Sharing the Loves: Understanding the How and Why of Online Content Curation

Changtao Zhong
King's College London
changtao.zhong@kcl.ac.uk

Sunil Shah
Last.fm
sunil@last.fm

Karthik Sundaravadivelan
King's College London
karthik.sundaravadivelan@kcl.ac.uk

Nishanth Sastry
King's College London
nishanth.sastry@kcl.ac.uk

Abstract

This paper looks at how and why users categorise and curate content into collections online, using datasets containing nearly all the relevant activities from Pinterest.com during January 2013, and Last.fm in December 2012. In addition, a user survey of over 25 Pinterest and 250 Last.fm users is used to obtain insights into the motivations for content curation and corroborate results. The data reveal that curation tends to focus on items that may not rank highly in popularity and search rankings. Yet, curated items exhibit their own skewed popularity, with the top few items receiving most of the attention; indicative of a synchronised community. We distinguish structured curation by active categorisation from a more passive bookmarking by 'liking' an item, and find the former more prevalent for popularly curated items. Likes, however, are initially accumulated at a faster pace. Finally, we study the social value of content curation and show that curators attract more followers with consistent activity, and diversity of interests. Interestingly, our user study indicates a divided opinion on the relevance of the social network.

Introduction

Social content curation is a new trend which has emerged following on the heels of the information glut created by the user-generated content revolution. Rather than create new content, content curation websites allow their users to categorise and organise collections of content created by others that they find online. These users or content curators provide an editorial perspective by highlighting interesting content. Typically, a social component is also involved: users can follow other content curators that they find interesting, as a way of gaining exposure to new and interesting content.

Although content curation has only recently become a buzzword, sites on the Web have supported the actual process of categorising and sharing content with followers for a few years now. For instance, delicio.us allowed users to categorise interesting URLs by tagging them, and sharing them with followers. digg.com and reddit.com have allowed sharing of news articles, and so on. In this paper, we take a broad view of content curation and seek to understand the basic process by examining two very different datapoints: Pinterest, arguably the most popular content curation website for sharing pictures and videos, and Last.fm, a popular social music recommendation service.

The two websites are used differently but have similar support for content curation: Users on Pinterest.com 'pin' images onto boards to categorise them, or 'like' the images to express an interest without categorisation. Curation is not the primary purpose of Last.fm, but relevant support has existed for over seven years. Similar to Pinterest, users can categorise music tracks by attaching tags to them, or 'love' tracks to express interest without categorising. Tags on Last.fm can be used to create playlists or personalised radio stations. Both sites support social networks where users can form social links, based on their interests.

Through these websites we seek to understand why people curate, how they curate, and what others, namely followers of the content curators, find useful. For our analysis, we collected nearly all of the curation actions and the entire social networks of active users from both websites for a period of time: for three weeks in January 2013 for Pinterest, and over the month of December 2012 for Last.fm. To complement the patterns we observe from the collected data and to better understand the motivations for content curation, we conducted a user study through online survey questionnaires and interviews, recruiting over 25 and 250 users of Pinterest and Last.fm respectively.

To understand why people curate, we look at the popularity distributions of highly curated items. The most popular curated items appear to be of niche interest that may not rank highly in other popularity rankings. For instance, the items most pinned or most liked on Pinterest are from websites with a low PageRank value or Alexa Global Traffic Ranking. Similarly, the most tagged or loved tracks on Last.fm tend to have a low rank in weekly music charts of total radio airplay, and greatest sales volume. We conjecture that curation might provide a personal value to the curators by collecting together items which may be difficult to find by other means. Our user studies provide support to the notion that curation provides personal value to the curators. Interestingly, despite their low popularity in other rankings, there appears to be a consensus on which items are most curated, and curation actions are highly skewed towards the top items on each site: The top 10% (0.1%) of items get over 70% of the curation actions on Pinterest (Last.fm), indicative of a synchronised community. These findings provide evidence for Clay Shirky's theories that "curation comes up when search stops working", and that "the job of curation is to synchronize a community so that when they're all talking about the same thing at the same time, they can have a richer conversation than if everybody reads everything they like in a completely unsynchronized or uncoordinated way" (Shirky 2010).

Next, we examine the different curation actions to better understand the process. Based on the similarity between the curation actions on Pinterest and Last.fm, we propose a distinction between two kinds of content curation actions: *unstructured curation*, which involves highlighting or collecting interesting content without categorising them (e.g., 'like' or 'love'), and *structured curation*, categorising content along with other "similar" items from some perspective (e.g., tag or 'pin'). We find that different users prefer different actions, with some preferring unstructured, and others structured curation. The proportion of users preferring each varies from site to site, and may be a result of differences in the way the site is structured and side effects associated with each action (e.g., the action of 'love'ing a track on Last.fm is used to recommend other similar tracks to the user). However, ranking items based on the number of unstructured or structured curation actions, we see that the top items in both rankings receive more structured curation actions than unstructured. In contrast, for all items, we see that the easier action of unstructured curation accumulates faster.

Finally, we study the social value of content curation. Bhargava, who appears to have coined the term content curator, defined it as "someone who continually finds, groups, organizes and shares the best and most relevant content on a specific issue online. The most important component of this job is the word 'continually.' "(Bhargava 2009). Consistent with his view, we find that curators who are regular and consistent in their activities accumulate the most number of followers on the respective websites. Diversity of interests is also similarly rewarded: Curators with an expertise in multiple genres of music in Last.fm or categories on Pinterest are similarly successful in attracting followers. Additionally, we find that in Pinterest, successful users are those who prefer structured curation or pinning to merely 'liking' items. On Last.fm, where the relative value proposition to the user of tagging and loving tracks are structured somewhat differently, we do not find the same advantage for structured curation.

Methodology and Data Description

We used two complementary approaches to examine content curation on both the Pinterest and Last.fm websites. The main approach was a quantitative analysis of datasets containing relevant curation actions over a fixed period of time. Our data-based findings were corroborated and complemented using a qualitative approach consisting of user studies and interviews.

Data and website description

We first provide a background about each of the websites we use, and the data we collect about their actions and their social relations.

Pinterest Pinterest is a photo sharing website that allows users to save images and categorize them on different collections. Images added on Pinterest are termed *pins*; we will use the terms pin and image interchangeably. A pin can be created by *pinning* or importing from a URL external to pinterest.com, or *repinning* from a existing pin on pinterest. Users organise their pins into collections called *pinboards* or *boards*. A board needs to be specified at the time of pinning; pins may be moved to a different board later on. A repin creates a new pin on the repinning user's board. Each board can belong to one of 32 globally specified *categories* on pinterest. Each category has a page on pinterest.com, highlighting the latest pins. In addition to pinning or repinning, users can *like* a pin or *comment* on a pin. *Likes* express an interest in or appreciation of a pin without adding it onto the liking user's collections. The most recent likers of a pin are listed on the pin's webpage on pinterest, and the likes of a user are collected on the user's profile. In addition to these content curation actions, users can also actively *follow* other users or boards they find interesting, effectively creating a directed social graph.

To analyze the curation activity on Pinterest, we collected nearly all activities by crawling the main site between 3 and 21 Jan, 2013. The crawl proceeded in two steps: firstly, to discover new pins, we visited each of the 32 category pages every 5 minutes, and collected the latest pins of that category. Secondly, for every pin obtained this way, we visited the webpage of the pin every 10 minutes. A pin's webpage lists the 10 latest repins and the 24 latest likes; we added these to our dataset, along with the approximate time of repins, likes and comments (if any). In this paper, we focus on repins and likes which comprise the vast majority of actions.

For any pin, if more than 10 repins or 24 likes had accumulated since our last visit, we may have missed some data. The danger of missing data is higher for popular images which may accumulate likes and repins faster than other images. However, if we find an overlap between the latest repins/likes on successive visits, then we can be sure of not having missed data. In practice, we find that even for popular images (those with more than 500 actions), we have missed data in less than 0.06% of visits for repins and 0.02% for likes. For all images, the fraction of visits which resulted in missed data stands at 5.7×10^{-6} for repins and 9.4×10^{-7} for likes.

In addition to these curation actions, we also obtained the social graph of Pinterest using a snowball sampling technique starting from a seed set of 1.6 million active users which we collected initially. In total 30.5 million users and 315.2 million directed edges between them were obtained. Users with a local clustering coefficient of 0 were filtered[1]

[1] This affects Pinterest results reported per-user (mainly in the final section, "What Other People Find Useful"). We justify this filtering on the basis that users with clustering coefficient of 0 have far fewer activities on the site. For example, in our seed set of 1.6 million users, the average number of pins for users with zero clustering coefficient is 324.3, while this value for users with non-zero clustering coefficient is 1686.3.

resulting in a smaller social graph of around 7.1 million users and 192.7 edges. For each of the remaining 7.1 million users, we collected statistics such as a string description, their boards, total number of pins and likes since joining the site, and numbers of followers and following users.

Last.fm Last.fm is a popular social music recommendations website which offers a radio service, a service to allow users to submit and track what music they have listened to and music recommendations calculated using collaborative filtering algorithms. The Last.fm social graph is a simple symmetric directed graph where users can *friend* each other. The friendship must be approved by each user. While there is no explicit following model through which content curated (i.e. loved or tagged) by friends is made visible to users, users can see friends' loved tracks via a link on their (personalised) Last.fm home page. A user's tagging and loving activity can also be seen by visiting their profile page. The action of submitting a track's name or identifier to Last.fm to record a listen is known as a *scrobble*. All scrobbled tracks are shown on the user's profile and are public unless deleted. Charts of their most scrobbled tracks are shown on their profile page. Users can also *love* a song - either retroactively via the Last.fm website or via any of the Last.fm client applications. Loved tracks are shown on their profile page. Similarly, users can *ban* a song when listening to it via Last.fm radio, either via the website or via any of the client applications. Bans are one of the most infrequently used social actions and although they are public, are not displayed on the profile page. All three actions are used to influence playlisting for Last.fm radio as well as music recommendations.

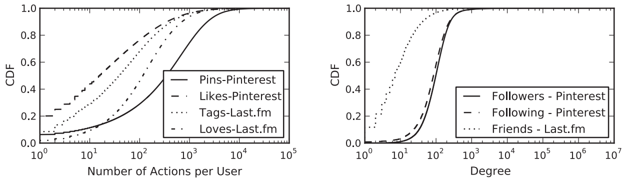
Additionally, users can *tag* a song via the Last.fm website. Tags are both global and local - any tags a user has applied will be shown on their profile page, as well as aggregated with other users' tags and shown on artist and track pages. Users can also comment on nearly every content page (catalogue and user) on the Last.fm website via the page's *shoutbox*. However, due to the complexity involved in extracting shouts, we have excluded them from our dataset.

Our dataset was generated directly from Last.fm source data using Apache Hive running on a Hadoop cluster. We consider all users of Last.fm worldwide that have both tagged and loved tracks during December 2012. We filter down this dataset to include all users who are have at least one other friend within the same dataset. This yields nearly 300,000 users who we consider to be power users of the site. For each of these users, we extract all stored scrobble, love and tag data along with approximately 5.9 million undirected edges between users.

Summary Table 1 provides a summary of the aggregate volume of data collected, and Fig. 1 provides an indication of the per-user distribution of the volume of the data.

	Pinterest	Last.fm
Timespan	03–21 Jan 2013	01–31 Dec 2012
Users	8,452,977	291,562
Relationships	96,390,143	5,887,159
Likes/Loves	19,907,874	89,338,529
Repins/Tags	38,041,368	59,622,487

Table 1: Dataset details



(a) Actions per user

(b) Degree distribution

Figure 1: Distribution of the number of curation actions and social relationships per user in Pinterest and Last.fm datasets

User study

A qualitative approach was undertaken through user surveys and semi-structured interviews. We questioned users of Pinterest and Last.fm on their general behaviour on each website and on their attitudes towards usage of various social signals on each website. These signals were primarily *liking* and *pinning/repinning* for Pinterest, and *loving* and *tagging* for Last.fm.

Other key areas of interest included the motivation for users to curate content and the underlying relationship between curation and social networking. Users were also questioned on the value they placed on the social aspect of each site and if this went beyond the scope of purely curation activities.

Pinterest The users who participated in the Pinterest study were sourced from the social circles of the authors and from posting requests on a Pinterest forum known as Pintester and on the official Pinterest Facebook page. 30 users participated in the survey and a further 3 were interviewed. Interviews were loosely structured around the survey aims and were presented to interviewees in themes. Interviewees were asked to narrate their behaviour given certain scenarios and were allowed to continue on their motivations and reasons behind said behaviour. They were also encouraged to provide examples of their past practices and suggest how and why these behavioural trends varied, if at all. Our main interest was to understand user perceptions and to find motivations behind their curating activities. Insight into how they perceived social connections within the site was also sought after. Interview responses were coded and then compared against trends prevalent in the survey responses.

Last.fm A much larger user population of 270 participated in the Last.fm survey since the survey was advertised over the official Last.fm Twitter account[2]. The questions posed to users followed the structure adopted by the Pinterest survey but were specific to the dynamics and infrastructure of Last.fm. Unstructured responses were obtained through open-ended questions and were analysed to provide reasoning behind predominant user tendencies.

[2] The study itself was carried out independently of Last.fm.

Why People Curate

In this section, we seek to find implicit reasons for why people curate by examining the characteristics of the content they curate. Our approach will be to compare different popularity ranks with basic ranks created by the volume of curation actions. First, we ask where the content in curation system is from, by correlating curation with traditional popularity ranks, and show that curation serves a different purpose than, say, search. Then, the distribution of curation activity is analysed and a highly skewed distribution is obtained, revealing that users synchronise and focus on the same small number of items. We draw on our user studies to provide support for and comment on these findings.

Curation highlights new kinds of content

A first question is whether curation serves a new and different purpose from other approaches to finding and highlighting interesting content. Popularity rankings traditionally highlight content which a community finds useful. Therefore, we compare curation with other traditional notions of popularity. In the case of Last.fm, we compare against weekly sales and radio airplay charts published by Music Week[3], a trade paper for the UK record industry and an established music data provider. In the case of Pinterest, we do not have a well accepted global popularity ranking of images. As a proxy, we use the website where the curated image was originally found, and compare the rank of a website on Pinterest (in terms of number of repins and likes), with its rank in search (PageRank value, obtained from Google via its Search API), and its global traffic ranking (according to Alexa[4]).

	Avg. Repins	**Avg. Likes**
Avg. Repins	/	0.912
Avg. Likes	0.912	/
Alexa Ranking	-0.010	0.032
PageRank	0.195	0.150

Table 2: **Curation highlights websites not popular in other rankings.** Low correlation coefficients between curation-based ranking of websites (ranking by the average number of repins or likes) and traditional websites rankings (Alexa Traffic Ranking and Google PageRank) reveal that curation serves a new purpose of highlighting non-traditional sites.

In the case of Pinterest, we find that websites with highly repinned or liked images tend not to have a high PageRank or Alexa Global Traffic Rank. In fact, Table 2 shows that, when considering all websites, there tends not to be a correlation between ranking based on number of repins/likes and traditional ranking based on Google PageRank or Alexa Global Traffic estimates. Thus, we conclude that curation highlights a different set of sites compared to search and traffic. The low correlation with PageRank also lends support to Shirky's theory that "curation comes up when search stops working"(Shirky 2010). Similarly, Table 3 shows a similar lack of correlation between highly ranked tracks through curation and the traditional Music Week rankings.

	Loves	**Tags**
Loves	/	0.39 (0.74)
Tags	0.39 (0.74)	/
Music Week Sales	-0.02 (0.05)	0.04 (0.14)
Music Week Airplay	-0.11 (-0.01)	0.04 (-0.04)

Table 3: **Curation highlights tracks not popular in traditional rankings.** Low correlation coefficients between curation-based ranking of tracks (ranking by number of tags or likes) and traditional music track rankings (obtained from Music Week) reveal that curation may be being used to find music that is "off the charts" (i.e., is not mainstream). The coefficients shown consider UK-based users only for the curation-based rank, for a fair comparison with Music Week, a UK-based rank. Numbers in brackets indicate correlation coefficients with a ranking considering users worldwide.

Curation for personal vs. social value

A second aspect of Shirky's theory is that the "job of curation is to synchronize a community so that when they're all talking about the same thing at the same time, they can have a richer conversation than if everybody reads everything they like in a completely unsynchronized or uncoordinated way". We find evidence for this by examining the distribution of curation actions in our corpus. Fig. 2 shows a highly skewed popularity distribution, with a large proportion of the user base curating a selected minority of items. However, that skewness is expected in popularity distributions, hence this is not in itself a confirmation of a community which consciously synchronises itself.

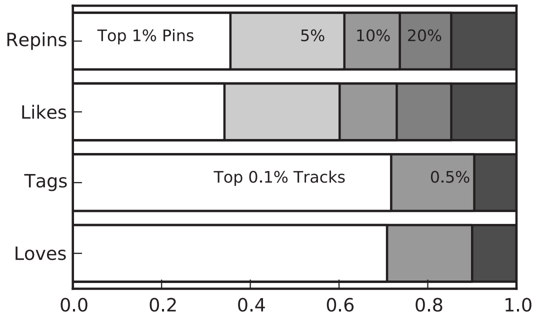


Figure 2: **Distribution of curation activity is highly skewed towards a few popular items.** In Pinterest, nearly 40% of curation activities (Repins and Likes) is for top 1% of images and over 73% are for top 10% images. In Last.fm,the top 0.5% of tracks account for about 90% of curation activities (Tags and Loves).

Rather than answer the difficult question of whether curation creates value for users by synchronising a community,

[3] http://www.musicweek.com/

[4] http://www.alexa.com

we turn to our user study to find if users perceive their community to be useful (and thereby, determine whether the social value of curation is a motivating factor for why people curate).

Some users value the ability to talk with others who share a similar taste. For instance, one Last.fm user states:

> Social connection with people who share a similar taste in music is an exciting thing. Say, for example, that you've found a band that you love, and you live in an area where people generally prefer listening to pop or maybe classical or filmi (Bollywood) tracks. So, you start looking for people around you who share the same taste as you in music and become fast friends. Last.fm has simplified that process.

A Pinterest respondent values the ability to serendipitously discover through other users' items which they might like, placing an implicit value in the Pinterest community:

> I like the feeling of stumbling on things which I did not know I would like but I do.

However, such views are from a minority of users. A number of users use curation sites as a personal tool: 85% of Pinterest respondents use it as a personal collection or scrapbook and only 48% of the population use the site to display their content to others (Note that our survey allowed multiple answers to be selected for this question). 39% of Last.fm users tag tracks for personal classification, whereas another 39% tag to create a global classification according to genres.

The majority of users shared in this Last.fm user's view:

> I find the social aspect more useful and interesting with people I know, rather than developing new interactions based on music taste.

One Pinterest user felt strongly about their aversion towards social interaction on the site:

> I don't really see a point (in communicating with a fellow user). And also the beauty of Pinterest, is the ability to pin things from strangers. Why would I want to get to know them.

Thus, we conclude that although the community of users may focus its curation actions on a few items (as seen from the popularity skew), this synchronisation is not a conscious effort. Users, largely, are not actively trying to curate for social value and do not try to integrate within their respective communities.

How People Curate: Understanding Curation Actions

As detailed before in the Methodology section, multiple curation actions are available to a user. For instance a user on Pinterest can pin an item, like it or comment on it. On Last.fm, tracks can be tagged, loved or banned or shouted. We note that these actions can be distinguished based on whether they simply highlight an item (love, like, ban, comment, shout), or they also organise the item onto user-specific lists (pinning an item onto a user's board, or attaching a user's tag to a track). We term the former as *unstructured curation* and the latter *structured curation* because of the organisational structure induced by pinning or tagging.

We use this framework to study curation actions: Do users have a preference towards one kind of action, do they use structured actions preferentially in one setting, what the relative dynamics of the different kinds of action are, etc.

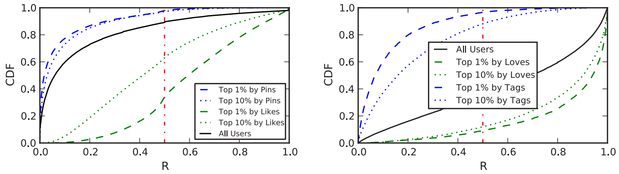
To investigate the relationship between the two forms of curation, we define an *unstructured curation ratio* R as:

$$R = \frac{Unstructured}{Unstructured + Structured} \tag{1}$$

Some users prefer structured, others unstructured

First we explore how users curate content, and whether they prefer structured or unstructured curation. We calculate the unstructured curation ratio R for each dataset and consider the top 1%, the top 10% and all users for each activity on both websites in Fig. 3[5].

We define users who prefer structured curation over unstructured curation (i.e., have $R < 0.5$) as *structured curators*. Conversely, users who prefer unstructured over structured curation ($R > 0.5$) are termed *unstructured curators*.



(a) Users on Pinterest

(b) Users on Last.fm

Figure 3: **CDF of users' unstructured curation ratio** R In both Pinterest and Last.fm, there are a mixture of structured ($R < 0.5$) and unstructured ($R > 0.5$) curators. However, generally Pinterest users participate in structured curation activities whilst most Last.fm users participate in unstructured curation activities.

In Fig. 3 we first draw attention to the difference between the proportion of structured and unstructured curators on each network. Fig. 3a shows that on Pinterest, more than 80% of all users are structured curators. Comparatively, Fig. 3b shows that less than 40% of all Last.fm users are structured curators. This corresponds with the findings of our user study. We found that the majority of Pinterest users surveyed would rather repin a post than like it if it matched their interests. This is irrespective of whether the post was from a user they were following or not. Similarly, a majority of Last.fm users would rather love than tag a music track.

However, as expected, when filtering for the top 1% and 10% of users for each curation activity, we see that the unstructured curation ratio moves closer in favour of that activity, on both websites: The most prolific likers on Pinterest are unstructured curators (i.e., $R > 0.5$ for these users, despite the prevalence of pinning on Pinterest); the most prolific taggers on Last.fm are structured curators (i.e., $R < 0.5$, despite the importance of loves on Last.fm).

[5] Since Pinterest does not distinguish between original pins and repins, both are included to represent the structured curation action.

The larger proportion of the top users by loves who are unstructured curators on Last.fm can be explained by the relative prevalence of loves to tags in our dataset, as well as one of the major side effects of loves. When a user loves a track on Last.fm, this action is fed back into their music recommendations and displayed to their friends. Loves are thus a more capable curation activity on Last.fm compared to likes on Pinterest. This is confirmed by our user study: 65% of surveyed Last.fm users have never tagged a track. Conversely, only 11% have never loved a track.

Structured curation is preferred for popular items

Next we explore how items themselves are curated, and whether the majority of items are curated in a structured or unstructured manner. We calculate the unstructured curation ratio R for each item in both datasets and consider the top content items by curation activity in Fig. 4. We observe that regardless of the ranking method used (i.e., whether the ranking is based on the volume of structured or unstructured curation action received), the majority of items have an $R < 0.5$: there are more structured curation actions for top items, whether they are the top items for structured or unstructured curation. In other words, even top liked items have more pins than likes on Pinterest (similarly for Last.fm, top loved items have more actions adding tags than actions 'loving' the track). This is further supported through our Pinterest user study where average R for popular content was 0.33 and for unpopular content was 0.5. The Last.fm survey did not address this question.

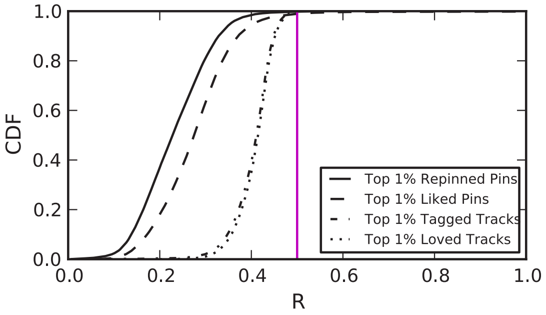


Figure 4: **CDF for unstructured curation ratio R of top items on Pinterest and Last.fm** Magenta line indicates $R = 0.5$. In Pinterest and Last.fm, all of the top items have $R < 0.5$, i.e. they are all subject to structured curation. Notice that even the top items for unstructured curation (i.e., top liked or loved items) have $R < 0.5$.

Unstructured curation is faster than structured

In this section, we discuss how items accumulate different curation activities over time. In order to compare these, we plot the action time - the time span between the n-th action and the time a content item was originally posted. We consider this time for both structured (pin/tag) and unstructured (like/love) curation activities.

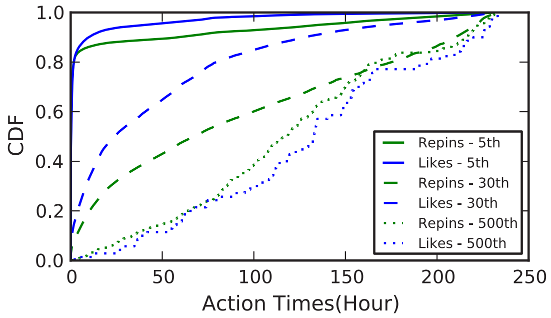
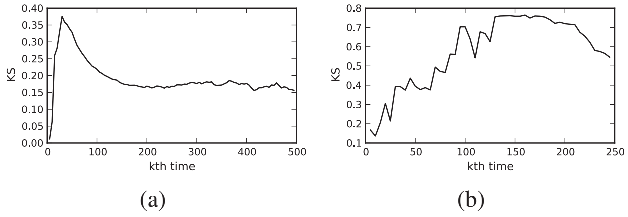


Figure 5: **CDF of Pinterest repin and like action times** The kth repin (like) time for a pin is the time between creation of a pin and the kth repinning (liking) in Pinterest. Likes accumulate quicker at first and there is a considerable difference between in the time it takes to get 30 repins and 30 likes. The distributions of the times for kth likes and repins converge as k increases to 500.



(a)

(b)

Figure 6: **Kolmogorov-Smirnov statistic for action times** Extending Fig. 5, for both Pinterest and Last.fm, there is a difference between structured and unstructured curation over successive actions. There is a noticeable peak for Pinterest, after which the difference between the two actions is minimised. For Last.fm, we see a similar tail off, albeit after many more actions.

Fig. 5 shows the time taken for items to reach their 5th, 30th and 500th curation actions on Pinterest. We find that the majority of pins reach 5 curation actions (whether repin or like) in several hours. As expected, it takes much longer to reach their 30th curation action. However, there is a considerable difference between the 30th action time for likes vs. repins: For 80% of items, accumulating 30 likes take approximately 100 hours whilst repins take approximately 200 hours. This difference decreases when we consider the 500th action times for each activity.

In Fig. 6, we summarise the difference between the distribution of $T_s(k)$, the kth action time for structured curation, and the distribution of $T_u(k)$, the kth action time for unstructured curation. This difference can be measured using the Kolmogorov-Smirnov statistic given by $D = \max(T_s(k) - T_s(k))$. For Pinterest, we see a quickly growing difference between likes and repins until a initial peak, after which the two converge again - suggesting that, initially, likes accumulate faster than repins. As items become more popular, repins catch up and the two grow at a similar rate. For Last.fm, we see a similar result - except that structured curation activity (tagging) does not completely catch up with unstructured curation activity (loving). This can be explained by what we show in Fig. 4 - structured curation is generally stronger for Pinterest items than those on Last.fm.

When do people curate?

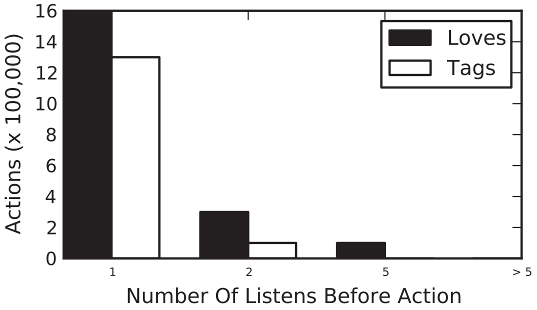


Figure 7: **Number of listens before first love and tag on Last.fm** reveals that most users who curate an item do so on their first listen.

A basic question that arises is when does a curation action happen during the content consumption cycle of a user. While it is difficult to answer this question in general, our data for Last.fm includes ‘scrobbles’ which record each listen. Using this, Fig. 7 shows the number of activities that a user participates in before loving or tagging a track on Last.fm for the most popular 1,000 tracks in our December dataset. Interestingly, we find that most users who love or tag a track predominantly on their 1st listen. A insignificant minority of users listen to a track twice before loving or tagging it and no user takes longer than 2 listens to tag a track and 5 listens to love a track.

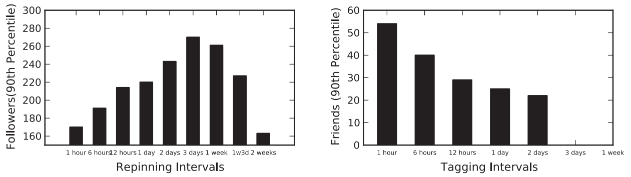
What Do Other People Find Useful?

Although as suggested previously, many users view curation as a highly personal activity, some users accumulate more followers than others. This section sheds light on what curation behaviours other people find useful by using the number of accumulated followers as a metric. In each case, we consider the per-user distribution of the values for some attribute of the user’s behaviour (e.g., interval between repins, number of music genres the user is interested in, or the unstructured curation ratio R). Firstly, we separate users into bins. Usually, we do this based on the user’s value of the attribute considered (e.g., based on the board categories of the user). Next for each bin, i.e., for each value of the attribute being considered, we compute the 90th percentile of the number of followers accumulated by users in the bin as a measure of how useful the bin’s value of the attribute is, to other users.

In summary, for both Last.fm and Pinterest, we find that regular curators who have a short interval between successive curation actions accumulate more followers, as do curators who have a diversity of interests. In Pinterest, we also find that users who prefer structured curation (i.e., those who prefer ‘pinning’ to ‘liking’) accumulate more followers. This result does not carry over to Last.fm, where structured curation does not have the same predominant role. We have verified that each of the results in this section are robust against the choice of 90th percentile as a summary measure. (Similar results hold for 80th and 50th percentile values as well).

Consistent and regular updates

Bhargava has suggested that the most important part of a content curator’s job is to continually identify new content for their audience (Bhargava 2009). Fig. 8 examines the role of regularity, by plotting the 90th percentile of the intervals between consecutive structured curation actions[6] for each user vs. the 90th percentile of the followers accumulated, and finds support for this theory. Note that for Pinterest, too short an interval between repins could detract followers. However, Last.fm does not exhibit this phenomenon. We conjecture that given the order of magnitude higher volume of curation actions on Pinterest (See Fig. 1a), followers on Pinterest may see too many repins as spam. Thus, Pinterest users must not only be consistent and regular but must also filter content by curating only the most interesting, in order to attract followers.

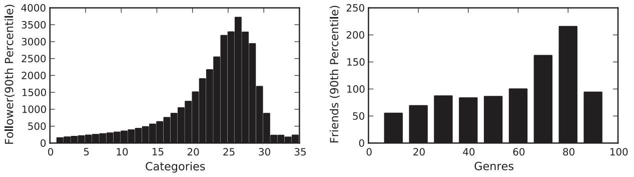


(a) Pinterest

(b) Last.fm

Figure 8: **Structured curation attracts followers when it is consistent and regular** Users with a short interval between successive repins (tags) attract a large number of followers (friends) on Pinterest (Last.fm). A similar result can be obtained for unstructured curation (loves/likes) as well (not shown).

Diversity of interests



(a) Pinterest

(b) Last.fm

Figure 9: **Diversity of interest attracts followers** (a) Pinterest users interested in nearly all the categories attract more followers; (b) Last.fm users interested in most of the music genres tend to attract more followers.

Next, we examine the role of diversity. We capture diversity of a user’s interest in Pinterest by counting the number of distinct categories (of the 32 globally recognised ones) that the user has boards in. Similarly, for Last.fm, we capture diversity of interest by counting the number of genre-specifying tags which have been used for tagging by the

[6] Because a user typically has many intervals between repins, we additionally use a 90th percentile method when selecting this attribute. That is, if a user’s structured curation intervals are represented as a list of intervals, I, this user will be put into a bin according to the 90th percentile value of I.

user. Genre-specifying tags were identified by selecting the top 150 tags on Last.fm and manually filtering out 22 non-genre tags, resulting in 128 distinct genre-specifying tags. Fig. 9 shows that users who have an extremely diverse interest attract a large number of followers. However, beyond a point, the number of followers falls off, for jack-of-all-trade curators who are interested in nearly all categories or genres.

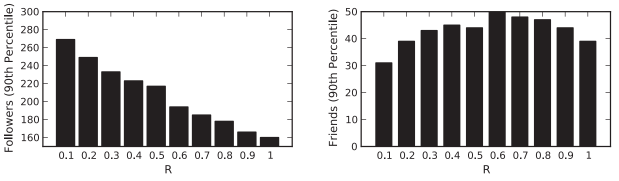
Note that there might be potential confounding factors: For example, being active in a number of categories might simply be a consequence of being more active on the site, and more active users might attract more followers, as shown above. To confirm that our finding about the importance of diversity of interests is not simply an artifact of diversity in usage, we verified that the result of Fig. 9 holds even when we observe limited subsets of users with similar numbers of pins (e.g., 1,000–2,000 pins, or 10,000–20,000 pins).

Structured vs. Unstructured Curation

In previous sections, we discussed structured and unstructured curation, and demonstrated that on Pinterest, most users would prefer to use structured curation, whereas the opposite is true for users of Last.fm. In this section, we try to find out which kind of curation action is more useful for other people.

In Pinterest, as shown in Figure 10, we find that with the increase of unstructured curation ratio R, the numbers of followers decrease. This shows that structured curation (repin) is more useful to others.

However, we do not observe a similar trend in Last.fm. We hypothesise that this is because repinning is the dominant curation method in Pinterest, but tagging is not in Last.fm. On the contrary, as explained previously, Last.fm users are rewarded for 'loving' a track because Last.fm recommends other tracks which might be interesting to the user. Thus, unstructured curation is much more prevalent in Last.fm; even users who tag extensively also use 'love's, increasing their R ratios.



(a) Pinterest

(b) Last.fm

Figure 10: **Structured curation attracts followers** Users with a low unstructured curation ratio R (i.e., those with a large proportion of structured curation actions) tend to attract more followers on Pinterest. A similar trend is *not* seen on Last.fm.

Related Work

Content curation (or digital curation) is an increasingly common phenomenon. A widely accepted definition of digital curation, introduced by the UK Digital Curation Centre, reads as: "Digital curation, broadly interpreted, is about maintaining and adding value to, a trusted body of digital information for current and future use" (Beagrie 2008).

In the Web, the idea behind curation is to link and/or excerpt the work of others (Carr 2012). Therefore, to some extent, the Web has always been about curation (Ovadia 2013), with users sharing links with each other. Many blogs can be considered as curated content, with bloggers sharing links and excerpts with readers. Platforms like Pinterest, Tumblr and Storify make it easier for users to share, showcase, and curate content they discover online. As such, these sites give an opportunity to deeply study the phenomenon of content curation.

Qualitative studies

Some researchers have qualitatively studied the phenomenon of curation on online social networks. For example, (Liu 2010) identifies seven curatorial activities (collecting, organizing, preserving, filtering, crafting a story, displaying, and facilitating discussions) based on an analysis of 100 web artefacts, and introduces the concept of socially-distributed curation, to emphasize the distributed nature of this curatorial process emerging from the social Web. (Rotman et al. 2012) explores the opportunities and challenges of creating and sustaining large-scale content curation communities through a case study of the Encyclopedia of Life (EOL)[7]. A qualitative study among the personnel of a newspaper (Villi 2012) indicates that engaging the audience in social curation is more important than involving the audience in content production.

However, without taking users' behaviour and attitude into consideration, why and how curation happens is still unclear. Therefore, in our user study, the motivation and process of curation are analysed. (Duh et al. 2012) also looks into motivations for curation, by manually inspecting 435 lists of Tweets curated on Togetter.com and identifying seven use cases for curation. Our paper takes a complementary approach, using orders of magnitude more data, and also directly surveying users, obtaining new insights.

Quantitative studies

Online social networks, especially social aggregation websites (e.g. digg, Slashdot, delicio.us and reddit), have supported the process of categorising and sharing content for a few years. Although quantitative studies on structured curation (e.g. tags (Li, Guo, and Zhao 2008)) or unstructured curation (e.g. likes (Sastry 2012)) are common, a comprehensive study of both kinds of content curation has not been carried out until now.

Several dataset-backed studies have used Twitter *lists* as a curation service. For instance, (Garca-Silva et al. 2012; Greene, O'Callaghan, and Cunningham 2012; Kim et al. 2010; Yamaguchi, Amagasa, and Kitagawa 2011) explore users' interest based on list names or through list aggregation. (Greene et al. 2011) proposes a method to identify members for Twitter lists on emerging topics, so that the list could contain the key information gatekeepers and present a balanced perspective on the story. (Ishiguro, Kimura, and

[7] `http://eol.org/`

Takeuchi 2012) assumes that the contents of a curated list are manually organized to fully convey the curators intentions and use contextual features in the curation list to understand images. However, many of these results are specific to the setting of Twitter lists, and cannot be directly extended. Our data-backed study of curation actions could, based on observed characteristics of curation activities, potentially help build similar applications either in the context of Pinterest and Last.fm, or in more contexts as well.

Summary and discussion

This paper used a quantitative analysis of several weeks of curation actions on two different websites, Pinterest.com and Last.fm, combined with user surveys and interviews, to characterise the phenomenon of content curation. First we showed that curation adds value by highlighting a different set of items than traditional methods such as search. Next, we discovered that collectively, the user base of each website focused most of its curation actions on a small number of items, resulting in an extremely skewed distribution of curation activity. This could be seen as evidence of a synchronised community focusing its attention. However, our user studies reveal that the majority of users view curation as a personal activity, rather than a social one. Thus, synchrony may emerge implicitly rather than as a conscious effort of the user base.

We then examined how people curate, and proposed a distinction between structured curation, which highlights an item *and organises it* (by pinning onto a specific board or tagging it) and unstructured curation, which simply highlights an item by liking or loving it. Our data shows that although users differ with some preferring unstructured, and others structured curation actions, popular items invariably see more structured curation activity than unstructured. Using data from Last.fm, we showed that curation tends to happen soon after first contact with an item.

Finally we asked what kinds of curation behaviours attract followers. Our data pointed to at least three factors: consistent and regular curation actions, diversity of interests, and a preference for structured curation (in the case of Pinterest).

References

Beagrie, N. 2008. Digital curation for science, digital libraries, and individuals. *International Journal of Digital Curation* 1(1):3–16.

Bhargava, R. 2009. Manifesto for the content curator: The next big social media job of the future ?

Carr, D. 2012. A code of conduct for content aggregators. The New York Times. Available from **http://www.nytimes.com/2012/03/12/business/media/guidelines-proposed-for-content-aggregation.-online.html**, last accessed 23 March 2013.

Duh, K.; Hirao, T.; Kimura, A.; Ishiguro, K.; Iwata, T.; and Yeung, C. M. A. 2012. Creating stories: Social curation of twitter messages. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Garca-Silva, A.; Kang, J.-H.; Lerman, K.; and Corcho, O. 2012. Characterising emergent semantics in twitter lists. In *Proceedings of the 9th international conference on The Semantic Web: research and applications*, ESWC'12, 530544. Berlin, Heidelberg: Springer-Verlag.

Greene, D.; Reid, F.; Sheridan, G.; and Cunningham, P. 2011. Supporting the curation of twitter user lists. *NIPS Workshop on Computational Social Science and the Wisdom of Crowds*.

Greene, D.; O'Callaghan, D.; and Cunningham, P. 2012. Identifying topical twitter communities via user list aggregation. *arXiv:1207.0017*.

Ishiguro, K.; Kimura, A.; and Takeuchi, K. 2012. Towards automatic image understanding and mining via social curation. In *2012 IEEE 12th International Conference on Data Mining (ICDM)*, 906 –911.

Kim, D.; Jo, Y.; Moon, I.-C.; and Oh, A. 2010. Analysis of twitter lists as a potential source for discovering latent characteristics of users. In *ACM CHI Workshop on Microblogging*.

Li, X.; Guo, L.; and Zhao, Y. E. 2008. Tag-based social interest discovery. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, 675684. New York, NY, USA: ACM.

Liu, S. B. 2010. The rise of curated crisis content. In *Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM 2010)*.

Ovadia, S. 2013. Digital content curation and why it matters to librarians. *Behavioral & Social Sciences Librarian* 32(1):58–62.

Rotman, D.; Procita, K.; Hansen, D.; Sims Parr, C.; and Preece, J. 2012. Supporting content curation communities: The case of the encyclopedia of life. *Journal of the American Society for Information Science and Technology*.

Sastry, N. 2012. How to tell head from tail in user-generated content corpora. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*.

Shirky, C. 2010. Talk about curation. Published as on online interview with Steve Rosenbaum. Available from **http://curationnationvideo.magnify.net/video/Clay-Shirky-6**, last accessed 15 Feb 2013.

Villi, M. 2012. Social curation in audience communities: UDC (user-distributed content) in the networked media ecosystem. *Participations: The international journal of audience and reception studies, Special section: Audience Involvement and New Production Paradigms* 9(2):616–632.

Yamaguchi, Y.; Amagasa, T.; and Kitagawa, H. 2011. Tag-based user topic discovery using twitter lists. In *2011 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 13–20.