponding category C_i. For the first three properties (gender, country, major category), each of the uniformity can be calculated by Equation 2. To calculate the uniformity for the interest feature, we use the cosine-similarity function between the interest features of two users instead of using δ in Equation 2. In addition to the properties of users, we further consider two properties of shared interests: (i) category and (ii) source of pins. Since the shared interests are represented by edges (which show pin propagation) in the interest graph, we calculate the uniformities for the category and source by replacing nodes with edges in Equation 2.

Table 3 shows the uniformity of each property. For comparison purposes, we randomly select two nodes (or edges) 10,000 times in the interest graph and calculate its uniformity (denoted by Random). As shown in Table 3, the uniformities of shared interests are significantly higher than those by Random; shared pins in the same community (i) belong to the same category or (ii) come from the same source with high probability. That is, users who have similar tastes or share similar pins tend to form a community in the interest graph. The user properties considering her tastes or interests (i.e., the major category and the interest feature) exhibit relatively higher uniformities compared to Random, which indicates that users with similar interests are likely to belong to the same community. Users in the same community are likely to be the same gender or belong to the same country with somewhat higher probability compared to those by Random.

7. APPLICATION

We have analyzed how and why pins are collected, organized, and propagated in Pinterest. Our empirically-grounded evidences suggest that most of pin propagations in Pinterest are driven by interest. In this section, we seek to answer the following question, which may be important for understanding the patterns of pin consumptions in Pinterest: *What pins will be consumed (pinned or repinned) by each user in the future?* To answer the question, we perform a trace-driven simulation study.

7.1 Trace-driven Simulation

To conduct a trace-driven simulation, we first select 4,667 target users, each of which has at least 10 pins in our dataset. Based on the prediction methods that we describe in Section 7.2, we make a candidate pin list (which consists of 20 pins) to be possibly consumed for each target user (as of Jul. 18, 2013). To validate whether the predicted pins are actually consumed by each target user after 125 days, we check all the pin information that the target user has (as of Nov. 20, 2013). To this end, we have collected additional dataset for 10 days from Nov. 20 to Nov. 30, 2013, which consists of 25 M web pages that contain the pin information of the target users. For evaluation, we measure precision, which is defined as the ratio of the number of pins that the user actually possesses in her boards to the total number of predicted pins.

7.2 Prediction on Pin Consumption

We first present two approaches for predictions: (i) a user-centric approach and (ii) a pin-centric approach. The user-centric approach first finds similar users, and then predicts pins based on the similarity among the users. The rationale behind this approach is that a user may consume pins that her similar users possess. The similarity between two users can be calculated in two ways: (a) $user-property-based$ and (b) $user-interest-based$. The $user-property-$ $based$ considers eight properties of users: the gender, coun-

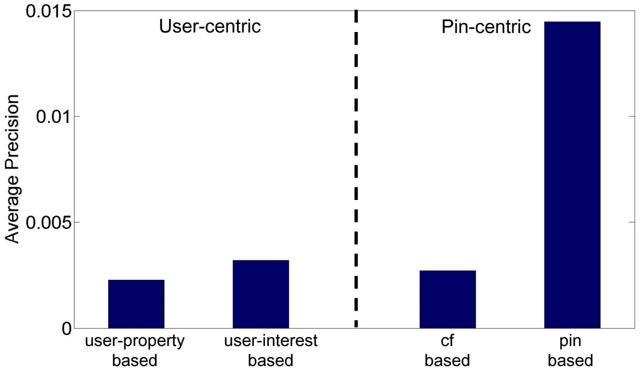


Figure 16: Average precision for each prediction method.

try, major category to which the majority of (the user's) pins belong, number of pins, number of boards, number of categories, number of followers, and number of followings. For the first three properties, the similarity in the $user - property - based$ is calculated by counting the number of matched entries. For the other five entries, we first sort all the users in descending order in each entry, who are divided into five partitions, and check whether the corresponding numbers of two users belong to the same partition. The $user-interest-based$ considers the interest feature (introduced in Section 6) that characterizes how user's interest is distributed across the 32 categories. We use the cosine-similarity between the two interest features of two users in the $user - interest - based$.

The pin-centric approach finds similar pins, not users. This approach assumes that a user may consume similar pins with ones she has. That is, the pin-centric approach finds 4 similar (candidate) pins for each of the five pins recently posted by the target user (20 pins in total). To find similar pins, we first adopt the *(item-to-item) collaborative filtering (CF)* [22, 28] technique. We call this technique $cf - based$, whose basic idea is to find similar pins based on the (collective) opinions of other like-minded users. A key advantage of this technique is that it does not require any knowledge of pins in advance. In the $cf - based$, the similarity between two pins is calculated by the cosine similarity [28]. We further suggest another simple but effective prediction method in this approach: $pin - based$. The $pin - based$ considers both the properties of a pin (i.e., category and source) and collective opinions of other users (i.e., number of likes of the original pin in its pin-tree) to the pin. This method reflects the lessons from the measurements in Section 5, which show that the propagation of a pin is substantially related to pin's popularity (i.e., number of likes), category, and source.

Figure 16 shows the average precision of each method. Our $pin - based$ method outperforms the others (around 4.5 times against the $user - property - based$, 3.3 times against the $user - interest - based$, and 4.5 times against the $cf - based$), which indicates that predicting based on the properties of the pins is more accurate than others. That is, the properties of pins are more important factors than those of users to predict pin consumption patterns in Pinterest; this can be an important implication on designing personalized services in Pinterest-like social networks.

8. CONCLUSIONS

We have conducted a comprehensive measurement study to understand how people collect, manage, and share pins in Pinterest. With the dataset, we analyzed: (1) how users collect and curate pins, (2) how users share their pins and why, and (3) how users are related by shared pins of interests. We found that pin propagation in Pinterest is mostly driven by pin's properties like its topic or content, not by user's characteristics like her number of followers. We also observed that there are notable differences in pin propagation patterns across different categories and sources. We further revealed that users in the same community in the interest graph of Pinterest share pins in the same category and from the same sources with high probability. Our empirically-grounded simulation demonstrated that the properties of pins are more important factors than those of users for accurately predicting pin consumption patterns in Pinterest; we believe this observation has an important implication on designing efficient personalized services in Pinterest-like social networks.

9. ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their valuable feedback. This research was supported in part by Basic Science Research Program through the "National Research Foundation of Korea (NRF)" funded by the Ministry of Science, ICT & future Planning (2013R1A2A2A01016562) and in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (No. 2013-R1A1A2010474). Jinyoung Han is currently a post-doctoral researcher in the RUBINET research group at University of California-Davis, and supported by National Science Foundation CNS-1302691 grant.

10. REFERENCES

[1] Alexa - the web information company. `http://www.alexa.com`.

[2] Can ben silbermann turn pinterest into the world's greatest shopfront? `http://www.fastcodesign.com/1670681/ben-silbermann-pinterest`.

[3] The demographics of social media users - 2012. `http://pewinternet.org/Reports/2013/Social-media-users.aspx`.

[4] Interest, meet pinterest: Site helps users catalog their passions. `http://edition.cnn.com/2012/01/26/tech/web/pinterest-website/index.html`.

[5] Pinterest hits 10 million u.s. monthly uniques faster than any standalone site ever -comscore. `http://techcrunch.com/2012/02/07/pinterest-monthly-uniques`.

[6] Pinterest traffic growth soars to new heights: Experian report. `http://www.huffingtonpost.com/2012/04/06/pinterest-traffic-growth_n_1408088.html`.

[7] Why pinterest is 2012's hottest website. `http://edition.cnn.com/2012/02/06/tech/web/pinterest-website-cashmore`.

[8] Wedding dresses and wanted criminals: Pinterest.com as an infrastructure for repository building. In *ICWSM*, 2013.

[9] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *WWW*, 2012.

[10] 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.

[11] M. Cha, F. Benevenuto, H. Haddadi, and K. Gummadi. The world of connections and information flow in twitter. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 42(4):991–998, 2012.

[12] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM*, 2010.

[13] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *WWW*, 2009.

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

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

[16] C. Hall and M. Zarro. Social curation on the website pinterest.com. *Wiley American Society for Information Science and Technology*, 49(1):1–9, 2012.

[17] R. A. Hanneman and M. Riddle. Introduction to social network methods, 2005. `http://www.faculty.ucr.edu/~hanneman/`.

[18] R. Joseph Lee and N. W. Alan. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

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

[20] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, 2010.

[21] K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *ICWSM*, 2010.

[22] G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[23] M. O. Lorenz. Methods of Measuring the Concentration of Wealth. *Publications of the American Statistical Association*, 9(70):209–219, 1905.

[24] M. E. J. Newman. The structure and function of complex networks. *SIAM REVIEW*, 45:167–256, 2003.

[25] 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.

[26] J. Park, M. Cha, H. Kim, and J. Jeong. Managing bad news in social media: A case study on domino's pizza crisis. In *ICWSM*, 2012.

[27] T. Rodrigues, F. Benevenuto, M. Cha, K. Gummadi, and V. Almeida. On word-of-mouth based discovery of the web. In *ACM IMC*, 2011.

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

[29] C. E. Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64, 1951.

[30] D. Wang, Z. Wen, H. Tong, C.-Y. Lin, C. Song, and A.-L. Barabási. Information spreading in context. In *WWW*, 2011.

[31] S. Ye and S. F. Wu. Measuring message propagation and social influence on twitter.com. In *International Conference on Social Informatics*, 2010.

[32] M. Zarro and C. Hall. Pinterest: Social collecting for #linking #using #sharing. In *12th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2012.

[33] S. Zoghbi, I. Vulić, and M.-F. Moens. I pinned it. where can i buy one like it?: Automatically linking pinterest pins to online webshops. In *ACM Workshop on Data-driven User Behavioral Modelling and Mining from Social Media*, 2013.