Understanding Online Collection Growth Over Time: A Case Study of Pinterest

Caroline Lo*
Stanford University
clo@cs.stanford.edu

Justin Cheng*
Stanford University
jcccf@cs.stanford.edu

Jure Leskovec
Pinterest
jure@pinerest.com

ABSTRACT

A common feature of content discovery applications is the ability of users to save and organize digital items into collections, *e.g.*, images into photo albums or songs into playlists. Understanding how these collections grow over time is important for user retention and collection as well as item recommendation. Here we study factors that affect collection growth over a long period of time. We conduct a large-scale longitudinal analysis of over 2.6 million collections, known as boards, on Pinterest, over a period of three years. We study the inter-event time distribution of pins saved to boards and find that it can be accurately described by a two-component lognormal mixture model. The mixture components reveal that board growth can be characterized by short-term fast-paced sprees of activity, and longer breaks between these sprees. Commonalities emerge in spree behavior; for example, sprees have consistent temporal dynamics and the content saved within the same spree is more focused compared to between sprees. Surprisingly, we observe that boards with longer initial sprees are less likely to have long-term growth. On the other hand, boards with more frequently occurring sprees continue growing for a longer time, and tend to have a larger size. Finally, we synthesize our findings into a series of predictive models which show that initial board evolution is a strong signal for long-term board growth in terms of size and lifespan. Overall, our research has important implications for the design of online content discovery applications and has immediate applications in user modeling and recommendation systems.

1. INTRODUCTION

An important aspect of the social web experience is the ability of users to create and curate collections of items [32]. People can use online platforms to collect not only digital representations of physical objects that they own or wish to own, but also digital representations of concepts (*e.g.*, quotations, poems), digital media (*e.g.*, books, pictures, movies, songs, videos), or even ideas (*e.g.*, recipes, to-do lists) [8]. Many content discovery platforms allow users to form collections ranging from digital playlists of songs on Spotify, to videos and movies on Youtube or IMDB, to real life products on Amazon or Etsy.

A unique aspect of these collections of digital items is that they remain useful and updatable over long periods of time [27]. For example, a Spotify playlist is meant to be enjoyed repeatedly and grown over time, and avid television fans would be hard pressed to empty their ever-evolving Netflix watchlists. The identification of collections that will grow consistently over a long period of time is therefore vital for collection recommendation in content discovery applications, because active and growing collections provide better value and lead to higher user engagement. Furthermore, an understanding of early signals that are indicative of long-term collection growth can inform user retention and application design. Specialized user interface flows and interaction techniques can be developed which enable users to experience more efficient content curation and exploration.

However, collection growth is currently not well understood. Prior work either treats a collection as a non-changing entity [19, 9, 22], or focuses on individual items in a collection [4, 18]. The temporal dynamics of collection growth has not yet been analyzed. What is missing from the picture are models of collection growth. Developing such models is important for understanding collection growth dynamics, predicting future growth, and identifying collections that will remain active for a long period of time.

Present work: Understanding collection growth over time. Here, we conduct a large-scale study of online collection growth over time, and analyze the process by which users save items into collections. We build on existing research on inter-event times to characterize the growth behavior of collections. We then analyze how initial collection growth, as defined by this characterization, can be used to predict future growth behavior.

In particular, we analyze the growth of collections on Pinterest [11], a content discovery application, over time. On Pinterest, users engage with visual bookmarks, called *pins*, and save them into collections of pins, called *boards*. We analyze over 2.6 million boards created over a one week period, and track each board's growth for exactly three years.

We first analyze the inter-event time distribution of pins added to boards, and find that this distribution can be accurately described by a two-component lognormal mixture model. After confirming that these two components exist on a per-user per-board basis, we find that the components of the mixture model correspond to (1) short-term *pinning sprees* with small inter-event times, and (2) longer *breaks* between these sprees.

We uncover patterns both within and between sprees. We discover that inter-event times between pins quicken towards the beginning of a spree, and slow down towards the end. Furthermore,

*Research partly done while at Pinterest.

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054189

we find that content within the same pinning spree is noticeably more focused and similar than content between different sprees.

Analyzing dynamics inside sprees, we find that initial growth is indicative of how collections will grow in the future in terms of final size and longevity. In general, factors that increase final board size also increase longevity. For example, we find that shorter breaks between initial sprees lead to larger and longer-lived boards, and that higher initial content similarity leads to smaller and more short-lived boards. However, while larger initial spree sizes indicate larger boards over time, they also surprisingly lead to substantially shorter-lived boards.

Finally, we connect our insights on initial collection growth together in a series of long-term and short-term growth prediction tasks. We develop a model that can predict whether a board of a given size will grow past its expected median size, with prediction performance improving as the board gets larger (ROC AUC=0.83). We are also able to identify boards that maintain a regular level of activity across an entire year (ROC AUC=0.88), and predict when a pinning spree will continue (ROC AUC=0.75). We find that different signals are important to different tasks; for example, while the average break time between sprees is extremely important in identifying boards with a regular level of activity, it is substantially less important when trying to identify spree continuations.

Overall, our work provides a first look at how collections grow over time in content discovery applications. We introduce a general method for analyzing the dynamics of collection growth, by fitting a mixture model to the inter-event time distribution, and show how this method can be used to quantify the relationship between a collection's initial and long-term growth. This approach can be used to better understand how collections evolve over time–in the case of Pinterest, we further show that a collection's initial growth is predictive of how it will grow in the future.

2. RELATED WORK

Next we briefly review related work on understanding the composition of online collections, modeling inter-event time distributions, and the evolutionary dynamics in the context of online networks and social computing applications.

Online collections. There exists a long and varied history of research on online collections on content discovery platforms. Studies have investigated the impact and usefulness of collections on sites such as del.icio.us, Twitter, and Netflix [29, 26, 23], as well as the motivation for creating collections on sites such as Flickr, museum websites, Youtube, and online shopping carts [24, 19, 8, 6]. Works have also focused on item recommendation to facilitate the growth of collections such as music playlists, online shopping carts, and digital bookmarks [4, 18, 20]. Another area of research has studied how to recommend collections themselves, *e.g.* Twitter user list recommendation [9, 22]. Rather than focusing on collections as an unchanging unit or on the specific items that are added to collections as prior work does, here we instead focus on the *temporal* aspect of collections. We treat a collection as an entity that grows and evolves over time, and analyze factors that characterize this growth over a long period of time.

Inter-event times. A large body of work focuses on characterizing inter-event time distributions for online services. Prior work models the inter-event distribution of email and web browsing [25, 2], as well as Netflix and and eBay usage [33]. While most studies observe inter-event times that follow a power law distribution [12, 34], there are also several studies that observe a two-component or three-component mixture when observing the inter-event distribution of platforms such as Wikipedia and del.icio.us [10, 28]. Here we build on this prior work to characterize the inter-event distribution of collection growth and find that it follows a two-component lognormal mixture model. In contrast to previous work, we show that the two-component lognormal mixture emerges not as a result of population averaging or resource constraints but due to a user's behavior on an individual collection. We build upon these insights, and relate them to long-term collection growth.

Online evolution. A variety of research analyzes online evolution over time. Works examine topics ranging from user changes in browsing behavior [15] to user group evolution [1, 21] to social network evolution in aggregate [17]. In recent years studies have focused on signals for size prediction of user groups [31, 13] and information cascades on Facebook and Twitter [5, 16]. Other research investigates signals for the lifespan prediction of user groups [13], as well as user participation in online communities [30, 7]. We add to this body of work by introducing collections as a concept whose evolution can be analyzed. We extend techniques and methodologies discussed in these works to study collection growth and evolution over time, and formulate prediction tasks that contribute insights in terms of collection size and longevity.

3. DATASET DESCRIPTION

Our analysis focuses on collections of items formed on Pinterest, a content discovery application. On Pinterest, users can browse and explore images–referred to as *pins*. These pins can represent real-life or abstract concepts. Pins that users like can be collected into *boards*, which are then displayed on a user's profile. A wide range of boards exist, ranging from collections of dinner recipes to inspirational quotes to photos of dream vacation destinations. The concept of pins and boards is unique to Pinterest, but these concepts are analogous to items and collections found on other content discovery platforms, for example, songs and playlists on Spotify, or products and wishlists on Amazon.

Here, we analyze boards on Pinterest that were created by non-spammers during the week of April 19-25, 2013. We track three years' (1095 days) worth of pinning data after board creation for each board, and include in our analysis boards that grew to at least 10 pins large during the observed time period. In total, our dataset contains 2.6 million boards created by 2.4 million unique users. These boards contain 282 million pins collectively, and 107 pins per board on average.

4. INTER-EVENT TIME DISTRIBUTION

To understand the dynamics that govern collection growth, it is important to understand the inter-event times of items that are added to collections. Here we develop a model that accurately describes the inter-event time distribution of pins added to boards in aggregate. We then infer board growth behaviors based on the results of this model.

4.1 Fitting the inter-event distribution

The first step is to examine the inter-event times of pins added to a board, using a time resolution of seconds. We do so by analyzing the probability distribution function (PDF) of inter-event times across all boards, as shown in Figure 1. Two peaks are noticeably observable in the PDF, one peaking at approximately 10^1 and the other at approximately 10^5 seconds.

The inter-event distribution follows a two-component lognormal mixture model. We find that the two observed peaks, both of which follow a lognormal distribution, are best described as a two-component lognormal mixture model (2-LMM). This mixture

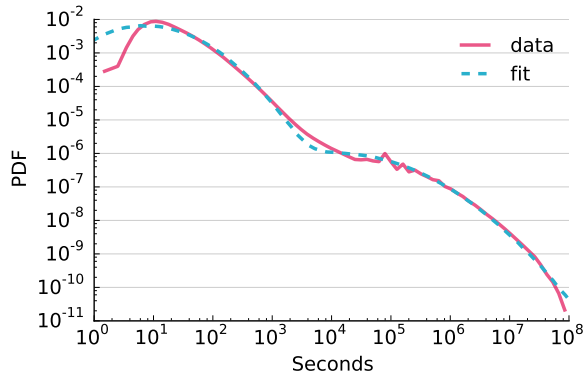


Figure 1: The inter-event distribution of pins added to boards (PDF). This distribution can be closely approximated by a two-component lognormal mixture model. Coefficients are displayed in Table 1.

model is parameterized by weights ϕ_1 and ϕ_2, which represent the likelihood that any given inter-event time sample is drawn from lognormal distribution $\ln\mathcal{N}(\mu_1, \sigma_1)$ or $\ln\mathcal{N}(\mu_2, \sigma_2)$, respectively. The PDF of a lognormal distribution is defined as

$$\ln\mathcal{N}(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right]$$

for $x > 0$.

Fitting the model. We use expectation maximization [3] to determine the parameters of our mixture model, as follows. For each inter-event time sample x_i and lognormal component $j \in \{0, 1\}$, we define weight $w_j^{(i)}$ to be the probability that inter-event time x_i is sampled from the j^{th} lognormal distribution. In our E-step, we update each weight as follows:

$$w_j^{(i)} := \frac{\ln\mathcal{N}(x_i;\mu_j,\sigma_j)}{\sum_{k=0}^{1}\ln\mathcal{N}(x_i;\mu_k,\sigma_k)} \quad (1)$$

In our M-step, we update the main parameters of our mixture model as follows (m is the number of inter-event time samples):

$$\phi_j := \frac{1}{m}\sum_{i=1}^{m} w_j^{(i)}$$

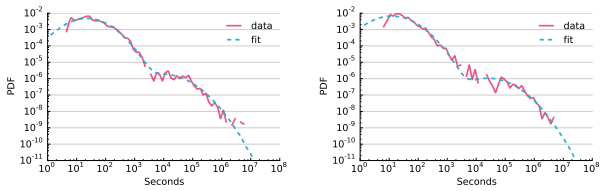
$$\mu_j := \frac{\sum_{i=1}^{m} w_j^{(i)} \ln x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}$$

$$\sigma_j := \frac{\sum_{i=1}^{m} w_j^{(i)} (\ln x^{(i)} - \mu_j)^2}{\sum_{i=1}^{m} w_j^{(i)}}$$

The fitted 2-LMM is shown in Figure 1. We define the *lower* lognormal component to be the one that peaks at a smaller value, which has a median inter-event time of 1.16 minutes. Likewise, the *upper* lognormal component peaks at a larger value, and has a median of 7.76 days. Parameters of the model are shown in Table 1. We find that the resulting fit is quite accurate; the Kolmogorov-Smirnov (KS) distance between the the observed inter-event time distribution and our fitted distribution is only 0.025 as opposed to 0.162 when compared to a fit using only a single lognormal.

Component	ϕ	μ	σ	median
Lower	0.55	4.24	1.50	1.16 minutes
Upper	0.45	13.43	2.03	7.76 days

Table 1: Parameters of the 2-LMM, fitted to the board inter-event time distribution, as well as the median value of each lognormal component. ϕ is the sample weight of each component, and lognormal parameters μ and σ are the location and scale parameters for each component.



(a) (b)

Figure 2: Explaining the origin of inter-event time distribution. (a) The inter-event time distribution of a sampled single board. (b) The inter-event time distribution for a user who has only ever created only one board. We conclude the two lognormal components exist on a per-board per-user basis and represent two modes of growth behavior.

4.2 Explaining the inter-event distribution

While we are able to accurately model the inter-event time distribution, one limitation is that we cannot immediately infer which populations generate the two observed components. Are the two lognormal distributions generated from different sets of users? Or, from different boards among the same users? To answer these questions and determine the origin of the two distributions, we conduct additional analyses.

The 2-LMM is observable in individual users. One potential explanation for the observed behavior in Figure 1 is that the two lognormal distributions are generated from two disjoint groups of users who behave differently: an active set of users whose boards contain shorter inter-event times drawn from the lower lognormal, and a less active set of users whose boards have longer inter-event times drawn from the upper lognormal. If this hypothesis is true, then a single board's inter-event time should follow either a lower or an upper lognormal distribution, because each board in our dataset only has one user saving pins to it. However, the inter-event time distribution of individual boards often contains two lognormal components; an example board's distribution is shown in Figure 2(a). Although the inter-event data is sparse, especially in boards with a small number of pins, we observe a clear 2-LMM even when we consider an individual board; the average KS distance for individual boards is 0.072 (0.0008 standard error) for a 2-LMM, versus 0.215 (0.001 standard error) for a single lognormal. The experiment demonstrates that it is not the case that a combination of two different classes of users, each with a different single lognormal inter-event time distribution, generated the resulting 2-LMM.

The 2-LMM is observable for users who only own one board on Pinterest. The second hypothesis is that the lognormal distributions occur due to "resource competition" between multiple boards owned by the same user; *i.e.*, inter-event times fall under the lower lognormal distribution when a user is focusing on a given board, and fall under the upper lognormal distribution when a user is fo-

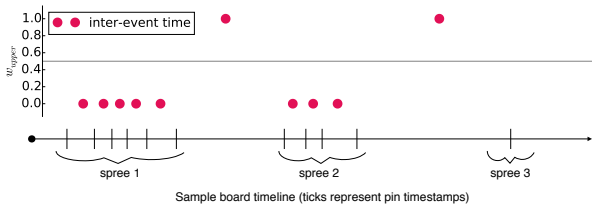


Figure 3: Sample board timeline from our dataset. Each tick represents a pin timestamp. We compute probability $w_{upper}(t)$ that inter-event time t is drawn from the upper lognormal distribution; $w_{upper}(t) > 0.5$ denotes a break between pinning sprees. Note that $w_{upper}(t)$ frequently tends towards 0 or 1.

cusing on other boards. However, we also observe a 2-LMM inter-event time distribution for users who have only ever created a single board on Pinterest, from the account creation time to the present date. An example is shown in Figure 2(b); the average KS distance for users with only one board is 0.113 (0.002 standard error) for a 2-LMM, versus 0.241 (0.006 standard error) for a single lognormal. Because there is no opportunity for resource competition if a user only has one board, we can conclude that individual boards' inter-event time distributions can be described by a 2-LMM, regardless of whether the board is one of many that a user owns.

4.3 Two types of board growth behavior

Our analyses above revealed that the two-component mixture model we observe describes board growth behavior on an individual basis. This means that the observed behavior is a fundamental property of individual user's behavior. The two-component nature of the inter-event time distribution reflects two types of behavior that board growth can be characterized by: short-term pinning sprees, and long-term breaks. A board experiencing a short-term pinning spree will have inter-event time(s) that fall under the lower lognormal distribution. A board experiencing a longer break will have an inter-event time that falls under the upper lognormal distribution, with a median of approximately one week.

Delineating sprees within a board. By defining longer breaks between pinning sprees to be inter-event times drawn from the upper lognormal distribution, we can neatly delineate when sprees begin and end within a board. Without loss of generality let the upper lognormal component be the $j = 1^{st}$ component defined in the expectation maximization algorithm. Then for any inter-event time t we define

$$w_{upper}(t) = \frac{\ln \mathcal{N}(t; \mu_1, \sigma_1)}{\sum_{k=0}^{1} \ln \mathcal{N}(t; \mu_k, \sigma_k)},$$

which is essentially the same as Equation 1, the update equation for our E-step.

If $w_{upper}(t) > 0.5$ for inter-event time t, then t is more likely to be drawn from the upper lognormal distribution and we mark the inter-event time as a break between pinning sprees. Concretely, in our dataset the threshold at which $w_{upper}(t) > 0.5$ lies at 67.8 minutes, or just over an hour. Thus events spaced less than 1h apart belong to the same pinning spree. Figure 3 displays a timeline of real board growth, with corresponding w_{upper} values. Here, $w_{upper} > 0.5$ for two inter-event times, and these times mark the boundaries between three individual pinning sprees. Notice that inter-event times clearly follow one of the two components. In other words, w_{upper} tends to take extreme values (0 or 1).

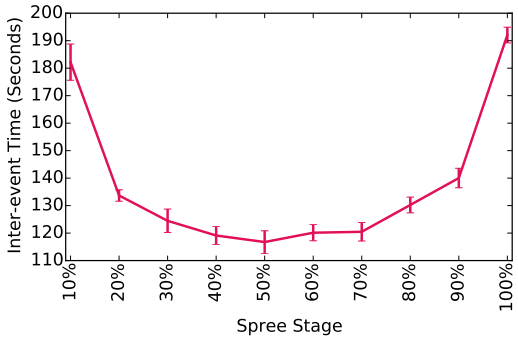


Figure 4: Average inter-event time at each spree life-stage. Inter-event times speed up in the beginning of a spree and slow down towards the end. Throughout this paper, error bars indicate standard error estimated by bootstrap resampling [14].

In summary, in this section we find that the inter-event time between pins that are saved to a board closely follows a mixture of two lognormal distributions. This means that when users are in a pinning spree they save a pin to a board every few minutes (median 1.16 minutes) and that breaks between pinning sessions last about 7 days (median 7.76 days).

5. UNDERSTANDING PINNING SPREES

Motivated by our analysis of collection inter-event times, in this section we study potential patterns that exist within and between sprees. Finding these patterns allows us to better understand why sprees occur and evaluate the validity of our spree segmentation approach. We focus on inter-event time changes throughout a spree, and on content similarity.

Pinning speeds up in the beginning of a pinning spree and slows down towards the end. One question to consider is whether a collection's growth speed remains constant throughout a single spree. Are there markers in terms of speed that suggest a spree is ending? To answer this question we examine the average inter-event time across various points in a pinning spree. As different sprees have different lengths, we define the life-stage of a given spree to be the percentage of pins added thus far out of the total number of pins that will be added in that spree. Thus a life-stage of 50% represents the half-way point of a spree in terms of number of pins.

In Figure 4 we plot the average inter-event times between pins at different life-stages, across all sprees of at least length 10. Users pin with increasing frequency (approximately 30%) in the beginning of a spree, maintain constant speeds in the middle, and then slow down back to the original starting speed towards the end. Interestingly, we observe that the eventual slowdown is noticeably more gradual than the initial speedup. A hypothesis is that as a spree progresses, users run out of content to add to boards very gradually as opposed to all of a sudden, suggesting that the success of collection growth may be dependent on how easily users are able to find content that they like.

Content is more similar within the same spree. To understand the relationship between content and a collection's sprees, we next examine *content similarity*. We define content similarity between pairs of pins to be the Jaccard similarity between each pins' annotated word vectors, which are generated based on the title and the description of the pin. To standardize across boards of different size, content similarity *within* pinning sprees and *between* pinning

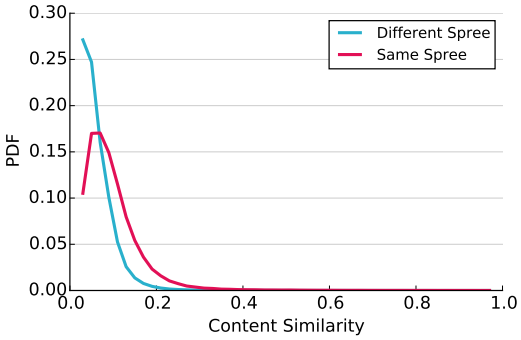


Figure 5: The PDF of content similarity for content both within the same pinning spree, as well as between sprees. Content is more similar within the same spree.

sprees is defined as the average similarity of 100 pairs of pins, sampled with replacement, within and between sprees, respectively.

We plot the PDF of content similarity within sprees and between sprees for each board with at least two sprees, shown in Figure 5. Content is noticeably more similar within the same spree compared to between sprees. In fact, we find that this is the case for a surprising 83.7% of all boards in our dataset. The average content similarity within sprees is 0.096 on average, over 50% larger than the average content similarity between sprees (0.058).

Overall, our findings in this section reveal commonalities that exist in spree-based behavior, validating our spree segmentation strategy for collections. We find that as a whole, inter-event times within a spree follow a predictable pattern, and that content is more focused within sprees compared to between sprees. These findings have applications in both user activity session analysis, as well as in content recommendation for collections.

6. RELATING SPREES TO BOARD GROWTH

Given the consistent patterns observed in the previous section, we next aim to discover whether the way that initial sprees in a collection grow might impact growth in the long term. Here, we specifically analyze two aspects of long-term growth for boards on Pinterest: final board size after three years of observation, and *effective lifespan*.

Motivating and defining effective lifespan. In terms of recommending collections to users, it is useful to surface collections that contain a regular stream of content. A board where one pin is added every month for a year, for example, should be considered much more active than a board that only added six pins in the beginning and six at the end of the year. To capture this sense of activity, given some inactivity threshold i, we define *effective lifespan* to be the longest period of time, starting from a board's creation, such that the longest inter-event time between pins is at most i days. In this section we set $i = 30$ days for the purpose of standardizing our effective lifespan analysis. We explore the effect of changing i in a later section.

Boards with larger initial spree sizes grow larger. Will a collection that grows quickly in the beginning continue doing so, leading to a bigger collection? To answer this question we examine the average initial pinning spree sizes of boards, by observing the first 30 days of pinning behavior for boards with at least 10 pins. The average final board sizes across various initial spree sizes are shown in Figure 6(a).

Boards with larger initial spree sizes indeed end up with a larger final board size as well; for example, a board with an average initial spree size of 10 will grow to be over 1.5 times as large on average than a board with only one pin added per spree. The difference may not be overly dramatic but the overall relationship is clear: faster early collection growth is a signal that the collection will grow larger after a long period of time.

Boards with larger initial spree sizes have shorter effective lifespans. Does fast initial collection growth translate to longer-lived collections as well? We next compare the average initial pinning spree size for each board, as defined earlier, with their final effective lifespans, with results shown in Figure 6(b). Surprisingly, it appears that boards that grow quickly (i.e. have larger initial average spree sizes) will "die" quickly as well; a board with six pins on average per spree lives only half as long as a board with one pin per spree. We call this phenomenon the "tortoise and the hare" effect, because slow and steadily growing boards ultimately win the race in terms of having higher effective lifespan.

One hypothesis for the tortoise and the hare effect is that on a content discovery application there is a finite amount of content that is relevant to any given collection, and users who grow their collections too rapidly exhaust their content options quickly. Another hypothesis is that collections that grow quickly or slowly serve different fundamental purposes for users.

To test the above hypotheses we compare the average initial spree size across different self-reported board categories, shown in Figure 7. We observe evidence that the second hypothesis might be true, as average initial spree size varies by a factor of up to 3.5 between different categories. Many categories with easily actionable purposes, such as "DIY & Crafts" and "Health & Fitness", have smaller pinning spree sizes than categories that have more abstract purposes, such as "History", or "Science & Nature". This isn't always the case, however, as seen with the "Quotes" category, which has a relatively small initial spree size.

Overall, however, while category may be a moderating factor, the patterns we observe earlier persist even when we analyze only boards from the "Food & Drink" category (plot not shown), one of the largest categories on Pinterest. This suggests that factors beyond category are at play when examining the relationship between spree size and final collection growth.

Boards with shorter breaks grow larger and live longer. We next wish to examine the relationship between how often a collection has a spree, and how the collection will ultimately grow. We first examine final board size by plotting the length of the first long break between two pinning sprees of a board, in days, versus the board's final board size, as shown in Figure 6(c). Overall, boards with a shorter break between sprees tend to grow much larger after three years. A likely explanation for this behavior is that shorter breaks over the same period of time means more pinning sprees, and therefore more opportunity to grow.

Interestingly, however, boards whose first two pinning sprees take place on the same day (i.e. the break is zero days) end up noticeably smaller on average than boards who instead return after a few days. Our hypothesis is that users of boards that return very quickly haven't yet "proven" that they will remain interested in the board in upcoming days and weeks, whereas those that return after a few days have proven that their interest at the very least spans several days. This is a concept that would be fascinating to explore in future work.

Similarly, we find that boards with shorter breaks live longer. By comparing the length of the first long break between pinning sprees with a board's effective lifespan (Figure 6(d)), we find that while

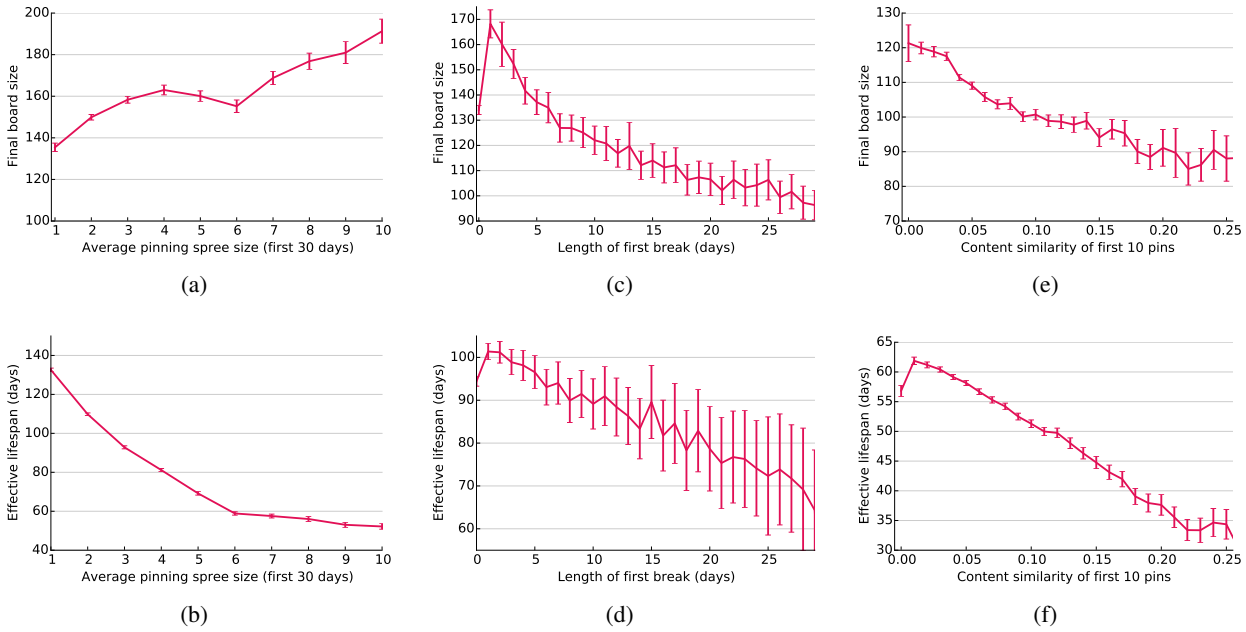


Figure 6: (a) Boards that have larger initial sprees have a larger final size. (b) Boards that have larger initial pinning sprees have shorter lifespans. (c) Boards that have a shorter first break have a larger final size. (d) Boards that have a shorter first break have longer lifespans. (e) Boards whose content is more similar to each other have a smaller final size. (f) Boards whose content is more similar to each other have shorter lifespans.

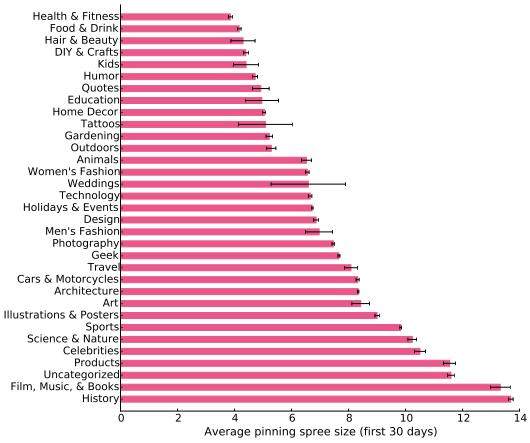


Figure 7: Categories impact average initial spree size. Broadly speaking, more actionable categories tend to have shorter initial spree sizes.

the anomaly at zero days persists, effective lifespan lessens by almost a third on average as the break length increases from one day to a month. Our findings suggest that short breaks between pinning sprees indicate board owners who have more long-term enthusiasm for their boards. Put another way, the same enthusiasm that drives users to grow their collections more frequently might make it less likely that their interest will wane over time, leading to longer effective lifespans.

Boards with dissimilar content grow larger and live longer. In order to gain insight into the relationship between the content of a collection and its size, we compare the the content similarity of the first 10 pins of a board's first pinning spree with the board's size after three years. We consider boards with content similarity between 0 and 0.25, which encompasses over 98.5% of all boards. To control for confounding variables, we also limit our analysis to boards whose first sprees are at least 10 pins long.

Figure 6(e) shows that boards that have higher initial content similarity tend to be noticeably smaller after three years of growth compared to boards with lower content similarity. For example, we see that boards with no initial content similarity grow to 120 pins on average, which is 33% larger than boards with an initial content similarity of 0.25. One explanation for this phenomenon may be that there is simply more potential content available for a collection with more diverse interests.

Similarly, we also find that boards with dissimilar content tend to live longer. Plotting the the board's effective lifespan as a function of the initial content similarity of a board (Figure 6(f)), we observe that the effective lifespan is up to twice as long when initial content is more diverse. One potential explanation for this behavior is that users "run out" of content more slowly when the collection they are growing is more diverse. Because content is less niche and easier to find, users with diverse collections might therefore find it more enjoyable to continue growing their collections.

In summary, our analysis in this section shows that initial collection growth patterns strongly impact future collection growth. In particular, we find that initial spree sizes, length of breaks, and content similarity are all useful signals when trying to measure final collection size and longevity. Our findings suggest that initial growth signals can be useful in prediction tasks, a concept that we explore in the next section.

7. PREDICTING BOARD GROWTH

Our analysis thus far points to signals in initial collection growth that indicate differences in final collection growth. To better un-

derstand the dynamics of how these signals relate to each other, we build on insights developed in Sections 5 and 6 to formulate a series of prediction tasks relating to final collection size, effective lifespan, and spree continuation.

Here, we are interested in observing which signals are most important for which tasks, as well as understanding settings under which our prediction tasks perform better.

Features and Model. Any sample in our prediction tasks consists of a subset of pins that have been added to a board, and contains at least 10 pins. Based on our findings in previous sections, we consider four broad classes of features:

- **General spree-based features:** Motivated by our findings in Section 6, we include the average observed spree size of our sample, as well as the average break time between sprees. We also include the fraction of inter-event times observed that fall under the upper lognormal distribution discussed in Section 4, representing a long break.
- **Pin-based features:** We include the raw number of pins we have observed thus far. We also include the content similarity of the first 10 pins in our sample, as described in Section 6.
- **Time-dependent features:** We include the timestamp, relative to the board's creation date, of the last observed pin in our sample, as well as the timestamp, if the sample contains n pins, of the $n/2^{th}$ pin, representing the halfway point.
- **Static features:** We consider demographic features of a user, *e.g.*, gender and location, as well as the category of the board.

All features are standardized. For each of our prediction tasks, we use a logistic regression classifier and perform ten-fold cross validation. To evaluate performance we compute the area under the ROC curve (ROC AUC).

7.1 Predicting final board size

The ability to identify collections that will grow larger over time is useful both in terms of collection recommendation as well as in identifying long-term engaged users. In our first task we aim to predict whether a collection of any given size will continue growing substantially in the future. Inspired by methodology introduced by [5], we predict, after having observed the first k pins, whether a given board will continue growing to its median expected size, given the fact that the board has at least k pins.

The precise setup is as follows: for a given k, we consider the set of boards with at least k pins. We compute the median final board size $med(k)$ after three years of observation, and predict, based on observing only k pins, whether the the final board size will be at least $med(k)$. The dataset is, by nature of medians, balanced. Results of our prediction task are shown in Figure 8. Overall performance is robust, with ROC AUCs ranging from 0.70 to 0.83. Interestingly, performance increases as a function of k, meaning it is easier to predict whether a board with 100 observed pins will reach its expected median size than it is to predict whether a board with 20 pins will do likewise. This suggests that in terms of identifying boards that will grow larger, it is most promising to consider more mature collections that have already reached a larger size.

Feature analysis. We now examine the relative importance of the features in our model by computing the Pearson/point-biserial correlation coefficient between each non-categorical feature and a binary variable indicating whether board size $\geq med(k)$, across different median board size k. Results are shown in Figure 9(a). Note that here, there are no correlation coefficients displayed for the feature *"# pins observed,"* because number of pins observed is always k for this prediction problem, and the correlation coefficient is therefore undefined.

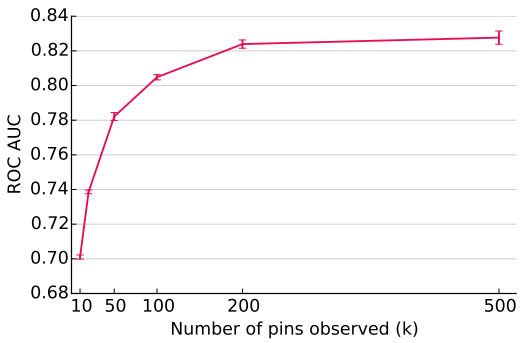


Figure 8: If we observe the first k pins of a board and want to predict if the board will reach its median expected size given that it has at least k pins, prediction performance improves as k grows.

Observe that across all k, *"timestamp: last pin"*, which serves as a proxy for how much time has elapsed so far since board creation, is strongly anti-correlated with growth past the median size (-0.322, $k = 10$). This suggests that younger boards tend to have more growth potential than older boards. On the other hand, *"content similarity"* is in comparison uncorrelated with whether board size $\geq med(k)$, particularly for large k (0.003, $k = 200$). This suggests that while initial content similarity may be related to final board size in general, as previously observed in Figure 6(e), once we control for the fact that a given board will become very large (k), content similarity is no longer as useful a signal.

From the perpective of changing k, we observe that the absolute value of correlation coefficients generally decreases as k increases, reflecting our findings that prediction performance improves as k grows. A notable exception is the *"avg time between sprees"* feature; while the correlation coefficient becomes more negative in the beginning while k grows, it then begins to steadily trend closer to zero. This surprising trend suggests that if a board manages to grow to a very large size k, it becomes increasingly less important how regularly sprees occur, while other factors become more important.

What is interesting when considering this and subsequent models is that across all our prediction tasks, correlation coefficients never reverse signs, *i.e.* features with positive correlation never become anti-correlated as k changes, and vice versa. We conclude that importantly, while the signal strength of a feature may vary as k varies, the underlying dynamics that govern the relationship between the feature and properties of the final board don't change.

7.2 Identifying boards with long lifespans

In terms of long-term collection recommendation, perhaps even more important than identifying future large collections is identifying collections that will grow consistently over a long period of time. These "active" collections are ideal in content discovery applications because they can regularly deliver new content to users who follow them. In Section 6 we set the maximum threshold of inactivity for effective lifespan to be 30 days, but in a practical settings different thresholds may be required for different purposes. Here we aim to find boards that are active for a very long period of time, for varying thresholds. Our prediction task is thus the following: given an inactivity threshold of k days, which boards have an effective lifespan of at least one year?

For various thresholds k ranging from 10 to 90 days, we sample an equivalent number of boards who grow at least once every k

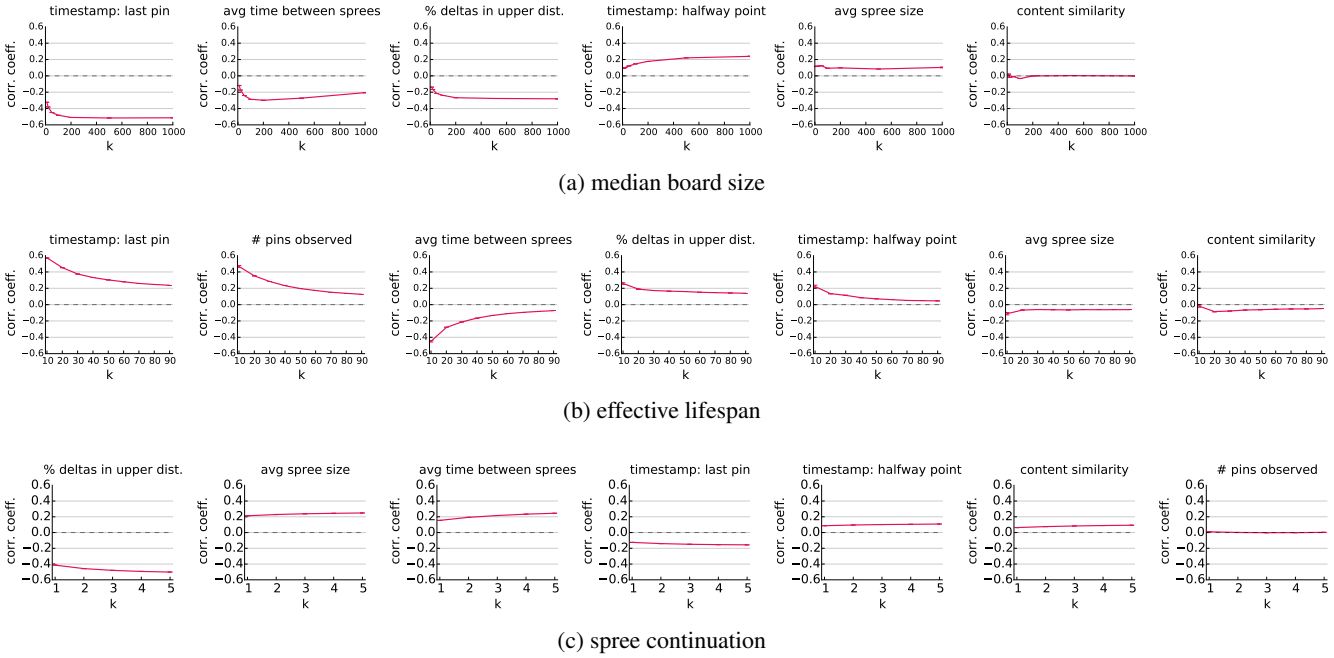


Figure 9: Pearson correlation coefficients between non-categorical features and binary variables indicating (a) whether board size $\geq k$, (b) whether a board's effective lifespan is longer than a year given an inactivity threshold of k, and (c) whether the next k pins added to a board will be the continuation of a pinning spree. Across each row, features are shown in decreasing order of maximum absolute correlation coefficient value, *i.e.* more strongly correlated or anti-correlated features are listed first.

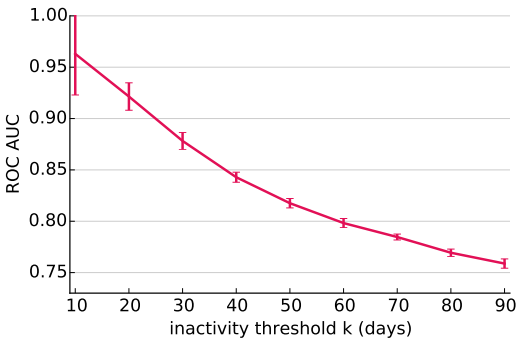


Figure 10: If we observe the first 30 days of a board and want to predict if a board's effective lifespan will be over a year given that the the board growth cannot remain inactive for more than k days, prediction performance is best for small k.

days across an entire year, and boards who don't meet this activity threshold. Our goal is to predict whether a given board's effective lifespan with threshold k will be at least one year (365 days), after observing only the first 30 days' worth of activity. Note that we do not require boards to actually meet the inactivity threshold during the observation period in our effective lifespan calculations, to prevent this task from being trivial for $k < 30$.

Figure 10 gives the results. Prediction performance is quite strong, with ROC AUCs ranging from 0.75 to 0.96. Note that performance is best when the allowed inactivity threshold is lower, likely because highly active boards are more likely to differentiate themselves from the rest of the board population in terms of initial behavior. Our results are promising for content discovery applications, for whom identifying the most active collections is likely to be more useful than identifying only moderately active collections.

Feature analysis. Correlation coefficients between features and whether a board's effective lifespan is longer than 365 days are shown in Figure 9(b). Here, we see that many features are strong signals for the prediction task. In particular, *"timestamp: last pin"* (0.573, $k = 0$), *"# pins observed"* (0.467, $k = 0$), and *"average time between sprees"* (-0.449, $k = 0$) have especially high absolute correlation coefficients. Furthermore, we observe that correlation coefficients for *"average spree size"* are negative, validating our surprising finding in Section 6(b) that smaller spree sizes lead to longer effective lifespans.

In general, we see that for multiple features (*"timestamp: last pin"*, *"% deltas in upper distribution"*, and *"average spree size"*), a positive correlation coefficient for this prediction task means a negative correlation coefficient when predicting final board size, and vice versa. This observation highlights the fact that large board size prediction and long effective lifespan prediction are two very different prediction tasks. Finally, we observe that correlation coefficients trend towards zero as the inactivity threshold k increases, which explains why prediction performance is best when k is smaller.

7.3 Identifying a pinning spree

We've shown that signals from initial collection growth can be extremely useful in predicting long-term collection growth, which raises an interesting question: can earlier signals predict short-term growth behavior as well? In our final task we aim to predict whether a given pinning spree will continue growing for the next k pins. This task is useful in a number of real-world settings; for example, content discovery applications may prefer to advertise to users of boards whose pinning sprees are likely to continue for some time,

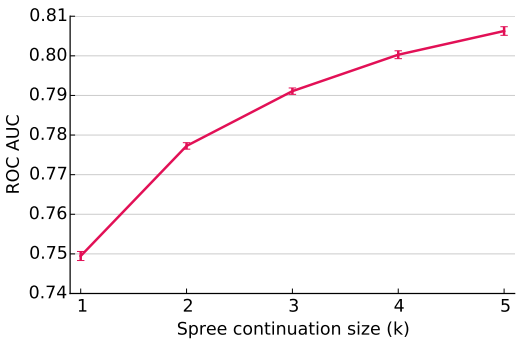


Figure 11: If we want to predict whether the next k pins added to a board will be the continuation of a pinning spree, performance improves as k increases.

because there is a higher chance that they will read and absorb the advertisement contents.

For k ranging from 1 to 5, we choose a random point for each board such that at least five pins have been added to the board by that point, so that there is data to construct features from. We aim to predict, knowing that the board will increase in size by at least k more pins, whether these k pins will be the continuation of a pinning spree. The dataset is balanced for each k.

Prediction results are shown in Figure 11. As with previous tasks, performance is solid, with AUCs ranging from 0.75 to 0.81. Furthemore, performance improves as k increases, meaning it is easier to predict whether a pinning spree will continue growing by 5 pins, than by a single pin. This matches our intuition that because a large spree continuation is rarer by definition, boards that tend to contain large sprees will differentiate themselves more easily.

Feature analysis. The correlation coefficients between different features and whether the next k pins will be the continuation of a pinning spree are shown in Figure 9(c). The first and most noticeable observation is that unlike the previous two models, here correlation coefficient doesn't decrease or increase dramatically as k changes. Instead, signal strength increases very gradually with k, suggesting that shorter and longer sprees are relatively similar in terms of underlying properties; after all, they are both sprees. We also observe that as expected, larger sprees are more likely to occur in boards with a smaller *"% deltas in upper distribution"* (-0.413, $k = 1$), as well as boards with a prior history of long sprees, as represented by *"avg spree size"* (0.213, $k = 1$).

In terms of the time between sprees, one might intuitively expect users of boards that do not grow for a long time to have lost interest in that board, causing smaller future spree sizes. Instead, we surprisingly find that *"avg time between sprees"* is positively correlated with larger sprees occurring in the future (0.154, $k = 1$). One potential explanation is that users who choose to grow a board after a long break may have a larger amount of previously unseen content to choose from, making it easier for them to pin more content and form longer sprees.

Taken together, our results show that with a set of carefully selected signals based on initial collection growth, content discovery applications can effectively model future growth behavior in a variety of growth-related prediction tasks.

8. DISCUSSION & CONCLUSION

In this paper we shed light on the way online collections grow over a long period of time. Through a large-scale study of over 2.6 million Pinterest boards across three years, we identified patterns and signals of long-term collection growth. We found that the inter-event time distribution of pins follows a two-component lognormal mixture model, characterizing collections as a series of short term sprees divided by longer breaks. We identified common patterns between sprees in terms of growth speed and content similarity. Using information gleaned from these sprees, we uncovered relationships between initial and final board growth for final board size and effective lifespan. Finally, we synthesized our insights into a series of predictive models of board growth that show that initial collection growth is incredibly useful in a variety of settings.

At a higher level, the goal of this work has been to provide a methodology by which the growth of online collections can be studied. There are also positive applications for content discovery platforms. Our insights make it easier to surface collections that are active, and recommend these active collections to users. Signals that identify the growth potential of collections are also useful for user retention, as specialized user interfaces and intervention strategies (*e.g.* notifications, email, changes in recommendations) can be designed to facilitate content curation and collection growth.

There are many fruitful avenues for future work. We believe that studying additional metrics for measuring the "success" of collection growth, such as social approval in the form of likes and follows, can uncover additional insights. While our work focuses on growth on a per-collection basis, it would be also be interesting to examine collection growth at the user level. Finally, it may be useful to study the evolution of content added to a collection over time using textual or image features, to better understand the motivations behind collection growth.

Acknowledgements

We thank Dan Frankowski, Marinka Zitnik, Tim Althoff, Austin Benson, Steven Bach, and David Lo for their help and useful discussions.

9. REFERENCES

[1] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 44–54, 2006.

[2] A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[3] S. E. Brockwell and I. R. Gordon. A comparison of statistical methods for meta-analysis. *Statistics in medicine*, 20(6):825–840, 2001.

[4] S. Chen, J. L. Moore, D. Turnbull, and T. Joachims. Playlist prediction via metric embedding. In *Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 714–722, 2012.

[5] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proc. of the 23rd Int'l Conf. on World Wide Web*, pages 925–936, 2014.

[6] A. G. Close and M. Kukar-Kinney. Beyond buying: Motivations behind consumers' online shopping cart use. *Journal of Business Research*, 63(9):986–992, 2010.

[7] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts. No country for old members: User lifecycle and linguistic change in online communities. In *Proc. of the 22nd Int'l Conf. on World Wide Web*, pages 307–318, 2013.

[8] M. Feinberg, G. Geisler, E. Whitworth, and E. Clark. Understanding personal digital collections: an interdisciplinary exploration. In *Proc. of the Designing Interactive Systems Conf.*, pages 200–209, 2012.

[9] D. Greene, F. Reid, G. Sheridan, and P. Cunningham. Supporting the curation of twitter user lists. *arXiv preprint arXiv:1110.1349*, 2011.

[10] A. Halfaker, O. Keyes, D. Kluver, J. Thebault-Spieker, T. Nguyen, K. Shores, A. Uduwage, and M. Warncke-Wang. User session identification based on strong regularities in inter-activity time. In *Proc. of the 24th Int'l Conf. on World Wide Web*, pages 410–418, 2015.

[11] E. Hwang. 100 million of the most interesting people we know, September 2015. `https://blog.pinterest.com/en/100-million-most-interesting-people-we-know`.

[12] J. L. Iribarren and E. Moro. Impact of human activity patterns on the dynamics of information diffusion. *Phys. Rev. Lett.*, 103:038702, Jul 2009.

[13] S. R. Kairam, D. J. Wang, and J. Leskovec. The life and death of online groups: Predicting group growth and longevity. In *ICWSM*, pages 673–682, 2012.

[14] P. Koehn. Statistical significance tests for machine translation evaluation. In *Proc. of the 2004 Conf. on Empirical Methods in Natural Language Processing*, pages 388–395, 2004.

[15] R. Kumar and A. Tomkins. A characterization of online browsing behavior. In *Proc. of the 19th Int'l Conf. on World Wide Web*, pages 561–570, 2010.

[16] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, and A. Kustarev. Prediction of retweet cascade size over time. In *Proc. of the 21st ACM Int'l Conf. on Information and Knowledge Management*, pages 2335–2338, 2012.

[17] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 462–470, 2008.

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

[19] P. F. Marty. My lost museum: User expectations and motivations for creating personal digital collections on museum websites. *Library & information science research*, 33(3):211–219, 2011.

[20] T. Menjo and M. Yoshikawa. Trend prediction in social bookmark service using time series of bookmarks. In *Proc. of DEWS*, volume 2, pages 156–166, 2008.

[21] J. Qiu, Y. Li, J. Tang, Z. Lu, H. Ye, B. Chen, Q. Yang, and J. E. Hopcroft. The lifecycle and cascade of wechat social messaging groups. In *Proc. of the 25th Int'l Conf. on World Wide Web*, pages 311–320, 2016.

[22] V. Rakesh, D. Singh, B. Vinzamuri, and C. K. Reddy. Personalized recommendation of twitter lists using content and network information. In *ICWSM*, 2014.

[23] T. Schnabel, P. N. Bennett, S. T. Dumais, and T. Joachims. Using shortlists to support decision making and improve recommender system performance. In *Proc. of the 25th Int'l Conf. on World Wide Web*, pages 987–997, 2016.

[24] B. Stvilia and C. Jörgensen. End-user collection building behavior in flickr. *Proc. of the American Society for Information Science and Technology*, 44(1):1–20, 2007.

[25] A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(3):036127, 2006.

[26] S. Velichety and S. Ram. Examining lists on twitter to uncover relationships between following, membership and subscription. In *Proc. of the 22nd Int'l Conf. on World Wide Web*, pages 673–676, 2013.

[27] M. Villi, J. Moisander, and A. Joy. Social curation in consumer communities: Consumers as curators of online media content. *NA-Advances in Consumer Research Volume 40*, 2012.

[28] P. Wang, X.-Y. Xie, C. H. Yeung, and B.-H. Wang. Heterogenous scaling in the inter-event time of on-line bookmarking. *Physica A: Statistical Mechanics and its Applications*, 390(12):2395–2400, 2011.

[29] R. Wetzker, C. Zimmermann, and C. Bauckhage. Analyzing social bookmarking systems: A del. icio. us cookbook. In *Mining Social Data Workshop*, pages 26–30. ECAI, 2008.

[30] J. Yang, X. Wei, M. S. Ackerman, and L. A. Adamic. Activity lifespan: An analysis of user survival patterns in online knowledge sharing communities. *ICWSM*, 10:186–193, 2010.

[31] T. Zhang, P. Cui, C. Faloutsos, W. Zhu, and S. Yang. Come-and-go patterns of group evolution: A dynamic model.

[32] X. Zhao and S. E. Lindley. Curation through use: understanding the personal value of social media. In *Proc. of the 32nd annual ACM conference on Human factors in computing systems*, pages 2431–2440, 2014.

[33] Z.-D. Zhao and T. Zhou. Empirical analysis of online human dynamics. *Physica A: Statistical Mechanics and its Applications*, 391(11):3308–3315, 2012.

[34] T. Zhou, H. A.-T. Kiet, B. J. Kim, B.-H. Wang, and P. Holme. Role of activity in human dynamics. *EPL (Europhysics Letters)*, 82(2):28002, 2008.