Pinned it! A Large Scale Study of the Pinterest Network

Sudip Mittal, Neha Gupta, Prateek Dewan, Ponnurangam Kumaraguru
Indraprastha Institute of Information Technology, Delhi (IIIT-D)
{sudip09068, neha1209, prateekd, pk}@iiitd.ac.in

ABSTRACT

Pinterest is an image-based online social network, which was launched in the year 2010 and has gained a lot of traction, ever since. Within 3 years, Pinterest has attained 48.7 million unique users. This stupendous growth makes it interesting to study Pinterest, and gives rise to multiple questions about it's users, and content. We characterized Pinterest on the basis of large scale crawls of 3.3 million user profiles, and 58.8 million pins. In particular, we explored various attributes of users, pins, boards, pin sources, and user locations, in detail and performed topical analysis of user generated textual content. The characterization revealed most prominent topics among users and pins, top image sources, and geographical distribution of users on Pinterest. We then tried to predict gender of American users based on a set of profile, network, and content features, and achieved an accuracy of 73.17% with a J48 Decision Tree classifier. We then exploited the users' names by comparing them to a corpus of top male and female names in the U.S.A., and achieved an accuracy of 86.18%. To the best of our knowledge, this is the first attempt to predict gender on Pinterest.

Categories and Subject Descriptors
H.3.5 [Online Information Services]: Web-based services

Keywords
Online social networks, Pin, Classification

1. INTRODUCTION

Online Social Networks (OSNs) like Facebook, Twitter, LinkedIn, and Google+ are web-based platforms that help users to interact, share thoughts, interests, and activities. These OSNs allow their users to imitate real life connections over the Internet. A report by the International Telecommunication Union states that the total number of online social media users has crossed the 1 billion mark as of May, 2012 [27]. According to Nielsen's Social Media Report, users continue to spend more time on social networks than on any other kind of websites on the Internet [31]. With this outburst in the number of social media users across the world, online social media has moved to the next level of innovation. While all the aforementioned conventional social media services are mostly text-intensive, some of them have gone beyond text, and have introduced images as their building blocks. Services like Instagram, and Tumblr have gained immense popularity in the recent years, with Instagram (now part of Facebook) attaining 100 million monthly active users, and 40 million photo uploads per day [17]. These numbers indicate successful entrance of image based social networks in the world of online social media.

Pinterest is one of the most recent additions to this popular category of image-based online social networks. Within a year of its launch, Pinterest was listed among the “50 Best Websites of 2011” by Time Magazine [29]. It was also the fastest site to break the 10 million unique visitors mark [8]. Number of users since then have increased, with Reuters stating a figure of 48.7 million unique users in February 2013 [37]. Although fairly new to the social media fraternity, Pinterest is being heavily used by many big business houses like Etsy, The Gap, Allrecipes, Jetsetter, etc. to advertise their products. Further, Pinterest drives more revenue per click than Twitter or Facebook, and is currently valued at USD 2.5 billion [37, 45].

The immense upsurge and popularity of Pinterest has given rise to multiple basic questions about this network. What is the general user behavior on Pinterest? What are the most common characteristics of users, pins, and boards? What is the sentiment associated with user-generated textual content? What is the geographical distribution of users? Is it possible to predict gender of Pinterest users? There exists little research work on Pinterest [2, 5, 22, 32, 42, 43]; but none of this work addresses the aforementioned basic questions. To answer these questions, and get deeper insights into Pinterest, we collected and analyzed a dataset comprising of user details (3,323,054), pin details (58,896,156), board details (777,748), and images (498,433). We applied multiple machine learning algorithms to predict gender on a true positive data-set of 6,309 male and 6,309 female Pinterest users living in U.S.A.

Based on our analysis, some of our key contributions are summarized as follows:

1. Topical analysis of user generated textual content on social media users living in U.S.A.
Pinterest: We found that the most common topics across users, and pins were design, fashion, photography, food, and travel.

2. User, pin, and board characterization: We analyzed various user profile attributes, their geographical distribution, top pin sources, and board categories. Less than 5% of all images on Pinterest are uploaded by users; over 95% are pinned from pre-existing web sources.

3. Gender prediction for American users: We extracted true positive gender information from Facebook, for over 66,000 Pinterest users from U.S.A., and were able to achieve an accuracy of 86.18% while predicting gender. We applied various machine learning algorithms using multiple content and network based features from Pinterest.

The rest of the paper is organized as follows. We discuss the related work in Section 2. We then discuss Pinterest as a social network in Section 3. In Section 4, we describe our data collection methodology. Analysis of the collected data and its results are covered in Section 5. Section 6 contains discussion, limitations, and future work.

2. RELATED WORK

Social network characterization.

Online social networks, in general, have been studied in detail by various researchers in the computer science community. Mislove et al. conducted a large scale measurement study and analysis of Flickr, YouTube, LiveJournal, and Orkut [30]. Their results confirmed power-law, small-world, and scale-free properties of online social networks. In a more recent work, Magno et al. performed a detailed analysis of the Google+ network, and identified some key differences and similarities between Google+, and existing social networks like Facebook, and Twitter [28]. Ugander et al. performed a large-scale analysis of the entire Facebook social graph and found that 99.91% of all the users belonged to a single large connected component [40]. They confirmed the 'six degrees of separation' phenomenon and showed that the value had dropped to 3.74 degrees of separation in the entire Facebook network of active users.

Pinterest Introduction.

Considering the rapid growth rate of Pinterest since its launch, there still exist only a few studies on this social network. In closely related work by Gilbert et al., authors presented a statistical overview of the Pinterest network, and showed that female users get more repins but lesser followers on Pinterest [12]. Their analysis was based on a smaller dataset of 2.9 million pins, and 989,355 users, in contrast to our dataset of over 58 million pins, and 3.3 million users. Chang et al. worked towards finding activity patterns for attracting attention on Pinterest. Some of the key findings of this work revealed that male users were not particularly interested in stereotypically male topics; sharing diverse content increases attention to a certain level; and homophily drives repinning. Their dataset consisted of 46,365 users, and 3.1 million pins [6]. Ottoni et al. analyzed Pinterest in a gender-sensitive fashion, and found that the network was heavily dominated by female users. Authors of this work found that females on Pinterest make more use of lightweight interactions than males, invest more effort in reciprocating social links, are more active and generalist in content generation, and describe themselves using words of affection and positive emotions. This study spanned across a large dataset consisting of over 2 million users [32]. Kamath et al. described a supervised model for board recommendation on Pinterest. They used a content-based filtering approach for recommending high quality information to users [21]. Du denhoffer et al. tried to use Pinterest as a library marketing and information literacy tool at the Central Methodist University. They reported that the number of followers viewing the library pinboards had outpaced the usage of the text-based lists in just one semester [10]. In another similar work by Zarro et al., authors talked about how digital libraries and other organizations could take advantage of Pinterest to expand the reach of their material, allowing users to create personalized collections, incorporating their content [42]. In their next piece of work, Zarro et al. found that Pinterest serves as infrastructure for repository building that supports discovery, collection, collaboration and publishing of content, especially for professionals [43].

Gender prediction on other social networks.

Rao et al. attempted to predict gender of Twitter users based on a rich set of profile, content, and network attributes, and achieved an accuracy of 72.33% using a SVM classifier. This was the first attempt to predict gender on Twitter [36]. Burger et al. achieved a 74% accuracy using Balanced Winnow2 classifier for predicting gender of Twitter users. Their corpus comprised of 4.1 million tweets, and 15.6 million distinct features [4]. Pennacchiotti et al. tried to extract gender information from Twitter users’ profiles by applying regular expressions on users’ bio field. Authors were able to extract gender information of 80% users from a sample of 1.4 million users, but with a very low accuracy. A manual annotation of over 15,000 users using only profile / avatar picture revealed that only 57% images were correlated with a specific gender [33]. Tang et al. applied a name-centric approach for predicting gender of New York City Facebook users, and achieved an accuracy of 95.2% [39]. Zeleeva and Getoor [44] proposed techniques to predict the private attributes of users in four real-world datasets (including Facebook) using general relational classification and group-based classification. Their accuracy for gender inference with their Facebook dataset, was 77.2% based on users’ group affiliations, and the sample dataset used in their study was quite small (1,598 users in Facebook). Other papers [15, 16, 26, 41] have also attempted to infer private information inside social networks. Methods they used are mainly based on link-based traditional Naive Bayes classifiers.

3. UNDERSTANDING PINTEREST

Pinterest is an image-based social bookmarking media, where users share images which are of interest to them, in the form of pins on a pinboard. It emphasizes on discovery and curation of images rather than original content creation. 2 This makes Pinterest a very promising conduit for the promotion of commercial activities online.

---


Similar to other OSNs, Pinterest also uses some specific terminology to refer to various elements and services it provides. Some terms are as follows:

1. **Pins**: A pin is an image that has some meta-data information associated with it. Pins can be thought of as basic building blocks of Pinterest. The act of posting a pin is known as **pinning**, and the user who posts a pin is the **pinner**. Similar to images on Facebook, pins can be **liked** and **shared**. Each of these pins has the following meta-data associated with it – **unique pin number**, **description**, **number of likes**, **number of comments**, **number of repins**, **board name**, **source**, and **content in comments**. The act of sharing an already existing pin is referred to as **repinning**.

2. **Pinboards**: They are a themed collection of pins, organized by a user. Each board (“boards” and “pinboards” are used interchangeably) has a name, a description (optional), category (optional, e.g. Animals, Art, Celebrities, Food and Drink, Design, Education, Gardening), and an option to make it Secret. Secret boards are only visible to the users who create them. This analogy of pins and pin-boards replicates the real-world concept of images on a scrapbook.

3. **Source**: Each pin on Pinterest has a source URL associated with it. As the name suggests, this is the actual URL from which the image has been pinned by a user. Images uploaded by users directly to Pinterest from their local computer, have **pinterest.com** as their source, whereas images which are pinned from an existing website (e.g. flickr.com) have this source website (flickr.com) as the source.

4. **Pin-It button**: A Pin-It button is a browser bookmark used to upload content to Pinterest. Some popular websites like Amazon, eBay, BHG, and Etsy also provide their own pin-it button next to their product images. This pin-it button makes it easier for a user to share the content that she likes on Pinterest.

### 3.1 User Accounts

A user begins by creating an account using her Facebook ID, Twitter ID, or an email address. On account creation, Pinterest asks each new user to follow 5 boards to complete the creation process, as a mandatory step to get started. Each user has a profile page (Figure 1) that is publicly visible to everyone, listing the user’s name, a description, location, connected Facebook account (if available), connected Twitter account (if available), a profile website, boards (which are not secret), and associated pins, likes, followers, and followees. A user also has a timeline where all pins from the users she follows, are displayed.

### 3.2 Social Ties

A user has the option to follow a particular user or a specific board of any other user. If a user follows another user, she gets updates about all the boards owned by that user. But, in case a user follows specific boards, she gets updates only from those particular boards. This relationship is quite similar to Twitter’s follower / followee relationship. Interactions on Pinterest are in the form of pins. A user pins an image, and can add a pin description to better describe the pin. Other users can then repin the shared pin, like it or share their views through a comment. These features are similar to Facebook’s share, like, and comment features respectively.

### 4. DATA COLLECTION

In this section, we discuss the methodology that we applied for data collection, and describe the data that we collected. Given the size of the entire Pinterest network (48.7 million users), it would have been hard, and computationally very expensive to be able to capture the entire network.

Pinterest does not provide a public API for data collection. Therefore, in order to collect data, we designed and implemented a breadth first search (BFS) crawler in Python. All data was collected using a Dell PowerEdge R620 server, with 64 Gigabytes of RAM, 24 core processor, connected to a 1 Gbps Internet connection. The entire data collection process spanned from December 26, 2012 to February 1, 2013. Broadly, this process (Figure 2) was split into three phases as described below:

#### 4.1 User Handles Collection

The data collection process was initiated by selecting the top 5 profiles in terms of the number of followers on Pinterest, as initial seeds, and feeding them into the crawler. The crawler first extracted 4,995,974 direct followers of these 5 input seeds, and then repeatedly crawled through the “followers of followers”. We collected a total of 17,964,574 unique user handles through this process, which is slightly over 30% of the entire Pinterest population [37]. We call this, the **userhandles** dataset. This technique of snowball-sampling is commonly used in online social media research [32].

#### 4.2 User Data Collection

Next, we started data collection for user profiles of the 17.96 million user handles collected in the previous step, and obtained a total of 3,323,054 user profiles, called the **userprofile dataset** (we present the analysis on 3.3 million userprofiles in this paper; though our data collection pro-
Figure 2: Flow diagram depicting the flow sequence of our data collection process. The darkened blocks represent our initial seed users, and primary datasets. The lighter blocks denote the additional information extracted through the primary dataset.

This userprofile dataset includes user display name, description field, number of followers, number of followees, number of boards, number of pins, boards, profile website, Facebook handle, Twitter handle, location, pins, and likes. Along with user profiles, we extracted 777,748 boards and their corresponding details (called the boards dataset). These details include board category, number of followers, and number of pins for each pinboard.

Many times users also mention their Facebook and / or Twitter profile URLs on Pinterest. Using this information from the userprofile dataset, we collected publicly available Facebook information of 1,667,973 users (50.19% of the userprofile dataset) and Twitter information of 49,416 users (1.4% of the userprofile dataset). Many Pinterest users also mention location in their profile. We found location details for 331,530 users (9.93% of the userprofile dataset). Some users mentioned only their country, whereas others mentioned their city as well. Some users gave their location as “The beach”, “mentally in lala land”, etc. In order to verify the credibility of such location information, we used Yahoo Placefinder API \(^3\) and obtained the correct details for 192,261 (57.99%) of these locations.

4.3 Pin Data Collection

Using user profiles as seeds, we collected 58,896,156 unique pins and their related information. We call this the pin dataset. This information consists of the pin description, number of likes, number of comments, number of repins, board name, and source for each pin. We also collected a random sample of 498,433 images (called the images dataset) from these pins. For each of these images, we extracted their Exchangeable Image File Format (EXIF) information for further analysis. \(^4\) Most common pieces of EXIF information available were date, time, image description, artist, copyright, and camera make / model. We also extracted information about pin sources for each pin, referred to as the source dataset.

5. ANALYSIS

We now present our analysis of the users, pins, and boards in detail.

5.1 User characterization

5.1.1 Profile description

From our userprofile dataset of 3,323,054 user profiles, we found that only 589,193 (17.73%) users had profile description. We observed that users revealed private details through this field, like age, marital status, personal traits, email IDs, phone numbers, etc. The profile description of one user said, “I am 35, happily married, love kids & cats, and have a disturbing sense of humor!” We extracted 100 most frequently occurring words from the profile description, and found topics like fashion, design, food, music, art, photography, and travel as the most popular user interests (Figure 3). We observed that the most common interests were in line with the most common professions (like artist, designer, cook, photographer) mentioned by the users. This shows that large proportion of Pinterest consumers make use of the network for professional activities.

![Figure 3: Tag cloud of the top 100 words taken from user’s profile description field.](image)

5.1.2 Social and commercial links

Another source of information on the user profile is the “website” field, where users can provide URLs to their personal websites, and blogs. In our dataset, we found that 177,462 (5.34%) users had mentioned a website. The top most domain was Facebook, where 9,697 (5.46%) users had mentioned a link to their Facebook profiles. Twitter, Etsy, YouTube, Flickr, About.me, LinkedIn, etc. were the other domains which constitute the top 10. Apart from the website field, Pinterest separately provides users with an option to connect their Facebook and / or Twitter accounts with their Pinterest profiles. Out of over 3.3 million user profiles that we collected, over 2.71 million users (81.78%) had connected their Facebook profiles with Pinterest. Only 328,570 (9.88%) users connected their Twitter accounts with Pinterest. Less than 4% (132,553) users had connected both Facebook and Twitter, while 12.3% (409,399) users had connected neither. Further, we found that 86,641 (26.36%) out of 328,570 users had identical usernames on Twitter and Pinterest. However, only 5,419 (5.02%) out of 107,910 users had identical usernames on Facebook and Pinterest. Two

---

\(^3\)http://developer.yahoo.com/boss/geo/docs/requests-pf.html

\(^4\)http://fotoforensics.com/tutorial-meta.php#EXIF
hundred and ninety seven (0.22%) users had identical usernames on all three networks. Analysis of usernames for the same user on various social networks can be useful for identity resolution across multiple OSNs [19].

5.1.3 Connections and popularity

The maximum number of followers for a user was found to be 11,992,745 (as of January 2013). Table 1 lists the description, number of followers and followees for the top 10 most followed users on Pinterest (to maintain users’ privacy, we do not mention usernames anywhere). The average number of followers per Pinterest user was found to be approximately 176, as compared to 208 followers per Twitter user [38]. With only one-tenth the number of users as Twitter, this average number of followers depicts that the Pinterest network is very-well connected.

We then plotted the ratio of number of followers versus the number of followees for all users (except for the users with 0 followees) on a log scale as shown in figure 4(a), and found that more than 70% users had more followees than followers. The graph depicts that a very small fraction of users had this ratio skewed, and most users on Pinterest in our dataset had a comparable number of followers and followees. Krishnamurthy et al. [23] found a similar relation between followers and followees for Twitter users.

From the 328,570 users who had connected their Twitter accounts with Pinterest, we extracted the number of Twitter followers and followees for 93,659 users. We then plotted the ratio of followers / followees for these users for both, Pinterest and Twitter, on a log scale, as shown in figure 4(b). As the plot suggests, the ratio of followers / followees on Pinterest was weakly correlated with the ratio of followers / followees on Twitter (correlation = 0.32). Users who were popular on Pinterest were not necessarily popular on Twitter (and vice versa).

5.1.4 Gender distribution

We extracted gender information from Facebook profiles of over 1.85 million users who had linked their Pinterest profiles with Facebook. Over 1.61 million users (87.15%) were females, and only 130,945 users (7.04%) were males. The rest (5.81%) did not have their gender information publicly available. This gender distribution is quite similar to the one observed by Ottoni et al. in their work on Pinterest [32].

5.2 Pin characterization

5.2.1 Pin description

To understand the most common type of pins on Pinterest, we extracted the textual content present in the “pin description” fields from all the pins, and analyzed the most frequently occurring terms. Figure 4(c) represents the tag cloud of the top 100 terms present in pin description. Similar to user descriptions, terms related to food and creative arts dominated the pin description. Other than food, decoration and wedding related pins were also found to be very common in pin description. For example, “Printable Snowflakes Wedding Invitations”, “Silk Bride Bouquet Peony Flowers Pink Cream Lavender Shabby Chic Wedding Decor.”, “Wedding dresses and bridal gowns by David Tutera for Mon Cheri for every bride at an affordable price”, “Wedding Dress Style”, “Vintage Wedding Decorating Ideas”.

5.2.2 Statistics and topical analysis

From our dataset of over 58 million pins, the average number of pins per user was 444.86 (min = 0, max = 100,135). The average number of repins per pin was found to be 0.72 (min = 0, max = 20,212). Almost 79% pins in our dataset never got repinned. The average number of likes per pin was 0.21 (min = 0, max = 5,640). Also, 90.32% pins were not “liked” by anyone. This low percentage of repins and likes shows that there is a limited set of pins that get popular, and that a majority of pins go unnoticed. In case of comments, the results are even more skewed compared to pins. The average number of comments on a pin was 0.0065 (min = 0, max = 3,345), and 99.53% pins had no comments. This shows lack of utility of the comment feature on Pinterest.

To get an insight about the content of these comments, we randomly crawled 643,653 (1.1%) pins from our pin dataset, and were able to extract 2,544 comments. We then applied the Linguistic Inquiry and Word Count (LIWC) tool [34] on these comments, pin descriptions (Section 5.2.1), and user profile descriptions (Section 5.1.1). We found that a large portion of the comments reflected positive emotion (Figure 5). A similar pattern of positive emotion was observed for user description, as well as board names. In general, the network was found to have a large fraction of social content suggesting active human interaction. Presence of sad emotion, anger, anxiety, and swear words was found to be minimal. Textual content depicting biological processes, work, and leisure activities was also found in substantial quantity. From all this analysis, we conclude that a user usually leaves a positive remark for a pin on Pinterest, and posts positive textual content in general.

5.3 Source Analysis

Each pin has a source embedded in it. This source is the original URL of the image from where it is “pinned”. However, if the user has directly uploaded an image to Pinterest, the source field is set as “pinterest.com”. Table 2 shows that the top source for images on Pinterest is the users themselves, i.e. a large portion of images are directly uploaded and pinned by the users. Out of all the pins in our dataset, 2,768,851 pins (4.7%) were uploaded by users, second spot was taken by Google, which included images from Google

---

Table 1: Top 10 user profiles on Pinterest based on number of followers (as of January, 2013). The table also shows number of followees for users, and interests / profession as captured from the about field.

<table>
<thead>
<tr>
<th>Followers</th>
<th>Followees</th>
<th>Interests / Profession</th>
</tr>
</thead>
<tbody>
<tr>
<td>11,992,745</td>
<td>149</td>
<td>Designer / Blogger / Food</td>
</tr>
<tr>
<td>9,099,998</td>
<td>143</td>
<td>Designer / Magic / Food</td>
</tr>
<tr>
<td>8,056,723</td>
<td>1,176</td>
<td>Interior Designer</td>
</tr>
<tr>
<td>7,519,854</td>
<td>205</td>
<td>Not Mentioned</td>
</tr>
<tr>
<td>6,004,793</td>
<td>1,106</td>
<td>Lifestyle Blog</td>
</tr>
<tr>
<td>5,023,007</td>
<td>242</td>
<td>Beauty Enthusiast / Blogger</td>
</tr>
<tr>
<td>4,793,914</td>
<td>310</td>
<td>Architecture Student/Blogger</td>
</tr>
<tr>
<td>4,409,097</td>
<td>66</td>
<td>Not Mentioned</td>
</tr>
<tr>
<td>4,126,895</td>
<td>1,001</td>
<td>Artist</td>
</tr>
<tr>
<td>3,658,844</td>
<td>383</td>
<td>Freelancer / Blogger</td>
</tr>
</tbody>
</table>

---

5Example of an image source: www.cookingchanneltv.com/recipes/spanish-tortilla-recipe/index.html
Figure 4: (a) Followers / followees for the users on Pinterest, on a log scale. (b) The follower / followee ratio on Pinterest had no correlation with the ratio on Twitter. (c) Pin description on Pinterest. Similar to user descriptions, pin descriptions were also dominated by terms related to food and creative arts, and partially overlapped with terms present in user descriptions.

Image Search, and other Google products, followed by Etsy, at the third spot. Not surprisingly, free image sharing platforms dominated the top 10 sources. Six out of the top 10 sources on Pinterest were among the top 1,000 most visited websites in the world [1]. Etsy, a commercial website being ranked high, shows that a reasonable amount of user traffic on Pinterest comes from e-commerce websites, and depicts that commercial activity is widespread on Pinterest.

Table 2: Top 10 image sources on Pinterest. W.A.R. = Worldwide Alexa Rank. Apart from free image sharing / social network platforms, top sources include commercial platforms like Etsy.

<table>
<thead>
<tr>
<th>Source</th>
<th>Count</th>
<th>W.A.R.</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pinterest.com</td>
<td>2,768,851</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Google</td>
<td>1,293,749</td>
<td>1</td>
<td>Search engine</td>
</tr>
<tr>
<td>Etsy</td>
<td>1,157,815</td>
<td>164</td>
<td>Commercial</td>
</tr>
<tr>
<td>Flickr</td>
<td>625,686</td>
<td>70</td>
<td>Image sharing</td>
</tr>
<tr>
<td>Imgfave</td>
<td>376,179</td>
<td>9,462</td>
<td>Image sharing</td>
</tr>
<tr>
<td>Weheartit</td>
<td>306,443</td>
<td>970</td>
<td>Image sharing</td>
</tr>
<tr>
<td>Someecards</td>
<td>296,908</td>
<td>6,648</td>
<td>E-cards</td>
</tr>
<tr>
<td>Houzz</td>
<td>294,065</td>
<td>958</td>
<td>Home decor.</td>
</tr>
<tr>
<td>Martha Stewart</td>
<td>292,128</td>
<td>2,439</td>
<td>Food / Art</td>
</tr>
</tbody>
</table>

5.4 Pinboard analysis

In addition to the above Pin analysis, we also analyzed the names of Pinboards. The most common terms occurring in board names were home, style, recipes, food, wedding, crafts, etc. Pinterest also provides an option with 33 different predefined categories for board creation. We analyzed the popularity of all these categories based on 3 factors, number of boards in each category, number of pins on these boards, and number of followers of these boards under each category. We saw that 69.37% boards were created with no standard category selected. Apart from these, the top three categories for board creation were food_drink (5.6%) followed by diy_crafts (2.5%), and hair_beauty (2.4%). Followers of boards in the “travel” category outnumbered all the other boards by a big margin, and had the highest ratio of followers per pin (23.69 followers per pin). The next most famous boards in terms of followers per pin were education (10.34 followers per pin), health_fitness (5.37 followers per pin), and home_decor (4.71 followers per pin).

5.5 Location analysis

We investigated location information to find the Pinterest population distribution across the world. From our dataset, we collected 192,261 valid user locations, and performed a lookup using Yahoo PlaceFinder API. We inferred the top 10 countries in terms of number of users (Table 3) from Yahoo’s API output. Similar to Facebook and Twitter [23], a majority of Pinterest users also came from the U.S.A., Canada, U.K., Brazil, India, and Europe. We found minimal users from Africa, Russia, and China. Table 3 also lists Pinterest’s regional traffic ranks taken from Alexa, on 2nd June 2013. These ranks show that Pinterest is among the top most popular sites in countries like U.S.A., Canada, U.K., Australia, Brazil, India, etc., which are also the top user locations in our dataset. After analyzing country-wise distribution, we did a city level location analysis for these top 10 countries (Table 3), and found that most Pinterest users belonged to big metropolitan cities. More than half of the cities in top 20 were from the U.S.A. Pinterest’s penetration was found to be quite low in smaller cities.

As most Pinterest users in our dataset were females (Section 5.1.4), we analyzed gender distribution with respect to location. We observed that approximately 88% of users from the U.S.A. were females, and approximately 7% were males. A similar trend was observed in U.K., Australia, Europe, and Brazil (Table 3). India was the only country in the top
Table 3: Top 10 countries, and top 20 cities in decreasing order of Pinterest population. Apart from India, all other countries were dominated by female users. The penetration of Pinterest is maximum in big metropolitan cities. P.R.R.= Pinterest Regional Rank.

10. Netherlands 29 75.88 16.52
8. France 183 70.36 22.53
7. Italy 142 62.91 27.04
6. Spain 54 66.83 24.56
5. India 20 45.30 46.64
4. Australia 23 80.59 11.05
3. U.K. 38 72.79 18.47
2. Canada 21 82.73 10.66
1. U.S.A 15 83.88 8.80

1. New York 5597 11. Dallas 1275
2. London 3424 12. Austin 1249
3. Los Angeles 3194 13. San Diego 1213
4. Chicago 2593 14. Houston 1169
5. Toronto 1752 15. Sidney 1157
7. Atlanta 1472 17. Melbourne 1034

5.6 Gender Prediction

As mentioned in section 5.1.4, we collected gender information of over 1.85 million Pinterest users (150,945 males, and 1.61 million females) from Facebook, who had connected their Facebook accounts with their Pinterest profile. Considering this information as true, we attempted to predict gender of Pinterest users on the basis of profile, network, and content based features. For this experiment, we limited our analysis to users from USA only. §

5.6.1 Dataset and feature description

For gender prediction, our training dataset comprised of 6,309 male users, and 60,047 female users from the USA. To maintain a balance between the class sizes for applying machine learning, and achieve better confidence, we picked up six random samples of 6,309 female users from the 60,047 total female users, and calculated an average accuracy over all of them. A similar technique was used by Benevenuto et al. while classifying spam on Twitter using unbalanced training data samples [3]. We used a total of 9 features for classification, as listed below:

1. **Number of followers**: The number of users who follow a given user.
2. **Number of followees**: The number of users who, the given user follows.
3. **Number of pins**: The number of pins pinned by the given user.
4. **Number of boards**: The number of boards created by the given user.
5. **Content from “about” field**: We extracted the top 1000 most frequently occurring terms in the “about” field of male and female users’ profiles, and normalized these frequencies with the number of users in their respective categories (male / female). Each data instance was then assigned a male-female ratio score as follows:

\[
\text{About Ratio}_{m/f} = \frac{\sum_{i=1}^{n} W_i \times (NM_{W_i})}{\sum_{i=1}^{n} W_i \times (NF_{W_i})}
\]

where

- \(W_i = \text{Words in the about field}
- NM_{W_i} = \text{Normalized frequency for word } W_i \text{ for males}
- NF_{W_i} = \text{Normalized frequency for word } W_i \text{ for females}

6. **Board names**: Similar to the previous feature, we extracted the top 1000 most frequently occurring board names from male and female users’ profiles separately, and normalized these frequencies with the total number of users in their respective categories. Each data instance was then assigned a male-female ratio score as follows:

\[
\text{Board Desc}_{m/f} = \frac{\sum_{i=1}^{n} P_i \times (NM_{P_i})}{\sum_{i=1}^{n} P_i \times (NF_{P_i})}
\]

where

- \(P_i = \text{Individual pinboard in user’s set of pinboards}
- NM_{P_i} = \text{Normalized frequency for pinboard } P_i \text{ for males}
- NF_{P_i} = \text{Normalized frequency for pinboard } P_i \text{ for females}

7. **Presence of a linked Twitter account with profile**: True, if the Pinterest user has connected his / her Twitter account; false otherwise.
8. **Presence of personal website**: True, if the Pinterest user has mentioned a website in their website field; false otherwise.
9. **Name**: We got a list of the most common male and female first names in the US population during the 1990 census, 7 and assigned a ternary integer score to each data instance according to the user’s name being present in the list of males, females, or both / none.

The last feature is independent of the Pinterest network. We wanted to examine the performance of Pinterest-specific features for predicting gender, as compared to features based on only names; which is a completely independent feature in itself. We performed all classification tasks using WEKA [14].

§We picked USA, since it had the largest proportion of users in terms of country-wise user distribution. See table 3.

7http://names.mongabay.com/
5.6.2 Classification results

First, we attempted to predict gender using a feature set $F_8$ of only the first 8 features, i.e., features extracted from Pinterest. We applied 3 classifiers on our dataset of 12,618 users, and achieved a maximum average accuracy of 73.17% with 10-fold cross validation using the J48 Decision Tree classifier. To enhance the prediction accuracy, we introduced the Name feature $F_{name}$ to our feature-set. Note that this feature completely relies on the name of the user, and is independent of Pinterest. We were able to achieve a better accuracy of 86.18% with the addition of this feature, using the J48 Decision Tree classifier. However, using only the $F_{name}$ feature for classification, we still achieved an accuracy of 83.64%. Table 4 summarizes the results.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Feature set</th>
<th>Accuracy ($\sigma$)</th>
<th>F-Measure ($\sigma$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>$F_8$</td>
<td>62.96% (5.85)</td>
<td>0.586 (0.100)</td>
</tr>
<tr>
<td></td>
<td>$F_8+F_{name}$</td>
<td>86.18% (0.29)</td>
<td>0.861 (0.003)</td>
</tr>
<tr>
<td>J48 DT</td>
<td>$F_8$</td>
<td>73.17% (0.39)</td>
<td>0.732 (0.004)</td>
</tr>
<tr>
<td></td>
<td>$F_8+F_{name}$</td>
<td>83.64% (0.27)</td>
<td>0.834 (0.003)</td>
</tr>
<tr>
<td>RF</td>
<td>$F_8$</td>
<td>71.38% (0.41)</td>
<td>0.713 (0.004)</td>
</tr>
<tr>
<td></td>
<td>$F_8+F_{name}$</td>
<td>83.64% (0.27)</td>
<td>0.834 (0.003)</td>
</tr>
</tbody>
</table>

Table 4: Classification results for Naïve Bayesian, J48 Decision Tree, and Random Forest classifiers. The accuracy and weighted average F-measure scores are averaged over a labeled dataset of 6,309 male users, and 6 random samples of 6,309 instances each, from 60,047 female users.

From the six random samples of training data we picked for female users, the J48 Decision Tree classifier performed the best across all the samples individually. We achieved a maximum accuracy of 73.51% using $F_8$ (Pinterest-specific features), 86.53% using $F_8 + F_{name}$ (all 9 features), and 83.99% using only $F_{name}$; across all samples. Table 5 represents the confusion matrix for these results. As expected, $F_{name}$ was the most informative feature, followed by board names, number of pins, presence of personal website, number of boards, content from about field, and presence of linked Twitter account. The number of followers and followers were the least informative features. Rao et al. [36] achieved a similar score of 72.33% while predicting gender of Twitter users, with the help of a rich feature-set using a SVM classifier. Since the ratio of male to female users on Pinterest is highly skewed [32] as opposed to Twitter (which is fairly balanced $^8$), the size of our training dataset was limited. We believe that with this limited training data, our classification accuracy is reasonable.

\footnote{http://www.beevolve.com/twitter-statistics/#a1}

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Cls</th>
<th>TP</th>
<th>FP</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Meas</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_8$</td>
<td>F</td>
<td>0.842</td>
<td>0.112</td>
<td>0.883</td>
<td>0.842</td>
<td>0.862</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>0.898</td>
<td>0.155</td>
<td>0.839</td>
<td>0.888</td>
<td>0.808</td>
</tr>
<tr>
<td>$F_{name}$</td>
<td>F</td>
<td>0.76</td>
<td>0.24</td>
<td>0.748</td>
<td>0.71</td>
<td>0.738</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>0.71</td>
<td>0.24</td>
<td>0.738</td>
<td>0.71</td>
<td>0.728</td>
</tr>
</tbody>
</table>

Table 5: Confusion matrix representing true positive, false positive, precision, recall, and F-measure scores for J48 Decision Tree classifier in the best case. #F = Number of features; Cls = Class (Male / Female); TP = True Positive score; FP = False Positive score.

These results imply that even though gender is not a publicly accessible attribute on Pinterest, it is not difficult to predict gender using a small number of other publicly available attributes.

5.7 Privacy and security issues

To study privacy implications of the public nature of Pinterest, we attempted to extract email addresses and phone numbers from the publicly available user description field from users’ profiles. We found that a total of 9,926 users in our dataset shared their email addresses publicly. We then searched for phone numbers, which are widely considered to be PII [25], and found a total of 1,046 phone numbers and/or BBM pins from the users’ profile description field. Research shows that it is possible for third-parties to link PII, which is leaked via OSNs, with user actions both within OSN sites and elsewhere resulting in privacy leakage [24]. A recent study also investigated the risks of sharing phone numbers publicly on Facebook, and Twitter, and highlighted the extent to which these phone numbers could be exploited to gather much more private information about a user [18].

While various brands are using Pinterest for legitimate commercial purposes by promoting their work through pinboards, Pinterest has also attracted spammers and malicious users. With the growth in the number of users, there has been a simultaneous growth in the number of spammers on Pinterest. Numerous online scams have been reported in recent times [7, 9, 20, 35], and Pinterest has taken measures to solve this problem. To get a better understanding of the presence of spam and malware on Pinterest, we used Google’s Safe Browsing API to check for malicious source URLs on the network. We analyzed the source URLs of a random sample of 5.5 million pins from our pin dataset and found 1,322 (0.024%) unique malware pins. Despite numerous reported incidents of spam and malware, such low numbers suggest that the techniques deployed by Pinterest to avert malware are indeed effective. Since we collected these pins in January 2013, we wanted to check if the captured malware continued to exist on Pinterest. We then crawled these 1,322 pins again in May 2013 and observed that 33 of these pins no longer existed. It is hard to predict if the users themselves deleted these pins, or Pinterest removed it. We re-checked the source URLs of the 1,322 malware pins in May 2013, and found that 223 source URLs no longer ex-

\footnote{http://mashable.com/2012/12/06/pinterest-spam-accounts/}
\footnote{http://blog.pinterest.com/post/37347668045/fighting-spam}
\footnote{https://developers.google.com/safe-browsing/}
ist. Corresponding to the 1,322 malware pins, we identified 1,171 unique users from our dataset. Re-crawling these user accounts in May 2013 revealed that 100 out of these 1,171 user accounts did not exist. This shows that other than removing malicious content, Pinterest also take measures to remove malicious user profiles.

6. DISCUSSION

In this work, we characterized the Pinterest social network, and tried to predict it’s users’ gender using profile, content, and network based features. We collected 17,964,574 unique user handles, 3,323,054 complete user profiles, 777,748 boards with their corresponding details, and 58,896,156 unique pins with their related information, using Snowball sampling [13]. Our analysis was based on a partial subgraph of the Pinterest network, and suggests that Pinterest is a social network dominated by “fancy” topics like fashion, design, food, travel, love etc. across users, boards, and pins. A large part of the network was found to have a comparable number of followers and followees. Only a small fraction of people had large number of followers as compared to followees and vice-versa. The largest contributors of content (images) on Pinterest were the users themselves, with 2,768,851 (4.7%) users uploading original content; the remaining content (95.3%) was pinned from pre-existing web sources. Google Images, and Etsy followed as the next most famous sources, from where images are pinned onto Pinterest. USA, Canada, and UK contributed the maximum proportion of users, together accounting for over 73% of the total Pinterest population.

We then focused our analysis on predicting gender of Pinterest users based on their profile, content, and network features. Our labeled dataset consisted of a total of 12,618 user profiles from USA, with equal distribution, and we were able to achieve an accuracy of 73.17% using only Pinterest specific features. Addition of the “name” feature increased the accuracy to 86.18%. Using only the name feature, we were able to achieve an accuracy of 83%, which shows that adding Pinterest specific features helps very little in predicting gender of a USA Pinterest user.

Finally, we did some preliminary analysis to explore the privacy and security implications associated with Pinterest, and found multiple instances of publicly available PII leakage due to the all-public nature of Pinterest. We also found presence of malware, and discovered that most of this malware continued to exist for at least 4 months; between our two crawls of the network. Given that Pinterest is fairly new in the social media fraternity, we suspect that the amount of malware would only grow in the near future.

We picked the initial seeds for our data collection process as the top 5 most followed users on Pinterest. We understand that this technique suffers from bias, and the sample taken is not completely random. We crawled only partial sub-graphs for all the 5 seed users. Similarly, on the next level of our BFS crawl, we crawled not more than 48 followees for each user. Since there is not much prior work on Pinterest, we do not have enough academic literature to claim that our dataset is representative of the whole Pinterest population. However, the previous work by Ottoni et al. [32], Gilbert et al. [12], and a report from Engauge, a digital marketing agency [11], show similar gender distributions for users, and similar topic distributions for boards and pins as our dataset.

In future, we would like to perform a deeper analysis for gender prediction on Pinterest. Our current feature set of 9 features can be expanded to accommodate features based on natural language, content, profile attributes, and network features. Users’ about field, and comments can be utilized for this purpose. More network based features like betweenness, and closeness centrality can also be explored. We would also like to generalize gender prediction over the entire geographic Pinterest population, rather than limiting it to users from USA only.

To the best of our knowledge, this is one of the first attempts to characterize Pinterest, and study its various components in depth, on such a large scale. For this analysis, we use profile information for only about 3.3 million users from over 17 million unique user handles that we had in our dataset. Our data collection process is still active, and we would like to redo our analysis on the largest connected component (LCC) of the complete Pinterest network. We would also like to perform a more detailed analysis of image-spam, and copyright violations on this network. Given that Pinterest has been the fastest growing social network in recent times, it would be interesting to see if malicious users are targeting Pinterest for spiteful purposes.

7. REFERENCES


I. Lunden. There are now over 1 billion users of social media worldwide, most on mobile. http://techcrunch.com/2012/05/14/it-there-are-now-over-1-billion-users-of-social-media-worldwide-most-on-mobile/, 2012.


