Social Curation of Content
Measurements and Models

Zhong, Changtao

Awarding institution:
King's College London

Social Curation of Content: Measurements and Models

Changtao Zhong
钟常涛

Submitted for the degree of
Doctor of Philosophy

Department of Informatics
King's College London

December 2016

Abstract

Social curation is a new trend which has emerged following on the heels of the information glut created by user-generated content revolution. Rather than create new content, social curation allows users to categorise content created by others, and thereby creating and resharing their personal taxonomies of the Web. In this dissertation, we collect a large dataset from Pinterest, arguably the most popular image curation service, and seek to understand the trend on three levels: content, friends and crowds.

We first take an empirical look at social curation by mining its content usage. Our data reveals that curation tends to focus on niche items that may not rank highly in popularity and search rankings. Yet, curated items exhibit their own skewed popularity, although most users, or curators, act for personal reasons. At the same time, it also shows that curators with consistent activity and diversity of interests show more social value in attracting followers.

This drives us to explore the role of social networks on social curation. We find that social users are more active and are more likely to return soon in Pinterest, indicating a bonding effect enabled by social networks. Then we divide the social network into two subgraphs, according to whether they are created natively or copied from some other established social networks (e.g., Facebook) via a *social bootstrapping* method. It shows that, when users just join the service, copied network can promote more social interaction, as it initiates a stronger and denser social structure than native network. However, social networks are *not* critical for information seeking, as a non-trivial number of users' content are curated from strangers with high interest matching. In fact, this trend also holds for social interaction: Users tend to wean from copied friends to interact more with interest-based native friends over a long-term view.

Finally, we understand social curation as a *distributed computation process*, and examine the relationship between curators and crowds. We show that despite being categorised by individual actions, there is generally a global agreement in implicitly assigning content into a coarse-grained global taxonomy of categories, and furthermore, users tend to specialise in a handful of categories. By exploiting these characteristics, and augmenting with image-related features drawn from a state-of-the-art deep convolutional neural network, we develop a cascade of predictors that together automate a large fraction of curation actions with an end-to-end accuracy of 0.69 (Accuracy@5 of 0.75).

Acknowledgements

This dissertation is the outcome of four years of extensive yet exciting and instructive exploration. I would like to take this opportunity to express my gratitude to all who have helped me to complete my dissertation.

First of all, I would like to thank my supervisors Nishanth Sastry and Meeyoung Cha for their guidance, encouragement and support throughout my PhD journey. I especially appreciate Nishanth. This dissertation would not have been possible without him and without the freedom and encouragement he has given me over the last four years I spent at King's.

Subsequently, I would like to express my gratitude to my colleagues in the NetSys group. Dmytro Karamshuk has been of untold support. My work with machine learning and deep learning would not have been that familiar to me without him. Thanks also to Aravindh Raman and Sagar Joglekar, for the infinite tea cups we consumed together. Menglan Jiang, Shuyu Ping, Zhenzhuang Miao and rest of people at CTR, with who I have supportive and stimulating conversations on a wide range of topics that made the lab not only a good place to work, but also a fun place to be.

I was lucky to have received great support from coauthors and collaborators outside King's. Each of them deserves my gratitude: Hau-Wen Chang (IBM), Marius Cobzarenco (UCL), Oana Goga (MPI-SWS), Krishna Gummadi (MPI-SWS), Nicolas Kourtellis (Telefonica), Dongwon Lee (PSU), Mostafa Salehi (University of Tehran), Sunil Shah (Mesosphere), Karthik Sundaravadivelan, Giridhari Venkatadri (MPI-SWS), Bimal Viswanath (MPI-SWS). I also thank to Miriam Redi, Diego Saez-Trumper, Neil O'Hare and Alejandro Jaimes for the great time at Yahoo Lab.

I owe a HUGE thanks to my best friend Jun, who was the first one suggested me to apply for this PhD. Without his encouraging, supporting and helping, it is impossible for me to be here. Thanks for being at the ends of a phone whenever needed.

Finally, but not least, I thank my parents, Lunshen Zhong (钟伦胜) and Shanxi Li (黎善喜): 虽然我一心想要像风筝一样飞上天空，您们的鼓励和支持才是支撑我翱翔蓝天的强大力量。谨以此文献给您们。

Table of Contents

List of Figures

List of Tables

無名天地之始；有名萬物之母。
故常無欲，以觀其妙；常有欲，以觀其徼。
此兩者，同出而異名，同謂之玄。
玄之又玄，眾妙之門。

——《道德經》老子
（“Tao Te Ching” or “Dao De Jing”, Laozi）

Chapter 1

Introduction

> Some of the most crucial steps in mental growth are based not simply on acquiring new skills, but on acquiring new administrative ways to use what one already knows.
>
> *Marvin Minsky*

1.1 Social Curation

Social curation corresponds to a relatively recent advancement in the space of the World Wide Web and, more broadly, Information Industry. It has emerged following on the heels of the information glut created by the user-generated content revolution. Rather than create new content, social curation services allow their users to categorise and organise collections of content created by *others* that they find online. These users provide an editorial perspective by adding context information or highlighting interesting content. For instance, on Pinterest[1], perhaps the most prominent social curation site, users can collect and categorise images (and the URLs of the webpages that contain them) by "pinning" them onto so-called

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

"pinboards". Similarly, Delicious[2] (formerly del.icio.us) allows users to categorise URLs by tagging them.

Content curation itself is not a new phenomenon. The concept is derived from the term "curator", a professional in museums and galleries, who is in charge of selecting and caring for objects that form collections or exhibitions [Pearsall, 1998]. In the library sector, curation represents the process of adding value to objects from collection-building around them, from the documentation which provides the relevant context or history of them, or from the knowledge of the curators [Beagrie, 2008]. For scientists and researchers, curation includes processes of annotation, linkage, validation and publishing research databases (e.g., human genome). Thus, *content curation is the activity of categorising or organising content created by others into collections, and thereby providing their own editorial views.*[3]

Social curation extends the concept into more expanded roles: The object of curation is not just art works or research data, but could also be any types of online content (text, image or video); at the same time, the subject of curation (i.e., curator) is extended from a few experts to everyone or *crowds*. Nowadays, millions of users around the world save and generate collections of content items at social curation platforms. The scale of these collections in terms of numbers of objects, curators and audiences is unprecedented. For example, a billion URLs from over 200 countries are curated to Delicious everyday [Delicious.com, 2016]. On Pinterest, there are more than 50 billion images curated onto more than 1 billion pinboards by March 2015 [Pinterest.com, 2015].

Therefore, a fundamental problem of social curation is how to exploit and utilise the massive curation activity data generated by millions of curators everyday. It is thought that social curation is a solution to the information overload problem created by the user-generated content revolution [Shirky, 2008; Grineva and Grinev, 2012; Anderson, 2015]. Indeed, millions of curators discover, select, organise and share content via social curation everyday, and thereby creating their personal taxonomies of the Web. These taxonomies are valuable, as they can be used to improve existing algorithm-based methods, such as search engines and recommender systems, and may also inspire new algorithms, although we are still far from achieving that.

[2] http://www.delicious.com

[3] We will provide a formal definition for content curation in §2.1

In the meantime, the exploitation of social curation datasets also offers a good opportunity for better understanding the concept of content curation, which is multi-disciplinary in nature and stretches from fields in the social sciences such as education, sociology and communication, to areas of computer science. In general, there are three types of interaction for curators in social curation services:

- **Interaction between curators and content**, including the content discovery, selection, organisation and sharing process;
- **Interaction between curators and friends**, including relationships with content creators, audiences and other curators;
- **Interaction between curators and crowds**, which indicates the agreement between one curator and others who are also interested in their personal content collections.

Empirical data on curation activity sourced from the new generation of social curation services can be helpful to compare and validate models and theories that have been enabled by scientists. For example, understanding the interaction between curators and content is useful for information retrieval, whileas curators and friends interaction is related to the study for the role of social networks on information seeking.

Despite the extensive literature published that attempts to explain human online activity patterns [Gilbert et al., 2013; Ottoni et al., 2013] and curation principles [Shirky, 2010; Bhargava, 2009a], the study to the social curation data is still at very early stage. Some basic questions, such as why and how people curate online, have not been answered yet, not to say big questions like how to utilise it to solve information overload problem. To understand the phenomena, from Pinterest, we collect one of the largest social curation datasets in the world which involves tens of millions of curators, billions of friendship links, and millions of curation activity records. With the dataset, we study social curation from interactions between curators and three different objects: content, friends and crowds. In the following sections, we will discuss these three interactions in detail, and introduce research questions of the dissertation.

1.2 Curators and Content Interaction

Interaction with content is a basic function provided by social curation platforms, and the major motivation of users to join them. As discussed before, social curation platforms allow users or curators to *discover*, *select*, *organise* and *share* content they are interested. In this dissertation, we will mine curation data generated from each of these curation steps, seeking to understand their usage patterns and provide a insight for the motivation of curators.

Content discovery and selection First, we seek to capture curators' motivation signals indicated by their content discovery and selection activity. Social curation is not the only service that assists users with content exploitation. For instance, search engines, that aggregate content items from all around the Web and then automatically present users the most relevant ones, are also widely used for content discovery. However, according to a popular theory of Shirky [2010], "curation comes up when search stops working". Thus, an interesting angle to understand social curation is to compare it with these existing content discovery methods, and to check whether there are any differences between them that motivate users to join social curation services.

A second aspect of Shirky [2010]'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". This can also be explored with the content discovery and selection data. In this study, we will analyse the data and seek to examine whether curators are talking about the same thing, and whether there is a synchronized community among them.

Content organisation. In most social curation websites, multiple curation actions are available to a user. For instance a user on Pinterest can pin an item, like it or comment on it. Similarly, on Delicious, URLs can be saved and tagged. We note that these actions can be distinguished based on whether they simply highlight an item (like, save, comment), 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 link). We term the former as *unstructured curation* and the latter *structured curation* because of the organisational structure induced by pinning or tagging.

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

Content sharing. Finally, we will study social value of social curation by examining curators' sharing activity, as sharing different kinds of content will attract different numbers (and types) of followers or audiences. Bhargava [2009a], who appears to have coined the term content curator, defined content curators 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.' " We will analyse users' content sharing data to validate the theory and provide insights for curators' social value.

In summary, we will study the curator-content interaction by examining the curation activity data collected from Pinterest (our **DATA-ACT** dataset, refer §3.2.1 for more details), and seek to answer following questions:

RQ 1 (Curators vs. Content) *Why do people curate? How do they curate? And what do others, namely followers of curators, find useful?*

1.3 Curators and Friends Interaction

Next, we discuss the interaction between curators and their friends. It has become *de rigueur* to create social networks amongst users on all kinds of Web 2.0 sites. Many websites now try to incorporate a social networking aspect to enhance user engagement and create active communities. Making a website "social" typically involves linking users together and providing some kind of awareness of the linked users' activities to each other. Social curation services are not exceptions. Most social curation websites (e.g., Pinterest, Delicious and Tumblr[4]) incorporate social features such as the ability to follow and react to the content shared by their friends.

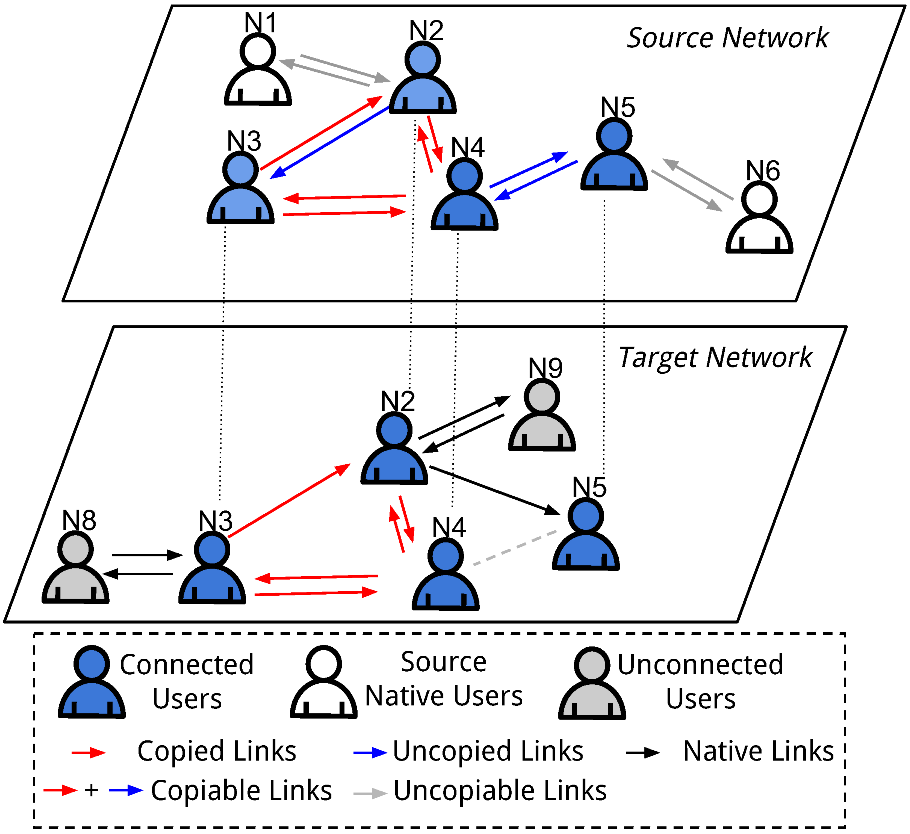
However, the role of social networks for content-driven services, such as social curation services, is still unclear. A key feature of social networks on content-driven websites is that links are intended to be made based on shared interests around an item or a category of items [Hendricks, 2014; Jamison, 2012]. However, recent research has shown that a majority of users do not participate in social aspects of content [Gelley and John, 2015] or product [Swamynathan et al., 2008] sharing websites. Further, many users with explicit friendship or follow links may not in fact have any content that they like in common [Gelley and John, 2015; Musial and Sastry, 2012]. These findings stand in direct contrast to numerous studies where social networks have been shown to help in community formation, in a diverse range of interest- and goal-oriented environments and applications such as learning [Baird and Fisher, 2005; Conole and Culver, 2010; Heiberger and Harper, 2008] , working [DiMicco et al., 2008; Lee et al., 2013] , medicine [Eysenbach, 2008] and online games [Choi and Kim, 2004; Ducheneaut and Moore, 2004] . In light of these conflicting results, we collect the complete activity history of 50 thousand of Pinterest users (our **DATA-USER** dataset, refer §3.2.3), and ask:

RQ 2 (Curators vs. Friends) *What is the value of social networks on social curation services?*

Another fundamental problem for the designers of social services is how to design online communities and maintain users participation. In creating a social experience on a website, designers face an important choice: should they create an entirely new social network embedded within the site? Or should they instead connect users who are already linked together on an established social network such as Facebook[5] and Twitter[6]? The latter option has recently become a possibility, with both Facebook and Twitter opening up their social graphs to third-party websites, who can write friend-finder tools that help users select and import friendship

[4] https://www.tumblr.com
[5] https://www.facebook.com
[6] https://www.twitter.com

Figure 1.1
The Structure of Social Bootstrapping



links from these established networks into their own service (e.g., through the open graph protocol [Facebook.com, 2016b]).

We term this act of copying existing friends from an established social network onto a third-party website as *social bootstrapping*, because this enables the third-party website to bootstrap social links from a mature social networks. Figure 1.1 is a toy example of social bootstrapping, where users copy links from the source network (such as Facebook) into the target network (such as Pinterest). Social bootstrapping has direct implications on how a new online social network community can grow quickly. However, this problem is complex to examine with real data because it involves user interaction across multiple heterogeneous networks. To this end, we gather massive amounts of data from Facebook and Pinterest involving tens of millions of nodes and billions of links as our **DATA-SOC** dataset (refer §3.2.2) and explore the benefits and limitations of social bootstrapping. We seek to evaluate how such bootstrapping could affect the user community and to what extent copying links contributes to social structure and user engagement as the new website matures.

RQ 3 (Social Bootstrapping) *What are the benefits and limitations of social bootstrapping?*

1.4 Curators and Crowds Interaction

Curators do not only interact with content and their friends in social curation website. Actually, a crucial aspect of social curation platforms is that content curated by one user is also (by default) made available to the rest of the users or *crowds* to curate. Thus, a curator also indirectly interacts with the crowd of other users on the site. For instance, on Delicious, links of another user can be copied onto one's own list, by tagging them. Users on Pinterest can copy images pinned by other users, and "repin" onto their own pinboards. Interestingly, such reappropriation and curation of content discovered by other users (termed as "repins") is by far the most common activity on Pinterest, constituting about 90% of user actions, as compared to directly discovering and pinning new images, which constitutes only 10% of actions.[7]

Even if curating a content item that has already been categorised, users typically have to re-adjust it for their own collections: For instance, in tagging systems, multiple studies [Golder and Huberman, 2006; Rader and Wash, 2008; Noël and Beale, 2008] have recognised that inter-user agreement can be low, and different users may choose different tags for the same URL (although an overall consensus vocabulary may emerge per-URL, due to tag imitation or background knowledge [Golder and Huberman, 2006]). Similarly, users' pinboards are highly personal and idiosyncratic represntations of their taste, and furthermore, users are free to choose to repin any image onto any pinboard. Consequently, curation on Pinterest remains a highly manual process as well. Based on our **DATA-IMG** dataset, we examine the agreement between curators and crowds and ask whether we could make it easier to re-appropriate and re-categorise content for personal use. That is, consider an image which has been introduced into

[7] As observed in our **DATA-ACT** dataset.

Pinterest by an original pinner and several other users may want to repin this onto their own pinboards, we are interested in examining:

> **RQ 4 (Curators vs. Crowds)** *What is the degree of agreement of curators' personal curation actions over crowds? And what is the predictability of those personal curation actions?*

1.5 Contributions and Chapter Outline

The contributions of this dissertation are threefold: First, we collect a large social curation dataset from Pinterest and share it with the research community. Then, we do large-scale empirical studies for the content and social network structure with our dataset, and find evidence for the motivations of curators and the value of social networks. Finally, we treat curation as a distributed computation process and use a machine learning approach to automate social curation, thereby obtaining an understanding of the end-to-end process of social curation on Pinterest.

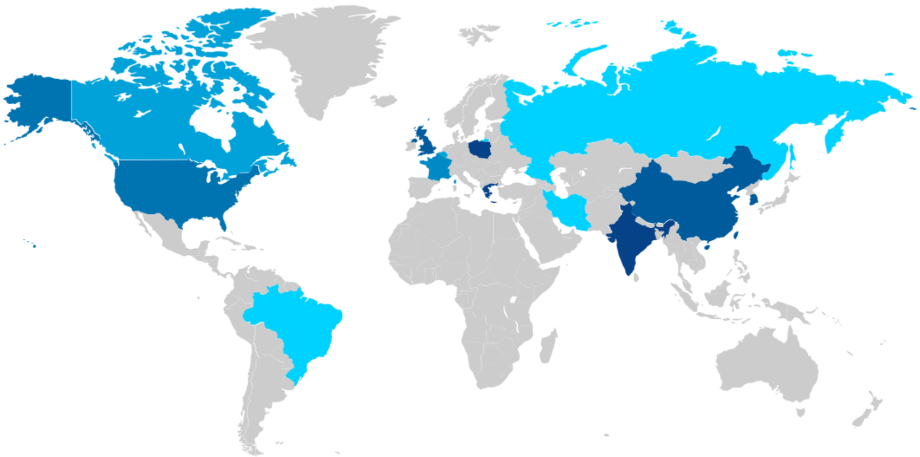
In addition to present a comprehensive understanding of social curation, this dissertation also shows a methodology of exploring online phenomenon with massive social data, from collecting data to providing insights, from empirical analysis to predicting future activity.

More specifically, the dissertation is organised as follows:

- In **Chapter 2**, we introduce the background and related work of the dissertation. First, we introduce a definition of content curation, and discuss its three properties: aggregation, organisation and reusability. Then, we introduce social curation, which is a new type of curation that expands curator from experts to crowds. We also summarise related work to our four research questions in this chapter.

- In **Chapter 3**, we introduce the terminology and dataset used in the dissertation. We first introduce a popular social curation service, Pinterest, and curation actions it provides. Then, we describe the process of our data collection from Pinterest. The social curation dataset is divided into four parts: **DATA-ACT** for content related empirical study, **DATA-SOC** and **DATA-USER** for social net-

Figure 1.2
Sharing the Pinterest Dataset



We have shared our dataset with researchers from over 70 research institutes of 15 countries until may 2016. The figure is generated using Matador Network (http://matadornetwork.com).

work related measurement research and **DATA-IMG** for curators and crowds interaction study. We have made the dataset available to research community for wider use[8], and it has been requested by researchers from over 70 different institutes (from 15 countries, see Figure 1.2) till May 2016.

- In **Chapter 4**, we explore the interaction between curators and content via our **DATA-ACT** dataset, seeking to understand the how and why of social curation (i.e., RQ1). We find that social curation platforms highlight niche content that can hardly be found through traditional ranking methods, such as search or website traffic ranking. Yet, curated items exhibit their own skewed popularity, although most of curators act for personal reasons. At the same time, curators with consistent activity and diversity of interests show more social value in attracting followers.

- In **Chapter 5**, we explore the value of social bootstrapping in social community construction (RQ3) and the role of social networks in social curation services (RQ2) with **DATA-SOC** and **DATA-USER** datasets. We find that social users are more active and are more

[8] http://www.inf.kcl.ac.uk/staff/nrs/projects/cd-gain/dataset.html

likely to return soon to social curation platforms, indicating a bonding effect enabled by social networks. But which type of social networks are more useful for user interaction, the network copied via social bootstrapping or the one created natively? Our empirical analyses indicate that copied network shows an more important role in promoting social interaction when users just join the platform, as it initiates a stronger and denser social structure than native networks. However, social networks also have limitations. We find that social networks are *not* critical for information seeking in social curation services, as a non-trivial number of users' content are curated from strangers. This can be explained by the fact that users curate based on their interest rather than friendship. A similar trend can also be found from users' interaction with their friends. That is, as users become more active and influential, users tend to wean from copied friends to have more interactions with interest-based native friends.

- In **Chapter 6**, we explore our RQ4 and ask to what extent we could reproduce the content curation automatically. We consider online content curation as a *distributed computation process* and find that there is a global agreement across all curators for the curation, despite most of curation actions occurring for personal reasons. At the same time, users tend to specialise in a handful of types of content. By exploiting these characteristics and augmenting with image-related features drawn from a state-of-the-art deep convolutional neural network, we develop a cascade of predictors that together automate a large fraction of curation actions with an end-to-end accuracy of 0.69 (Accuracy@5 of 0.75).

- Finally, in **Chapter 7**, we summarise the findings and identify directions for future work in social curation and its applications.

1.6 List of Publications

Some of the research related to this thesis has been published (or is under review) in various peer-reviewed conferences and journals. These publications are as follows:

Chapter 4:

- C. Zhong, S. Shah, K. Sundaravadivelan, and N. Sastry. Sharing the Loves: Understanding the How and Why of Online Content Curation. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, ICWSM, 2013.

Chapter 5:

- C. Zhong, M. Salehi, S. Shah, M. Cobzarenco, N. Sastry, and M. Cha. Social Bootstrapping: How Pinterest and Last.fm Social Communities Benefit by Borrowing Links from Facebook. In *Proceedings of the 23rd international conference on World Wide Web*, WWW, 2014.
- C. Zhong, N. Kourtellis, and N. Sastry. Pinning Alone?: A Study of the Role of Social Ties on Pinterest. In *Proceedings of the 10th International AAAI Conference on Weblogs and Social Media*, ICWSM, 2016. (Poster Paper).

Chapter 6:

- C. Zhong, D. Karamshuk, and N. Sastry. Predicting Pinterest: Au tomating a Distributed human computation. In *Proceedings of the 24th international conference on World Wide Web*, WWW, 2015.
- C. Zhong, D. Karamshuk, and N. Sastry. Automated predictive curation of items, May 17 2015. US Patent Application No. 62/162778.

Others:

- C. Zhong and N. Sastry. Copy content, copy friends: Studies of content curation and social bootstrapping on pinterest. *SIGWEB Newsletter*, (Summer):4:1–4:6, July 2014.
- C. Zhong and N. Sastry. Systems Applications of Social Networks. *ACM Computing Survey*, 2016. (Under Review).
- C. Zhong, M. Chang, D. Karamshuk, D. Lee, and N. Sastry. Wearing many (social) hats: How different are your different social network personae?. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing* , CSCW, 2017. (Under review).

- G. Venkatadri, O. Goga, C. Zhong, B. Viswanath, K. P. Gummadi, and N. Sastry. Strengthening Weak Identities Through Inter-Domain Trust Transfer. In *Proceedings of the 25th international conference on World Wide Web*, WWW, 2016.

Chapter 2

Background and Related Work

Curate *verb*
Select, organize, and look after the items in (a collection or exhibition).

- Derivatives **Curation** *noun.*
- Origin late 19th century: back-formation from **Curator** (*a keeper or custodian of a museum or other collection*).

(Definition from the New Oxford Dictionary of English, 1998)

Chapter layout. In this chapter, we will survey papers related to the dissertation, and give an overview of research on content curation. First, in §2.1, we discuss content curation activity. We will provide a definition for content curation and discuss its properties. Then, in §2.2, we focus on social curation, which allows everyone to be curators, and compare it with other types of curation. To this end, in the rest of this chapter, we present a literature review around our research questions proposed in §1. In §2.3, we review studies related to our RQ1, i.e., the empirical studies of online content curation activity. In §2.4, we survey studies related to the role of social networks and social bootstrapping which are related to our RQ2 and RQ3, respectively. Finally, we present studies related to the curator-crowds interaction and the predictability of social curation for our RQ4 in §2.5.

2.1 The Definition of Content Curation

Curation lately made its appearance in lots of articles of dealing with content marketing. But what is curation? To answer this question, we start from a professional named "curator".

From its Latin etymology, curator is someone who "takes care" [Pearsall, 1998]. Expending this, curator is "someone who takes an inordinate mass of material" and "turns noise into signal" [Guerrini, 2013]. It is traditionally used to describe a type of museum professional (which we term "museum curator" to distinguish from curators in other fields). Chambers [2006] examined the job descriptions, job qualifications and personal statements of more than 200 museum curators and wrote an article named titled "Defining the Role of the Curator". According to the article, curators are vehicles for exploring and interpreting some aspects of culture by building collections, managing collections, conducting academic research, developing exhibitions and creating educational programs. There is a general overview of the duties of museum curators in the 2016-17 Occupational Outlook Handbook published by U.S. Department of Labor:

> "Curators direct the acquisition, storage, and exhibition of collections, including negotiating and authorizing the purchase, sale, exchange, and loan of collections. They may authenticate, evaluate, and categorize the specimens in a collection." [Bureau of Labor Statistics, 2015]

Therefore, museum curators are behind art exhibitions. Typically, they have four jobs [Guerrini, 2013]: *selecting* the best or most representative items (paintings, sculptures etc.), *verifying* the origin and authenticity of them, *organising* them with a certain way - maybe in chronological order, a certain theme or other criteria - and *presenting* them to the audience in the most effective way. Thus, curators do not just present the individual artistic items, they also add value to items by building themed collections and providing contextual documentation [Beagrie, 2008].

In recent years, the concept of "curation" has become a buzzword in online content marketing. The idea shifted from "museum curator" to "digital curation", "content curation" and "social curation".

The term digital curation is introduced by librarians to convey the concept to creating and managing digital assets in the age of information glut. A widely accepted definition was introduced by the UK Digital Curation Centre, which reads as follows:

> "Digital curation, broadly interpreted, is about maintaining and adding value to, a trusted body of digital information for current and future use" [Beagrie, 2008].

The definition emphasises five points:

- The word "maintaining" shows that the information needs to be preserved for future access.
- The phrase "adding value to" emphasises that the activity is beneficial for the original information.
- The word "trusted" stresses that the information need to be verified and is trustable.
- "Digital information" shows that the object of this curation is digital (not physical) content.
- The phrase "for current and future use" is the aim of curation activity, and emphasises the accessibility and adding value again.

However, there are several aspects of the definition that are not clear. For example, the output of digital curation (usually is collections) and basic processes has not been mentioned by the definition.

In this dissertation, we adapt and extend this definition from "digital curation" to "content curation" as follows to cover more curation activities:

> **Content curation** is about discovery, selecting and organising (physical or digital) trusted content items into collections for current and future use.

In the definition, we emphasise that a typical curation process includes activity of discovery, selecting and organising. The curation output usually is "collections". The target of this activity is "for current and future use".

2.1.1 Properties of content curation

Next, we discuss the concept in detail. We summarise three major properties of content curation: aggregation, organisation and reusability.

Aggregation. From our previous discussion, we see the initial interpretation of curation involved aggregation, as the major job of curators is to take care of collections, which entails the aggregation of items and documentation. It is worth noting that curators themselves usually do not create content. Instead, they provide ideas by aggregating items, as stated by Bhargava [2009b] in a blog titled "How To Use Curation To Make Your Blog Better: Lessons From Postsecret":

> "Curation evokes that powerful idea of working on something larger than yourself. Museum staff curate the works of art and historical significance that line their walls. National archives that store the lessons of the world's past are similarly curated."

For example, one of the first curation-based news website, the Huffington Post[1] is working on a strategy of aggregating information from lots of different sources together. It does not reply on a specific third-party for the content. Instead, the website presents a mix of original content made by professional journalists, posts written by thousands of (unpaid) bloggers and popular contetn from Facebook, Twitter or other social networks [Guerrini, 2013].

Organisation. However, aggregation is not the end of curation. A lot of online platforms also support content aggregation, such as search engines (e.g., Google[2] and Bing[3]) and RSS readers (e.g., flipboard[4] and feedly[5]),

[1] The Huffington Post launched in 2005, with its US version as http://www.huffingtonpost.com, and UK version as http://www.huffingtonpost.co.uk.

[2] https://www.google.com

[3] https://www.bing.com

[4] https://www.flipboard.com

[5] https://www.feedly.com

but they are not curation tools (in our definition). Shatzkin [2009] compares aggregation and curation and states that “aggregation without curation is, normally, not very helpful; curation creates the brand.” Actually, curation does not just aggregate content from different sources, a more important part is curators showing their own ideas via the aggregation. Liu [2011] states:

> “A well-curated collection aggregates content . . . may have not been initially related or connected; . . . it is the aggregation of disparate items that can provide new insights that were not visible before.” [Liu, 2011, p. 204]

In addition to putting items together, curators provide their editorial views by filtering, categoriting and organising content, as stated by Givhan [2007]:

> “Museums don’t own culture, but they sort through it, rank it and attempt to make some sense of it. Theirs aren’t the only valid points of view, but they are especially valued because they’re the result of research, dispassionate analysis and intellectual curiosity.”

Another example is journalists. Guerrini [2013] considers the job of journalists as curation, with the reporter discovering and selecting sources into a draft of content and the editor assembling and reshaping the content. They also need to filter, organise and re-write original information into news articles (which can be considered as news collections), but, more importantly, they add their editorial perspectives and make the content more understandable for readers. As Liu [2011] states, curators have “the ability or tools to extract the significance of the aggregated collection of content so that consumers can easily digest and make sense of it.”

Reusability. Here, the reusability of content curation is twofold: reusability for *the original content* and for *the output collections*.

At first, the reusability requires the original content being well preserved. For the physical items, it requires “curators have the ability to restore and reconstruct artifacts to maintain their physical quality and ensure their authenticity and security” [Liu, 2011]. Reusability is also important for digital content, because it may be removed by the uploader or due to copyright infringement and some other reasons. For example, SalahEldeen and Nelson [2012] analysed six event-centric datasets from social media from 2009 to 2012, and found that nearly 11% of shared resources would be lost. Thus, curators should have ability to backup and maintain curated information.

More importantly, the reusability also requires that the audience can access the output collections. It is a difference between content curation tools and other tools [Guerrini, 2013]. For example, in Twitter[6], one may collect all pictures, videos, links and other user generated content posted online related to an event, and by tweeting them in a sequence. But when the event is over, it would be very difficult to reconstruct the event, since Twitter privileges things happening in real time. However, if those content items are curated to some curation tools like Tumblr or Pinterest (refer §3.1), one could create different collections for different events by adding tags or creating so-called pinboards, so that those collections will still be reusable to the audience after the event.

Other properties. In addition to above three properties, content curation also have some other properties. For example, journalists have to perform various checks and investigations to make sure that the content found from social media or other online platforms is genuine and not an elaborated forgery made up by someone trying to promote his own agenda [Guerrini, 2013]. Although verification is an important properties in journalistic curation, it is not so critical for curators in Pinterest and Tumblr, as most of them curate for their personal interests [Zhong et al., 2013; Linder et al., 2014].

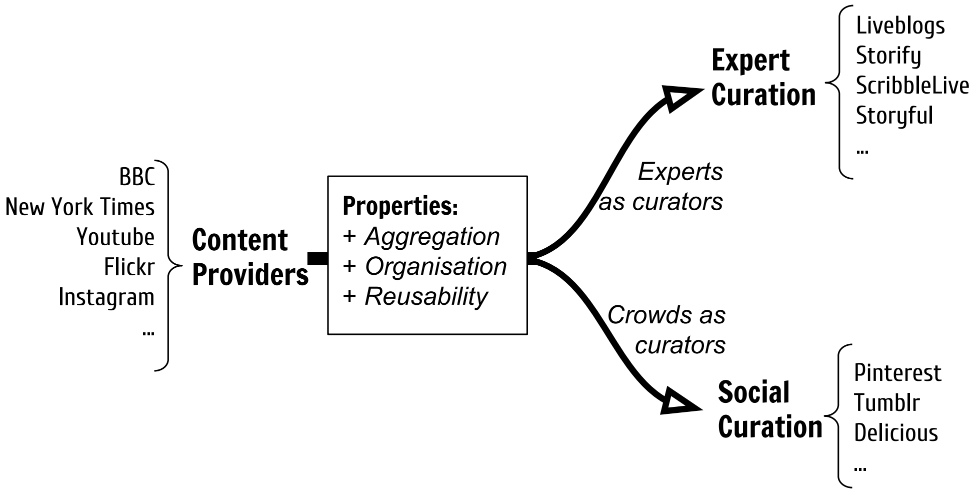
2.2 Social Curation

In previous section, we have defined content curation, and discussed some of its major properties. Now, we focus on a new trend emerged recently on the Web, social curation, and compare it with other online curation activity.

In Figure 2.1, we compare several online platforms: content provider, expert curation and social curation. As we have discussed in previous sec-

[6] Twitter (`https://www.twitter.com`) is a popular online microblogging service that allow users to send and read short 140-characters messages. Messages at Twitter are called “tweets”, and the action of posting messages are called “tweeting”.

Figure 2.1
Content Provider, Expert Curation and Social Curation



tion, content curation platforms usually do not create new content; most of content are imported from content providers, which can be traditional content providers (e.g., BBC[7], New York Times[8]), or user-generated content platforms (Youtube[9], Flickr[10], Instagram[11]). Online content curation platforms give the power of aggregating, organising and reusing imported content to their users. In general, we can divide online content curation platforms into two types according their major users: for experts and for crowds.

Expert curation sites. Curators are traditionally considered as experts of some specific fields. There are a lot of online tools are built for their curation activities, especially for journalistic curation. *Livebolgs* are the major tools used by news organisations. A liveblog is a single post made up of short micro-updates (which could consist of text, pictures, videos, links or other items such as tweets) and constantly updated by one or more authors [Guerrini, 2013]. For example, when the Paris terror attacks happened in November 2015, the Guardian created an liveblog[12] for the event. It is a popular method journalistic curation method now. The Guardian alone publish on average 146 Liveblogs a month [Thurman

[7] https://www.bbc.co.uk
[8] https://www.nytimes.com
[9] https://www.youtube.com
[10] https://www.flickr.com
[11] https://www.instagram.com
[12] http://www.theguardian.com/world/live/2015/nov/14/paris-terror-attacks-live-news-updates-isis-france

and Walters, 2013]. Although BBC and the Guardian have built their own internal platforms, some other news organisations, such as Reuters[13], The New York Times and Bloomberg[14], reply on external platforms like ScribbleLive[15] and Storyful[16] [Guerrini, 2013].

Social curation sites. Social curation is a new trend that has emerged in the social media landscape. In the age of information glut, "no one entity — be it a wire service or a news organization — can possibly track what's appearing in all papers, large and small, blogs, magazines and Web sites; nor could any one news organization acquire the rights to all of this material." [Korr, 2008] To solve the problem, Korr emphasises that we need to rely on the rise of collective intelligence. Social curation is a such system. In the system, the curation activity becomes a distributed and crowdsourced process. It also enable a wide social network to make sure that unconnected curators can still share their knowledge [Liu, 2011].

Instead of relying on experts to curate content, social curation sites give the power to the crowds by allowing their users to import external content, to organise those imported content into collections, and share the collections with the audience. Pinterest and Delicious are two popular social curation websites. Pinterest attracted images and videos from all kinds of websites around the world. For example, in our **DATA-ACT** dataset (refer §3.2.1), we find images from more than 300K websites have been curated to Pinterest in just two weeks. Delicious, a URL curation website, attracted a billion URLs from over 200 countries everyday [Delicious.com, 2016].

2.3 Empirical Studies of Content Curation

We have discussed content curation and social curation in previous two sections. In the rest of this chapter, we will review literature related to our research questions. In this section, we focus on studies on the empirical studies of online content curation activity, as they are related to our RQ1.

Content curation is an increasingly common phenomenon. In the Web, the idea behind curation is to link and/or excerpt the work of others [Carr,

[13] http://www.reuters.com

[14] http://www.bloomberg.com

[15] http://www.scribblelive.com

[16] http://www.storyful.com

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 Delicious 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.

2.3.1 Qualitative studies

Some researchers have qualitatively studied the phenomenon of curation on online social services. 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] explore the opportunities and challenges of creating and sustaining large-scale content curation communities through a case study of the Encyclopedia of Life (EOL)[17]. 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. Duh et al. [2012] also look into motivations for curation, by manually inspecting 435 lists of Tweets curated on Togetter.com[18] and identifying seven use cases for curation.

There have also been some qualitative studies on Pinterest recently. For example, Hall and Zarro [2012] analyse comments on Pinterest, and find evidences for four types of motivations: sharing comments & judgment, engaging in dialog, sharing a personal history with the image and providing additional narrative details. They also show that activities in Pinterest are mainly content-based, and the major topic is on information use, reuse and creation. Zarro et al. [2013] identify four types of curation activities on Pinterest: discovery, collecting, collaborating, and publishing. Linder et al. [2014] consider those curation activities as processes of everyday ideation, i.e., users explore interesting content and develop ideas from them for shaping their lives.

[17] http://www.eol.org
[18] http://www.togetter.com

However, without taking users' actual behaviour and actions into consideration, how curation happens is still unclear. Therefore, when we explore answers for our RQ1 in §4, both motivation and the actual action of curation are analysed. We use orders of magnitude more data to obtain new insights.

2.3.2 Quantitative studies

Online social networks have supported the process of categorising and sharing content for a few years. Although quantitative studies on curation actions like tags [Li et al., 2008] and likes [Sastry, 2012] are common, a comprehensive study of both kinds of content curation has not been carried out until now. In §4.2, we will compare and discuss the differences of those curation actions.

Several dataset-backed studies have used Twitter lists[19] as a curation service. For instance, [García-Silva et al., 2012; Greene et al., 2012; Kim et al., 2010; Yamaguchi et al., 2011] explore users' interest based on list names or through list aggregation. Greene et al. [2011] propose 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 et al. [2012] assume 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.

With the popularity of Pinterest, there are also some quantitative studies on online social curation websites. For example, Ottoni et al. [2013] and Gilbert et al. [2013] examine the effect of the gender of curators to their activity pattern. They find that females tend to be more active for curation actions and social interactions, but males are more likely to specialise on a few types of content that reflects their personal taste. Ottoni et al. [2014] compare the temporal activity pattern and linguistic usage at social curation websites (e.g., Pinterest) and microblogging websites (e.g., Twitter). Bakhshi and Gilbert [2015] find the colour of images may affect their popularity in Pinterest.

[19] Twitter list is a curated group of Twitter *accounts*. Users can create their own lists or subscribe to lists created by others [Twitter.com, 2016].

Most of these existing literature focus on the activity patterns of social curation. In §4, we will also do a large-scale quantitative study for the activity pattern of curators at Pinterest. But we will take a different angle by identifying two types of curation actions: structured and unstructured curation. We will also dig further into the source of content, and seek to understand why people use social curation services.

2.4 Studies on Social Networks for Curation

Next, we do literature review relating to our studies on social networks for social curation. We first provide a summary of research on the role of social networks on content-driven websites (our RQ2). Then, we survey papers related to social bootstrapping for our RQ3.

2.4.1 Studies on the role of social networks

Next, for our RQ2, there is a series of studies that investigate the motivation of users in creating social network links in general. One study identified that social links on Facebook have high predictive power in determining which newcomers will continue to engage with the platform in the future Burke et al. [2009]. Other studies have demonstrated that properties such as reciprocity and clustering promote interaction in natively created social graphs [Macskassy, 2012; Teng and Adamic, 2010].

Social capital is an influential concept in studying social network benefits for several fields such as economics [Knack and Keefer, 1997], development and policy [Woolcock and Narayan, 2000], organization theory [Nahapiet and Ghoshal, 1998; Tsai and Ghoshal, 1998], and, most relevant to us, modern sociology [Portes, 1998]. A number of social science researchers have studied social capital in the context of online communications and social networks, but most such studies have focused on Facebook (e.g. [Ellison et al., 2006, 2007; Steinfield et al., 2008; Wellman et al., 2001; Valenzuela et al., 2009; Cheung et al., 2011; Burke et al., 2011, 2010]). In particular, two studies by Burke et al. [2011, 2010] link specific user activities on Facebook with the feelings of social capital by combining data from self-reports and logs of Facebook. They find that directed communication, including wall posts, comments and likes, is associated with greater *bonding* social capital, while receiving messages from friends is associated with greater *bridging* social capital.

However, some recent studies have shown that a majority of users do not participate in social aspects of content. For example, [Gelley and John, 2015] show that most of Pinterest users are not very actively interact with their friends. In fact, many users with explicit friendship or follow links may not in fact have any content that they like in common on content-driven websites [Gelley and John, 2015; Musial and Sastry, 2012]. Swamynathan et al. [2008] find that most of users in a online retailer, Overstock[20], have not engaged into the social network provided by the platform.

In §5, we will attempt to understand better the limitations and benefit of social network for content curation by examining its effects at both information seeking and user retention.

2.4.2 Studies related to social bootstrapping

Many websites try to incorporate a social networking aspect to enhance user engagement and community interaction. Social networks are known to facilitate the formation of learning communities, foster student engagement and reflection, and enhance overall user experience for students in synchronous and asynchronous learning environments [Baird and Fisher, 2005]. Social networks are also the core of the design of new user-driven communities around health issues [Eysenbach, 2008] and have been utilised to facilitate community formation in environments ranging from professional settings [Lee et al., 2013] to online games [Choi and Kim, 2004; Ducheneaut and Moore, 2004]. Including the above studies, most existing research on the formation and evolution of online social network communities has focused on single networks. In contrast, our studies on social bootstrapping (i.e., our RQ3) in §5 evaluated interactions between two *different* networks: a generic social network and a target network on the content-driven website.

Multilayer networks (or also called multiplex, heterogeneous, interdependent, or multi-relational networks in literature) describe the fact that users may belong to different social networks (or layers) at the same time in real world. Each network layer could have particular features different from the others. A number of studies have looked into multilayer networks, including modelling of the formation and evolution of multilayer networks based on preferential attachment models [Magnani and

[20] https://www.overstock.com

Rossi, 2013; Nicosia et al., 2013; Podobnik et al., 2012]. In particular, resilience of cooperative behaviours is known to enhanced by a multilayer structure [Gómez-Gardeñes et al., 2012] and in some cases, cascading failures may occur in interacting networks [Buldyrev et al., 2010]. Multilayer structures are also known to speed up diffusion in networks [Gómez et al., 2013].

In §5, we will present a large-scale empirical study across two different network (Pinterest and Facebook) based on our **DATA-SOC** dataset. Empirical analysis of multiple networks is relatively uncommon. Szell et al. [2010] collected data from an online game and extracted networks of six different types of one-to-one interactions between the players. Then, both reciprocity and clustering were studied for each layer of the network. In contrast, our dataset shows the process of copying links between two independent websites, the source and target, where the original purpose of the link in the source network may be quite different from the intended purpose for the copied link in the target network.

Finally, this study is also related to the series of studies that investigate the motivation of users in creating social network links. One study found that professionals use internal social networking to build stronger bonds with their weak ties and to reach out to employees they do not know [DiMicco et al., 2008]. Another study identified that social links have high predictive power in determining which newcomers will continue to engage with the service in the future [Burke et al., 2009]. Other studies have demonstrated that properties such as reciprocity and clustering promote interaction in natively created social graphs [Macskassy, 2012; Teng and Adamic, 2010]. We will explore if such positive effects of social links also apply to links copied from unrelated *external* social networks.

2.5 Studies on Predicting Social Curation

Concurrent with the rapid rise of Pinterest, there have been several studies of Pinterest as a social curation platform [Hall and Zarro, 2012; Zarro et al., 2013; Ottoni et al., 2013; Gilbert et al., 2013; Ottoni et al., 2014; Chang et al., 2014]. Using a variety of approaches ranging from qualitative user studies [Zarro et al., 2013] and textual or content analyses [Gilbert et al., 2013; Ottoni et al., 2014], to large-scale exploratory data analysis [Han et al., 2014; Ottoni et al., 2013; Chang et al., 2014] and complex or multi-layer network studies, these works have shed light on a number of important aspects, such as the relationship between Pinterest and other social networks such as Twitter [Gilbert et al., 2013; Ottoni et al., 2014], the role of homophily and user specialisation [Chang et al., 2014], various differences in activities when conditioning on the gender of the user [Ottoni et al., 2013; Gilbert et al., 2013], the motivations of the content curators [Zarro et al., 2013; Han et al., 2014] and how to model user interests [Yang et al., 2015].

However, to explore answers for our RQ4, we will employ a different method in §6. More specifically, we will analyse the *end-to-end process of social curation* on Pinterest and extensively use a *machine learning approach* to automate social curation. Han et al. [2014, §7] also explore pin prediction, the scope is a more limited problem of predicting 20 pins for each of 4667 users (in comparison, our **DATA-IMG** dataset has 214K pins with 1.2M repins by 230K users) and checking whether these are in the user's pinboards after 125 days, without the multi-class classification into specific pinboards. Also, the best Han *et al.* models obtain an average precision of 0.015; we are able to obtain much higher values (c.f. Table 6.6). Also using a supervised machine learning approach is the preliminary work of Kamath et al. [2013], who propose a model for recommending boards to Pinterest users.

Chapter 3

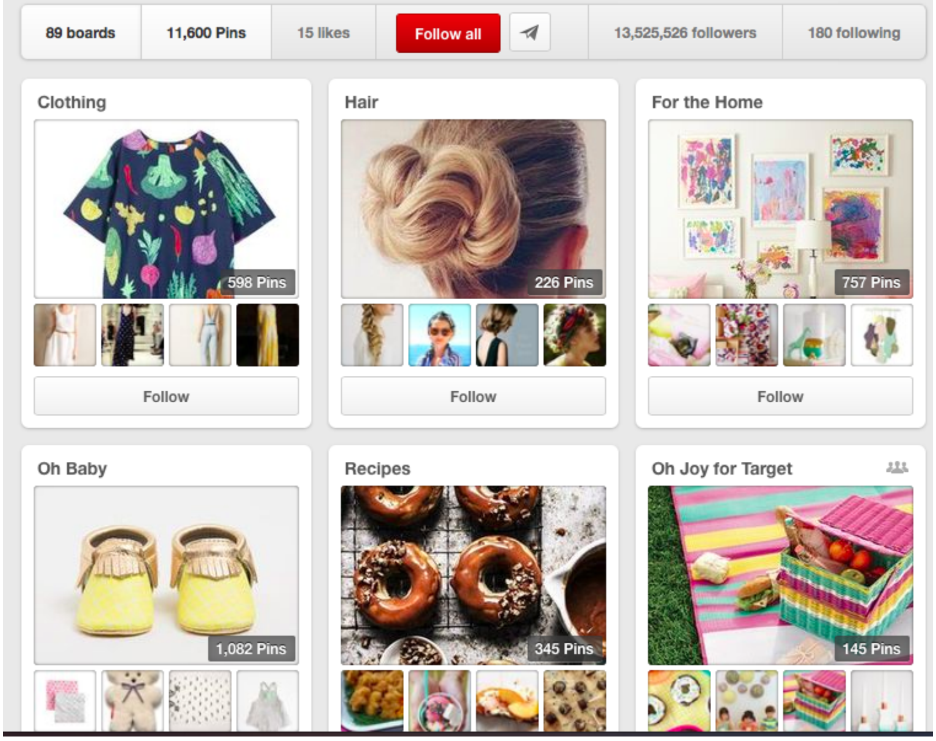
Preliminaries: Data and Terminology

Chapter layout. In this chapter, we introduce the data that we collected for this study and some related terminologies. In §3.1, we introduce Pinterest which is the source of our data. Then, in §3.2, we describe the collection process and some basic statistics of our social curation dataset. We have divided the dataset into four parts for our four research questions: **DATA-ACT** for content related empirical study, **DATA-SOC** and **DATA-USER** for social network related measurement research and **DATA-IMG** for our machine learning based automation of social curation.

3.1 Pinterest

Pinterest is a photo sharing website that allows users to save images and categorise 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*, uploading or importing from a URL external to Pinterest, or *repinning* from a existing pin on Pinterest. Users organise their pins into collections called *pinboards* or *boards*. A pinboard needs to be specified at the time of pinning; pins may be moved to a different pinboard later on. A repin creates a new pin on the repinning user's pin-

Figure 3.1
An Example of Pinterest User Profile Webpage



board. Each pinboard can belong to one of 32 globally specified *categories* on Pinterest, such as “Design”, “Products”, “Home Decor”, “Animals and Pets”, etc. Each category has a page on Pinterest, 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.

Pinterest incorporates social networking features to allow users to connect with other users with similar interests. Users can also actively *follow* other users or pinboards they find interesting, effectively creating a directed social graph. Users can create connections to other users on Pinterest in two ways. The first is to explore the website and follow users they find interesting. We call social links created in this way *native links*, as they are created natively on the platform. The second way is using the “Find Friends” function (i.e., via social bootstrapping, refer §5.1). Users can connect their Facebook and Twitter accounts with their Pinterest accounts and the Friend Finder function will provide a list of Facebook and Twitter friends who are also registered on Pinterest. Users can select some of them to follow on Pinterest, and create *copied links*.[1]

The social network thus created has been one of the fastest growing social networks of any kind in recent years [Dube, 2015; Chaffey, 2016]. To understand the effects of the social network on pinning, we distinguish and quantify the effects of pins involving the social network of the pinner, which we call *social repins*, from *non-social repins*, which involve users to whom the pinner is not connected socially (i.e., repins of *non-friends* or *strangers*). We further distinguish between repins made from users who are friends with the original pinner as native and copied, based on the type of friendship link. Note that the nomenclature is always relative to the user who is (re)pinning an image onto their own pinboard.

3.2 Pinterest Social Curation Dataset

To explore our research questions on online social curation, we collected a unique dataset from Pinterest, which includes four parts: **DATA-ACT** with nearly all Pinterest activities within a few of weeks, **DATA-SOC** with a snapshot of the Pinterest social network, **DATA-USER** contains the entire history of randomly selected 50K users since they joined Pinterest and **DATA-IMG** with thousands of visual features of more than 200K images curated to Pinterest.

3.2.1 Activity dataset (DATA-ACT)

To implement an empirical study for the content involved in social curation as for our RQ1, 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[2]; 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. In total, 8.5 million users (termed as

[1] We will focus on the connection between Pinterest and Facebook in this dissertation, as more than 60% of Pinterest users have connected with Facebook accounts, while only about 10% of users connected to Twitter accounts according to our **DATA-SOC** dataset.

[2] This setting has been changed in April 2013

Table 3.1
The DATA-ACT Dataset Details

	Pinterest
Timespan	03–21 Jan 2013
Users	8,452,977
Likes	19,907,874
Repins	38,041,368

active users), 38.0 million repins and 19.9 million likes were included (see Table 3.1). We also obtained the basic statistics of active users, such as the number of pins, likes, followers and followees.

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 this way, we have collected about 50M repin and like actions. This allows us to compare the usage pattern of those two types of actions. At the same time, we have collected the basic information of images (such as the domain they originally imported from), which is useful for us to understand the motivation of curators. Table 3.1 provides a summary of the aggregate volume of data collected. We will use the dataset to answer our RQ1 in §4.

3.2.2 Social network dataset (DATA-SOC)

Next, we explain how we construct a social network dataset for our study on social bootstrapping (i.e., our RQ3).

As we discussed, the social graph of Pinterest is created through users *following* other users or pinboards they find interesting. We call social links created in this way *native links*. In addition to this method, users are able to connect with their Facebook and Twitter accounts and import their social links into Pinterest via social bootstrapping. The *Find Friends* func-

Table 3.2
The DATA-SOC Dataset Details

(a) Target social graph

	Nodes	Links
Pnt network	68,665,590	3,871,570,784
Fb-copied	40,472,339	983,520,986
Pnt-native	28,193,251	2,888,049,798

(b) Facebook network

Nodes	links
2,322,473	444,216,279

tion provides a list of Facebook and Twitter friends who are also registered on Pinterest. Users can select some of them to follow on the Pinterest website, which we call *copied links*.

Table 3.2 summarises our **DATA-SOC** dataset, consisting of the social graph on Pinterest and the corresponding nodes and edges on Facebook. To obtain the Pinterest social graph, we used a snowball sampling technique, starting to crawl from a seed set of users which we collected in **DATA-ACT**. In total 68.7 million Pinterest users and 3.8 billion directed edges between them were obtained. For each user, we checked whether there was a connected Facebook account, and gathered basic profile information such as gender and profile, as well as basic statistics such as the number of pins, likes, followers, and followees. Of the 68.7 million, 40.4 million were Facebook-connected users, who have 2.4 billion links between them on Pinterest.

We next separate the 2.4 billion edges into those which are present on Facebook (i.e., are *Fb-copied*), and those which are native to Pinterest (*Pnt-native*). To identify the *Fb-copied* portion of the network, we used the Facebook API to individually check whether a Pinterest link between two connected users was also present between the corresponding Facebook accounts[3]. We find that 0.98 billion links between connected users are

[3] Note that checking whether a pair of users are friends is affected by users' privacy setting. That is, it is unknown for us whether two users are friends or not if both of them had set their friend lists as private. Also, we assume that a link which exists both on Facebook and the target networks is a copied link, first made on Facebook and then copied to the target network. Although we expect this to be the case normally, it is possible for user pairs to link to each other separately on Facebook and Pinterest, or link first on Pinterest, and subsequently on Facebook. We are unable to distinguish these

also on Facebook. These form our *Fb-copied* network. *Pnt-native* links were identified by excluding the *Fb-copied* network from our Pnt network.

Then we check the 8.5 million active users from the **DATA-ACT** dataset, there are 5.2 million connected to Facebook. We crawled the Facebook pages of these 5.2 million connected active users, and attempted to obtain their Facebook friend lists. Due to privacy settings, only 2.3 million users' social links could be obtained. Together, this collection of Facebook edges constitutes a subgraph of 444.2 million edges (Table 3.2b). Of these, 141.9 million are *copiable* links, i.e., edges between connected users who are on both Facebook and Pinterest.

To this end, we constructed a social bootstrapping dataset, which allows us to explore the social structure of copied and native networks. At the same time, together with **DATA-ACT** dataset, we could also examine users' interactions over social links and thus check the benefit and limitation of social bootstrapping. We will implement those studies with this dataset in §5.

3.2.3 User dataset (DATA-USER)

To analyse users' activities and understand how they use social network for RQ2, we created a user dataset which has two parts. First, in **DATA-USER**-1, we crawled the entire pinning activity history of 50,000 users from the time they joined Pinterest until an arbitrarily chosen end date of April 1, 2014. To ensure that the sample was as unbiased as possible, the user IDs were randomly sampled from a near complete snapshot of the Pinterest social network **DATA-SOC** collected in Jan 2013. Thus, all users in **DATA-USER**-1 have at least 15 months of activity on Pinterest. Some of these user accounts have been suspended or deleted, and some have no pins. This left us with 48,185 users, who collectively have 3.9 million social links and 10.3 million pins.

To analyse the various effects of the social network, we looked for users in **DATA-USER**-1 who had activities (repins) involving links that were created natively on Pinterest, as well as links copied from their Facebook friendships. This resulted in 10,312 users, who were linked to (followed by or were following) 573,015 other users. We further obtained all 174,170,718 pins of these friends made between Jan 1, 2014 and Apr 1, 2014, resulting in a second, larger, dataset **DATA-USER**-2.

cases from links copied using friend finder tools.

Apart from the pins, for each user in **DATA-USER**-1 and **DATA-USER**-2, we also obtained user metadata such as number of pins, likes, follower, following, and boards. For each pin in **DATA-USER**-1 and **DATA-USER**-2, we obtain pin metadata such as a “created at” timestamp, the board it was pinned in, the category of the pinboard, and numbers of repins, likes and comments.

In this way, we use **DATA-USER**-1 to explore the usage of social networks in users’ activity history. **DATA-USER**-2 allows us to compare the activity of friends over time. With this data, in §5, we examine the role of social networks on information seeking and user retention.

3.2.4 Image dataset (DATA-IMG)

Finally, for our RQ4, we will implement a machine learning based study for social curation, hoping to obtain a mechanistic understanding of the end-to-end process of curation on Pinterest. For this, we collected the **DATA-IMG** dataset. We wish to understand how the content features of the image and the pinner affect the activity of curation. Therefore, we focus on users with more than 10 pins in our **DATA-ACT** dataset, and on pins which have been repinned at least 5 times, ending up with a set of 214,137 pins, 237,435 users and 1,271,971 repins for analysis. Then, we downloaded all of those 214K images and extract their visual features (refer §6.2.2) as the **DATA-IMG** dataset.

Chapter 4

An Empirical Look at Social Curation

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, Delicious allowed users to categorise interesting URLs by tagging them, and sharing them with followers. Digg[1] and Reddit[2] have allowed sharing of news articles, and so on. In this chapter, we take a broad view of online content curation and seek to understand the basic process by examining our **DATA-ACT** dataset. We exploit this, seeking to understand our RQ1, i.e., why people curate, how they curate, and what others, namely followers of the content curators, find useful.

Summary of findings. 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 (re)pinned or most liked on Pinterest are likely from websites with a low PageRank value or Alexa Global Traffic Ranking. We conjecture that curation might provide a personal value to the curators by collecting together items which may be dif-

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

[2] http://www.reddit.com

ficult to find by other means. This is supported by Shirky [2010] who said "curation comes up when search stops working". 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% of items get over 70% of the curation actions on Pinterest.

Next, we examine the different curation actions to better understand the process. Based on the similarity between the curation actions on Pinterest and other platforms, we propose a distinction between two kinds of content curation actions: *structured curation*, categorising content along with other "similar" items from some perspective (e.g., "repin" and tag), and *unstructured curation*, which involves highlighting or collecting interesting content without categorising them (e.g., "like" and "love"). We find that different users prefer different actions, with some preferring unstructured, and others structured curation. 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. 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 categories on Pinterest are similarly successful in attracting followers.

Chapter layout. In §4.1, we check the characteristics of the content people curated, and exploit the implicit reasons for why people curate. §4.2 shows the differences of two types of curation actions, and we explore how people use them. Then, we examine content curation from the aspect of social values, and ask how curators can attract more followers in §4.3 . Finally, §4.4 summarises the chapter.

4.1 Why Do 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

Table 4.1
Curation Highlights Content Not Popular in Other Rankings

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

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 nontraditional sites.

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.

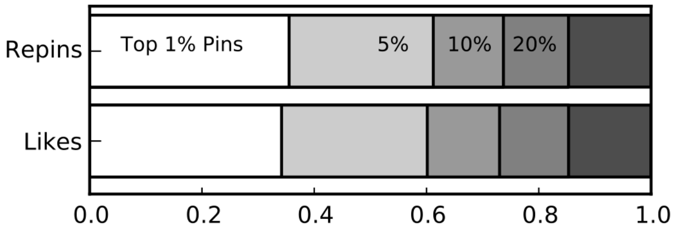
4.1.1 Curation highlights new kinds of content

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 Pinterest, since 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[3]).

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 4.1 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

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

Figure 4.1
Distribution of Curation Activity



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.

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].

We also validated the experiment in Last.fm[4], a popular social music recommendation service that has provided curation relevant supports for over nine years, and published the result in a paper [Zhong et al., 2013, Table 3]. In the experiment, we first ranked tracks with a curation-based method, i.e., by the number of tags (structured curation) and likes (unstructured curation) made in UK. Then, we compared it against weekly sales and radio airplay charts published by Music Week[5], a trade paper for the UK record industry and an established music data provider. A similar lack of correlation between highly ranked tracks through curation and the traditional Music Week rankings was also observed in the experiment.

4.1.2 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" [Shirky, 2010]. We find evidence for this by examining the distribution of curation actions in our corpus. Figure 4.1 shows a highly

[4] The validation experiments on Last.fm dataset were mainly undertaken by Sunil Shah and were published in a paper [Zhong et al., 2013] with me as first author. We omit them in this thesis, as we focus on Pinterest data.

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

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.

This is a very difficult question of whether curation creates value for users by synchronising a community, and just a data analysis would provide an incomplete picture. Therefore we augment the data analysis (our main contribution) with a user study[6] to examine 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.

According to the survey, some 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. According to the user study, 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 the survey allowed multiple answers to be selected for this question). 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.

This is consistent with another user study conducted by Linder et al. [2014], which suggests that users feel that they have a "separate space", that "they are not pressured by extrinsic judgments on the quality of their Pins and repins".

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

[6] The user study was mainly undertaken by Karthik Sundaravadivelan, with myself and other authors participating in the survey design. The paper has been published in a paper [Zhong et al., 2013] with me as first author.

trying to curate for social value and do not try to integrate within their respective communities.

4.2 How Do People Curate: Understanding Curation Actions

As discussed before in §1.2, multiple curation actions are available to a user, and we could distinguish them based on whether they simply highlight an item ("like", "comment", "save"), 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 post). We term the former as *unstructured curation* and the latter *structured curation* because of the organisational structure induced by pinning or tagging. In this section, we will 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:

$$R = \frac{Unstructured}{Unstructured + Structured} \tag{4.1}$$

First we explore how users curate content, and whether they prefer structured or unstructured curation. We calculate the unstructured curation ratio R in our **DATA-ACT** dataset and consider the top 1%, the top 10% and all users for each activity in Figure 4.2.[7]

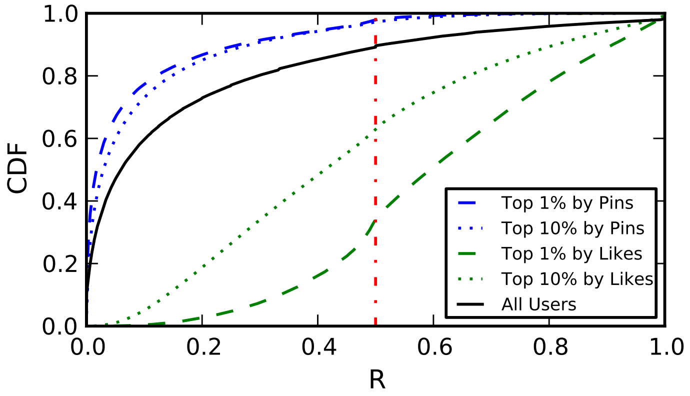
We define users who prefer structured curation over unstructured curation (i.e., have $R < 0.5$) as structured curators. Conversely, users who prefer unstructured over structured curation ($R > 0.5$) are termed unstructured curators.

4.2.1 Some users prefer structured, others unstructured

In Figure 4.2 we first draw attention to the difference between the proportion of structured and unstructured curators on each network. Figure 4.2 shows that on Pinterest, more than 80% of all users are structured curators. Comparatively, the same experiment on Last.fm [Zhong et al., 2013,

[7] Since users can use "pin" and "repin" to categorise images in Pinterest, both are included to represent the structured curation action.

Figure 4.2
Users' Unstructured Curation Ratio



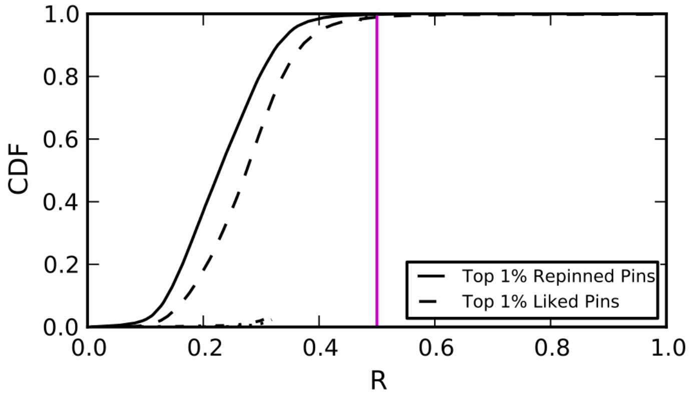
CDF of users' unstructured curation ratio R in Pinterest,there are a mixture of structured ($R < 0.5$) *and unstructured* ($R > 0.5$) *curators. Generally Pinterest users participate in structured curation activities.*

Figure 3b] shows that less than 40% of all Last.fm users are structured curators. This suggests that Pinterest generally users prefer structured curation activities whilst most Last.fm users prefer in unstructured curation activities. This corresponds with the findings of the user study in [Zhong et al., 2013]: the majority of Pinterest users surveyed would rather repin a post than like it if it matched their interests; while a majority of Last.fm users would rather love than tag a music track.

The larger proportion of the top users by loves who are unstructured curators on Last.fm can be explained by 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 the user study: 65% of surveyed Last.fm users have never tagged a track. Conversely, only 11% have never loved a 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).

Figure 4.3
Unstructured Curation Ratio R of Top Items



The CDF for unstructured curation ratio R of top items on Pinterest. Magenta line indicates $R = 0.5$. In Pinterest 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 items) have $R < 0.5$.

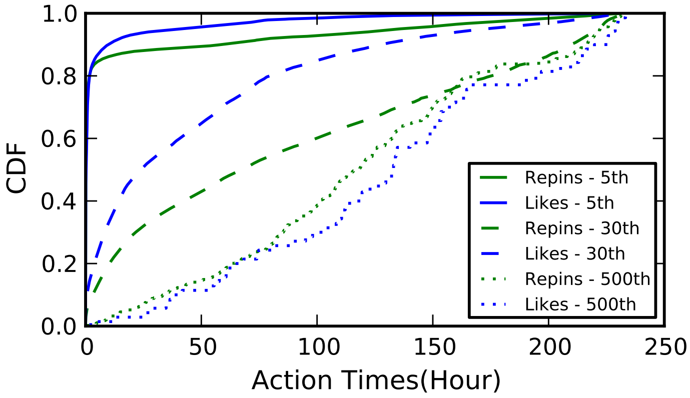
4.2.2 Structured curation is preferred for popular items

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 and consider the top content items by curation activity in Figure 4.3. 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 [Zhong et al., 2013, Figure 4]). This is further supported through the Pinterest user study where average R for popular content was 0.33 and for unpopular content was 0.5.

4.2.3 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

Figure 4.4
Pinterest Repin and Like Action Times



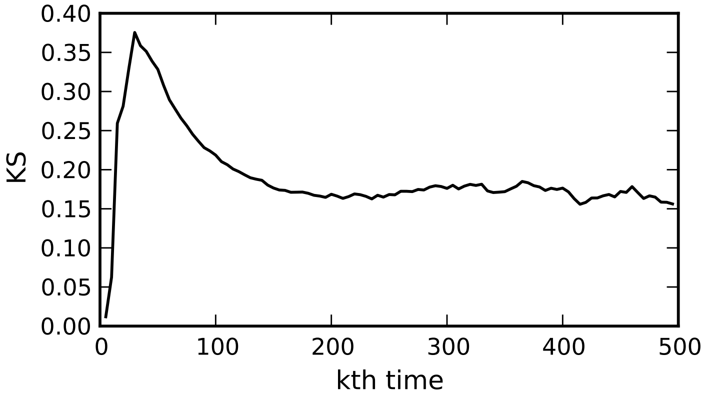
CDF of Pinterest repin and like action times The kth repin (like) time for a pin is the time between creation of a pin and the kth repinning (liking) in Pinterest. Likes accumulate quicker at first and there is a considerable difference between in the time it takes to get 30 repins and 30 likes. The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

posted. We consider this time for both structured (pin) and unstructured (like) curation activities.

Figure 4.4 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 Figure 4.5, we summarise the difference between the distribution of $T_s(k)$, the k-th action time for structured curation, and the distribution of $T_u(k)$, the k-th action time for unstructured curation. This difference can be measured using the Kolmogorov-Smirnov statistic given by $D = max(T_s(k) - T_u(k))$. 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.

Figure 4.5
Kolmogorov-Smirnov Statistic for Action Times



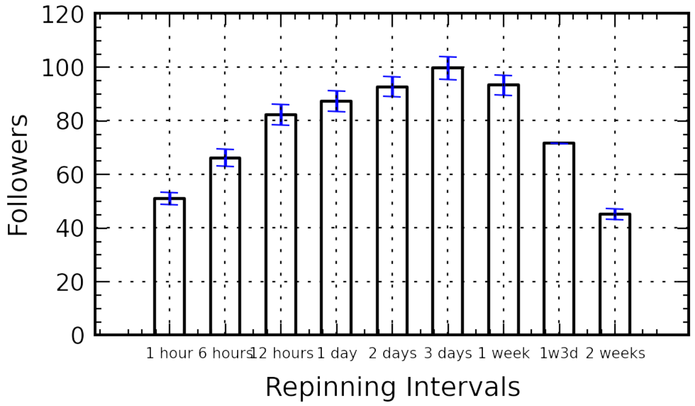
Kolmogorov-Smirnov statistic for action times Extending Figure 4.4. There is a difference between structured and unstructured curation over successive actions. We could observe a noticeable peak for Pinterest, after which the difference between the two actions is minimised.

4.3 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 random sample 1000 users and compute the mean of the number of followers accumulated by these users as a measure of how useful the bin's value of the attribute is, to other users. We repeat the experiment for 1000 times, and report the average result of all of 1000 experiments.

In summary, in 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. We also find that users who prefer structured curation (i.e., those who prefer 'pinning' to 'liking') accumulate more followers.

Figure 4.6
Consistent Structured Curation Attracts Followers



Structured curation attracts followers when it is consistent and regular. Users with a short interval between successive repins attract a large number of followers on Pinterest.

4.3.1 Consistent and regular updates

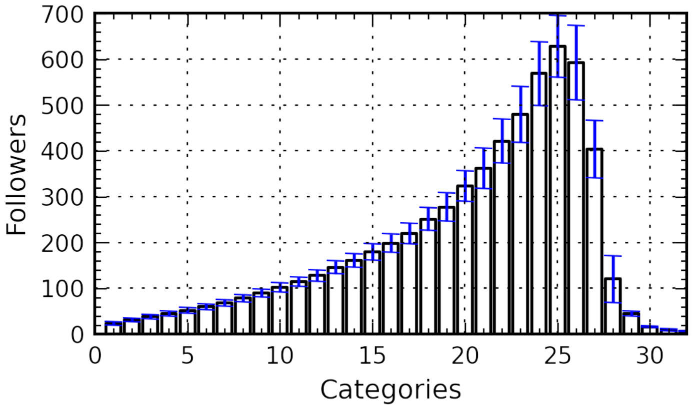
Bhargava [2009a] has suggested that the most important part of a content curator's job is to continually identify new content for their audience. Figure 4.6 examines the role of regularity, by plotting the average of the intervals between consecutive structured curation actions[8] for each user vs. the average of the followers accumulated, and finds support for this theory. Note that for Pinterest, too short an interval between repins could detract followers, while the same experiment on Last.fm does not exhibit this phenomenon [Zhong et al., 2013, Figure 8b]. We conjecture that given the order of magnitude higher volume of curation actions on Pinterest, 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.

4.3.2 Diversity of interests

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. Figure 4.7 shows

[8] Because a user typically has many intervals between repins, we additionally use a average 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 average value of I.

Figure 4.7
Diversity of Interest Attracts Followers



Pinterest users interested in nearly all the categories attract more followers.

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. For Last.fm, we captured diversity of interest by counting the number of genre-specifying tags which have been used for tagging by the user and observed similar trend [Zhong et al., 2013, Figure 9b].

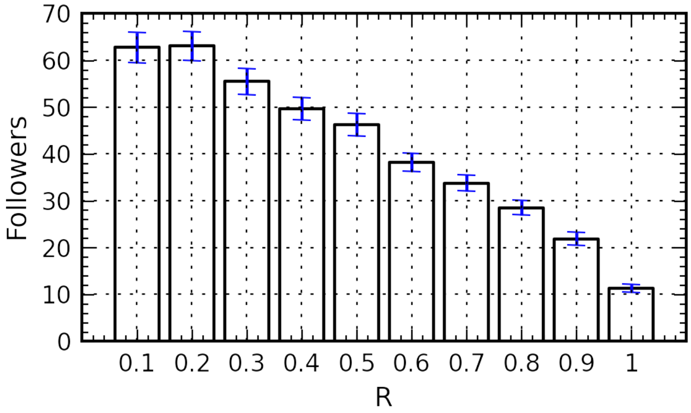
Note that there might be potential confounding factors: For example, being active in a number of categories might simply be a consequence of being more active on the site, and more active users might attract more followers, as shown above. To confirm that our finding about the importance of diversity of interests is not simply an artifact of diversity in usage, we verified that the result of Figure 4.7 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).

4.3.3 Structured vs. Unstructured curation

In §4.2, we discussed structured and unstructured curation, and demonstrated that on Pinterest, most users would prefer to use structured curation.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 4.8, 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.

Figure 4.8
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.

However, we did *not* observe a similar trend in Last.fm [Zhong et al., 2013, Figure 10b]. 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.

4.4 Summary and Discussion

This chapter used a quantitative analysis of several weeks of curation actions on a social curation websites, Pinterest.com, to characterise the usage pattern of social 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 the 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, some user studies [Zhong et al., 2013; Linder et al., 2014] reveal that the majority of users view curation as a personal activity, rather than a social one. Actually, this is a limitation of our study that it is based solely on data analysis. It would be interesting to validate whether the conclusions we draw correspond to user motivations. Our choice of relying solely on data analysis was driven by the observation that even if users believed they were not trying to follow the pattern, the data might indicate an unconscious bias in user choices towards fitting in with a (potentially subliminally) perceived pattern. Thus, we are of the opinion that the data analysis in and of itself can be a valuable first step towards understanding social curation actions. Disambiguating between users' conscious ideas about fitting in and actions observed through a data-based approach would require careful research design and can be the subject of a follow-on work.

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.

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). This study throws a light on the social network of social curation services by considering the number of followers as the metric to the user influence. But it is still unclear how curators interact with their friends, and what are the benefits and limitations of social networks. Threrefore, in next chapter, we will dig into the social aspect of social curation services and seek to find answers to these questions.

Chapter 5

The Role of Social Networks

It has become *de rigueur* to create social networks amongst users on all kinds of Web 2.0 sites, especially those involving content sharing. Several prominent social curation or content-driven sites, ranging from Pinterest (image-based sharing) and Vimeo (video sharing) to last.fm (music sharing) and Etsy (social shopping) incorporate social features such as the ability to follow other users' activities, and to like or repost (share) content or products that they like. A key feature of social networks on these sites is that links are intended to be made based on shared interests around an item or a category of items [Hendricks, 2014; Jamison, 2012].

However, recent research has shown that a majority of users do not participate in social aspects of content [Gelley and John, 2015] or product [Swamynathan et al., 2008] sharing websites. Further, many users with explicit friendship or follow links may not in fact have any content that they like in common [Gelley and John, 2015; Musial and Sastry, 2012]. Our study in previous chapter also show that most people curate for personal reasons.

These findings stand in direct contrast to numerous studies where social networks have been shown to help in community formation, in a diverse range of interest- and goal-oriented environments and applications such as learning [Baird and Fisher, 2005; Conole and Culver, 2010; Heiberger and Harper, 2008], working [DiMicco et al., 2008; Lee et al., 2013], medicine [Eysenbach, 2008] and online games [Choi and Kim, 2004; Ducheneaut and Moore, 2004].

In light of these conflicting results, it is natural to ask:

RQ 2 (Curators vs. Friends) *What is the value of social networks on social curation services?*

In this chapter, using two parts of our social curation dataset, **DATA-SOC** and **DATA-USER**, we explore the benefits and limitations of social networks for online social curation services.

How to design online communities is another problem of the designers of social curation services. In creating a social experience on a website, designers face an important choice: Should they create a social network at all? If creating a social network, should they create a brand new social network that is customised and optimised for the site, or instead borrow links from other social networks such as Facebook or Twitter, and connect user who are friends on the other sites. As discussed in §1.3, the latter option (which we term *social bootstrapping*) has recently become a possibility, with both Facebook and Twitter opening up their social graphs to third-party web- sites, who can write friend-finder tools that help users select and import friendship links from these established networks into their own service. Social bootstrapping has direct implications on how a new online social network community can grow quickly. However, this problem is complex to examine with real data because it involves user interaction across multiple heterogeneous networks. To this end, we gather massive amounts of data from Facebook and Pinterest involving tens of millions of nodes and billions of links (i.e. our **DATA-SOC**) and explore following question:

RQ 3 (Social Bootstrapping) *What are the benefits and limitations of social bootstrapping?*

Summary of Findings. First, we examine the benefit of social networks on social curation websites. We find that the Pinterest social network serves an important purpose for bonding users: users who do engage with the social network, and in particular, users who have relatively close friends whom they know from offline contexts and from another social network (Facebook), are the most likely users to return to the platform.

These users are also the most active subset of users and contribute the vast majority of pins.

Surprisingly for a goal-oriented and interest-based social network in Pinterest, we find a non-trivial proportion of information seeking happens through non-social means which indicates a decreased importance for social networks in content discovery in Pinterest. We explain this by considering the interest-oriented nature of information seeking: User interests evolve over time, and appear to be satisfied better by recent pins which are featured on the Pinterest homepage rather than their friends. Indeed, we find strong evidence that when one user repins another user, the similarity between them peaks, and as users' interests evolve, it may lead them to repin strangers who are more similar to their current interests than their own friends.

Then, we ask which part of the social network are more important for user engagement, the subgraph created natively in Pinterest or the one copied from Facebook via social bootstrapping. Using our **DATA-SOC** dataset, we study the structural properties of the copied subgraph, comparing it to native subgraph. We examine some structure properties that are thought to be related to user interactions, such as reciprocity and clustering coefficient. Our results show that copying enriches reciprocity and clustering of the local structure, indicating that social bootstrapping successfully promotes user engagement.

Copying links yields diminishing returns. As users become more active and influential on the new website, they create proportionally more native links than copied ones. Native links offer a benefit over copied links: users connecting natively on Pinterest tend to be more similar to each other in their tastes than with the ones copied from Facebook. This is an important observation for long-term user engagement, as prolific users tend to engage more with native links and fine-tune the local relationships to meet their interests. As a result, we conclude that while "copying" links is essential to bootstrap one's network, the opposite "weaning" process is equally important for long lasting user engagement.

Chapter layout. The rest of this chapter is organised as follows. First, in §5.1, we introduce our terminology on social bootstrapping and measures for our analysis. In §5.2, we examine the benefits and limitations of social networks. To understand the role of social networks, we also explain non-social interactions with users' interest alignmentin §5.3. Then, in §5.4, we further ask which subgraphs of Pinterest social network are more important for the user engagement, the native network or the copied one. Finally, in §5.5, we take a long-term view and examine the effects of social bootstrapping as users become more active and influential.

5.1 Terminology and Methods

Before exploring our research questions, we first introduce some terminologies and methods that will be used in this chapter.

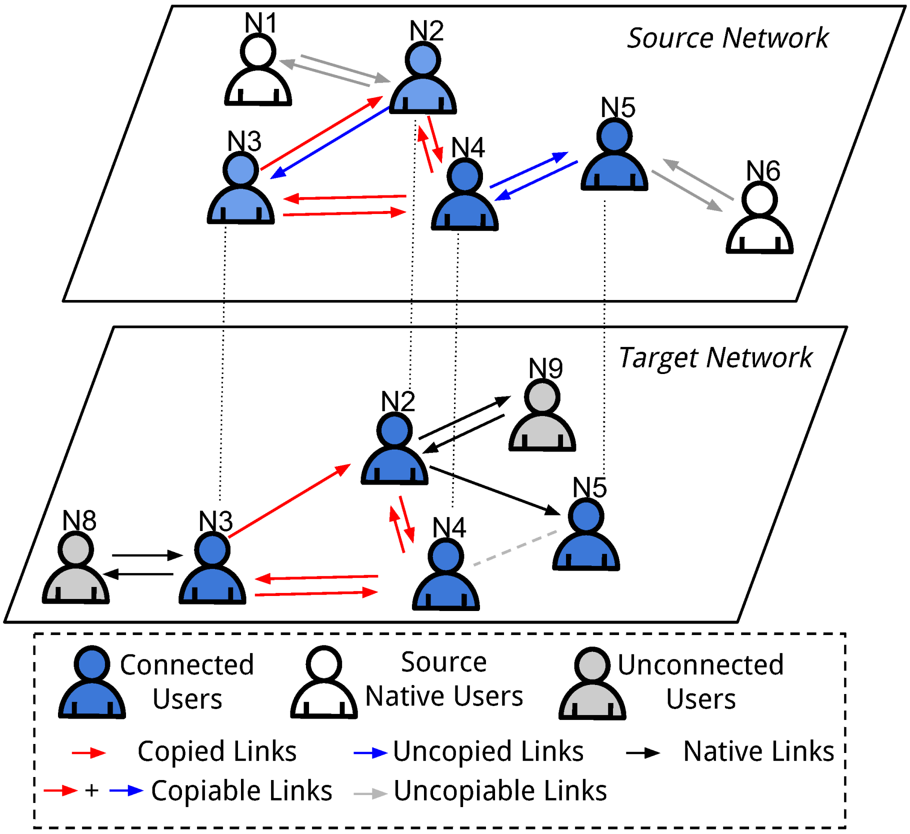
5.1.1 Social bootstrapping

As discussed in §1.3, We term the act of copying existing friends from an established social network onto a third-party website as *social bootstrapping*. Here, we define several sub-networks to describe this phenomenon for our later study (we reshow Figure 1.1 here for the convenience of readers):

Source network: The social graph of an established social network like Facebook (*Fb* for short), which contains a significant number of nodes and links (e.g., 1.65 billion monthly active users as of March 31, 2016 [Facebook.com, 2016a]). The source network is displayed as the upper layer in the toy example in Figure 5.1. Note that some users, such as $N1$ and $N6$ are *source native* and are present only in the source network.

Target network: The relatively new third-party network that allows users to copy links from established networks, displayed as the lower layer in Figure 5.1. *Connected nodes* are the subset of all nodes in target network that have used the "Friend Finder" tool to connect their accounts to the source network. In the toy example, blue nodes, i.e., $N2$, $N3$, $N4$ and $N5$ are the connected nodes. Grey nodes, i.e., $N8$, $N9$ are *unconnected nodes* who either exist only on the target network or have chosen not to connect their accounts on the source network to their identity on the target network. Within the target network, social links copied from the source network are called *copied links* and those created natively are called *native links*. Copied links in

Figure 5.1
The Structure of Social Bootstrapping (same as Figure 1.1)



the target network may be directed even if they are copied from the undirected source network. Copied links are a subset of *copiable* links, the set of all links between connected nodes in the source network. We take Pinterest (*Pnt* for short) as target network of interest.

Copied network: The social subgraph of the target network solely containing copied links and all *connected nodes*. In Figure 5.1, the copied network contains the red edges and all blue nodes. We call the network copied from Facebook as *Fb-copied*.

Native network: The subgraph of the target network that only contains native links and the corresponding nodes at either end of each native link. In the toy example, the native network is the subgraph made up by black edges and nodes linked by them, i.e., $N2$, $N3$, $N5$, $N8$ and $N9$. Nodes can be in copied and native networks at the same time, but links are either copied or native. We call the native networks for Pinterest *pnt-native*.

5.1.2 Measurement Methodology

It is thought that social networks will affect users from three aspects [Putnam, 1995; Ellison et al., 2007; Burke et al., 2009]: retention to the platform, information sources and interaction with friends. In this chapter, we will focus on these three factors: First, we will examine how social networks affect the *user retention* of platforms and the *information source* of users to explore the role of social networks (RQ2). Then, in our study of RQ3, we compare the effects of copied and native networks based on *social interaction*.

User retention. We follow Java et al. [2007], and mark users who have an activity in a given week as *active*. An active user is considered as *retained* if she also pins or repins in *each* of following X weeks. In our experiments, we measure the fraction of users retained among all active users in each week. In this chapter, we will only show the results of $X = 3$, although similar results can be obtain for $X = \{1, 2, 4\}$.

Information source. Creating new pins (or repins) is by far the most common activity on Pinterest, and presents a quintessential information seeking activity according to our **DATA-ACT** dataset. Users may find out about the new pin external to the website and upload the pin themselves. Or, they may repin an existing pin they find on Pinterest. In the latter case, they may repin a pin from a friend of theirs (i.e., a social repin), or they may repin an image pinned by someone with whom they are not connected socially (i.e., a non-social pin). We check what fraction of a user's pins come from each of the three sources – *uploaded pins, social repins and non-social repins*. In theory, each of these sources can be important to different extents in different kinds of information-seeking activities. For instance, at the time of the first pin of a user in a *new category*, the user may be less knowledgeable about that category. Similarly, the first pin in a *new board* may be seen as a new 'sub category' or thematic collection, and may have different information seeking patterns, in comparison with subsequent pins. We will examine these measures to evaluate how social networks change users' content consumption.

Social interaction. In our studies on the effects of the social network on interaction, we distinguish and quantify the effects of pins involving the social network of the pinner, which we call *social repins*, from *non-social repins*, which involve users to whom the pinner is not connected socially (i.e., repins of non-friends or strangers). We further distinguish between repins made from users who are friends with the original pinner as native and copied, based on the type of friendship link.

5.1.3 User groups

A basic strategy we employed in this chapter is to compare above measures for different user groups. The first method we employed to divide users is according to their social structure or interaction, which allow us to identify *non-social*, *social*, *natives*, *expats* and *bi-network* users. The other method is to divide users according to their *maturity*, which measures how long a user has been on Pinterest. This method can be used to show the long-term effects of social networks.

Social structures based groups. To compare users with different kinds of social network structures, we could divide them into four groups as follows:

- *Non-social*: users who do not have any social friends.
- *Facebook expats*: users whose social links are entirely copied from Facebook.
- *Pinterest natives*: users who do not copy any links.
- *Bi-networked*: users with a mixture of native and copied links.

This method requires only social relationships, thus is available for all of 68 million users in our **DATA-SOC** dataset.

Social interaction based groups. Next, rather than only considering social relationships, we take social interactions into the group division as follows:

- *Non-social*: users who do not have any social repins. All other users are social users.
- *Facebook expats*: users whose social repins are entirely made from copied friends.

- *Pinterest natives*: users who only have native repins and do not make any copied repins.
- *Bi-networked*: users with a mixture of native and copied repins.

Since this requires social interaction information, this method is only available for 50 thousand users in our **DATA-USER** dataset.

User maturity. To quantify the level of user maturity on Pinterest, we evaluate users' activity and influence. For the activity level, we employ the number of pins made (including repins of other users' pins) by users as the major measure, since it is the most popular activity on Pinterest (according to our **DATA-ACT** dataset, refer Table 3.1) and available in both **DATA-SOC** and **DATA-USER** datasets. We will also use the numbers of boards created and likes to others' pins as measures for validation purpose. The level of influence of a user is measured as the activity of *other* users directed towards that user, i.e., the number of repins and likes received by that user for her pins. Note that there is only three weeks influence information in our **DATA-SOC** for active users.

5.2 The Benefits and Limitations of Social Networks

Now, we explore the benefit of social networks. As introduced in previous section, we will compare the *user retention* and *information seeking* of users in different groups to examine the benefits of social networks for platforms and users.

5.2.1 Social users are more active

We start with a macro-scale analysis of users with different kinds of social links, and examine the number of pins they are responsible for in our **DATA-USER** dataset, to understand whether the social users are fringe or core users of the platform. Table 5.1 shows the results for user groups based on social interaction. *Social users* – those who have repinned at least one image from their friends – are the users who power Pinterest: there are only about 61% of users of this kind, but nearly all Pinterest (re-)pin activities (97%) are made by them. At the same time, we also notice that

Table 5.1
Social Users are Active

Conditions	# of users	# of pins
All users	48,185	10,394,396
Users with social repins	29,194 (61%)	10,080,257 (97%)
Connected users (i.e., users logged in with FB ID)	24,130 (50%)	8,288,315 (80%)
Connected users with **social** repins	17,164 (36%)	8,180,340 (79%)
Connected users with **native** repins (i.e., Pnt natives + Bi-networked)	15,488 (32%)	8,063,550 (78%)
Connected users with **copied** repins (i.e., FB expats + Bi-networked)	11,978 (25%)	7,453,134 (72%)
Connected users with **copied and native** repins (Bi-networked)	10,302 (21%)	7,336,344 (71%)

We filter users in our ***DATA-USER*** *dataset, and show a small group of users with social repins contributes the majority of pins. Note that the user groups are defined according to "interaction version" of social status.*

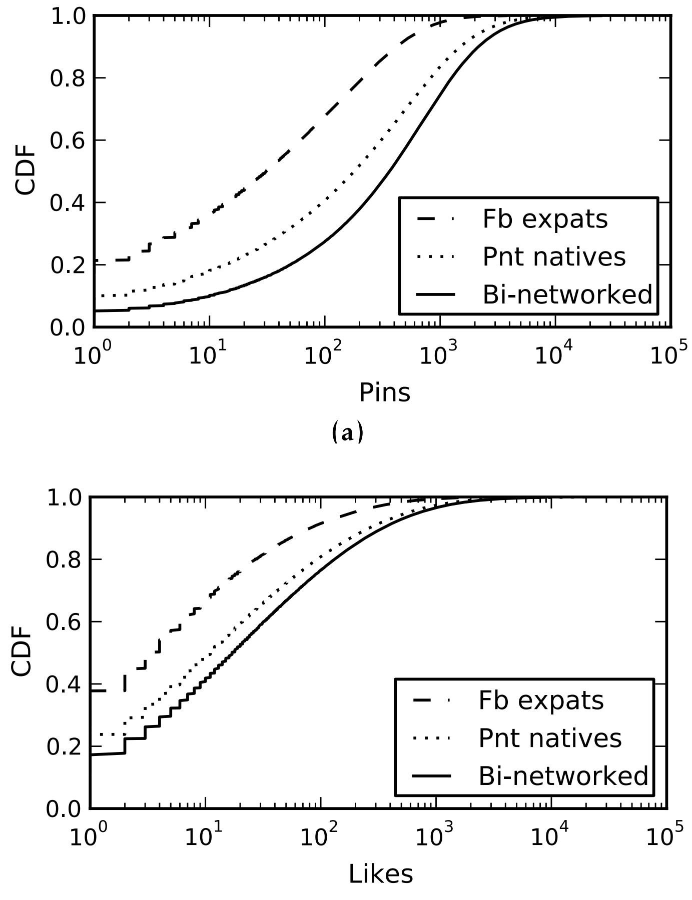
connected users – users who login to Pinterest via their Facebook account by social bootstrapping – are active and contribute 80% of all activities.

Then, we ask which type of connected users are more active. In Table 5.1, we can see that, among connected users, those who have made social repins contributed most (79% out of 80% activity). In other words, users with some form of actual social interaction are more active than users who have simply formed links.

So we further divide users into three types according to their social repins: *Facebook expats*, who have interaction with copied friends, *Pinterest natives*, who interact with native friends and *bi-networked* with a mixture of native and copied social repins. In Table 5.1, it is clear that bi-networked users with both native and copied repins (21% of all users) contribute most (71%) of Pinterest activities.

This result is based on our **DATA-USER** dataset, which includes the entire activity history of 50 thousands random sampled users. But is this complete? Actually, we also validate the result with 68 millions users in

Figure 5.2
Bi-networked Users Exhibit More Activity



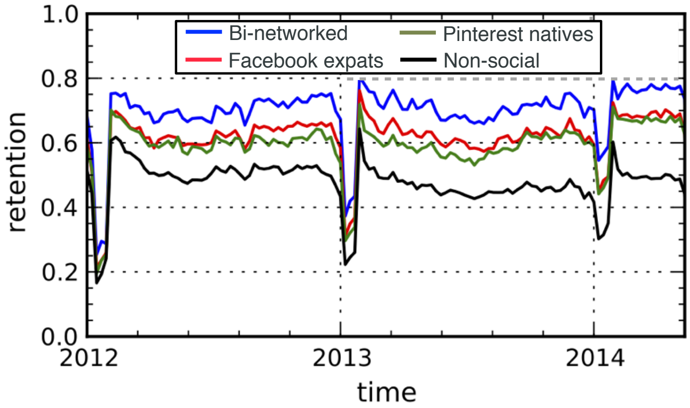
(a)

(b)

Here we measure the activity level of users according to the number of (a) pins (b) likes users made. It shows that Facebook expats whose social links are entirely copied from Facebook are the least active, whereas bi-networked users with a mixture of native and copied links are the most active.

our **DATA-SOC** using the social structure based user groups. Since the **DATA-SOC** dataset is collected with a snowball crawling method, non-social users are not available. Thus, we compare the activity of these three types of users in Figure 5.2 and find that *Facebook expats* whose social links are entirely copied from Facebook are the least active, whereas *bi-networked* users with a mixture of native and copied links are the most active. *Pinterest natives* who do not copy at all are in the mid range.

Figure 5.3
Social Users are More Likely to Return



The retention rate of social users is higher than when users have no social repins. At the same time, users who did both copied and native repins (Bi-networked) have the highest retention rate.

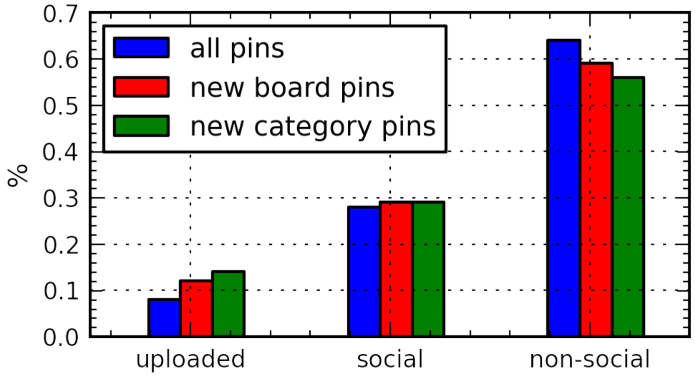
5.2.2 Social users are more likely to return soon

We next check whether the active users are also consistent, by measuring user retention as defined in §5.1.2. Figure 5.3 shows the retention rates every week amongst different kinds of users. Apart from the annual drop in retention rates corresponding to the end of the year holiday season, retention rates for each kind of user stays at roughly the same level throughout the year. First, we divide active users in each week into social users and non-social users, according to whether they have made any social repins in the given week (i.e., interaction version of definition). These results show that retention of social users is higher than non-social users. We find that only about 50% of users who have not made any social repins ("non-social", black line) return to the platform within a week, whereas more than 60% of users that interacted with their social friends do return. That is, users with social interactions are more likely to return and be engaged with the platform. This is statistically significant ($p = 1.68 \times 10^{-35}$).

Furthermore, we divide users into three types. In Figure 5.3, we can see that bi-networked (blue line) still show the highest retention rates, compared with Facebook expats and Pinterest natives.

In summary, above results suggest that the Pinterest social network serves the important purposes of bonding and social grooming: *Users will exhibit distinct behaviour patterns in social and non-social information seeking. The core and highly active members will be engaged socially, and return*

Figure 5.4
The Source of Pins in our DATA-USER Dataset.



We examine whether users' images are "uploaded" by users themselves, repinned from "social" friends, or repinned from "non-social" strangers. We term the first pins users' boards as "new board pins", while the first pins of users' category as "new category pins".

to Pinterest for social activities.

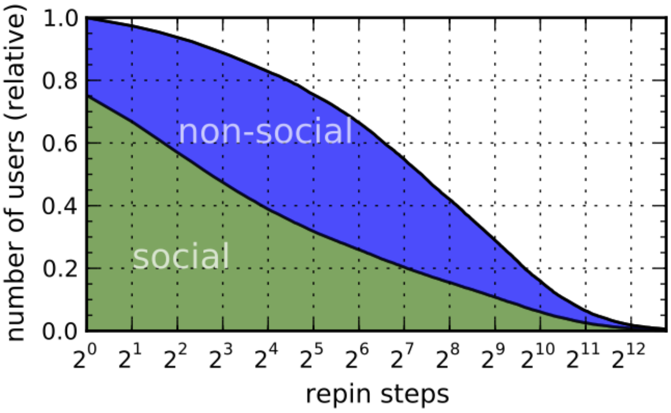
5.2.3 Social network is *not* critical for information seeking

So far, we have shown that social networks are important for users. However, does it mean that social networks are always useful? In this section, we check what is the role of social networks on information seeking. A core function of content curation sites such as Pinterest is to enable users to find the information that suits their interests. Therefore, we might expect the ability of social networks to provide access to new information [Putnam, 1995], would be important on Pinterest. To study this, we first compare social vs. non-social means of acquiring new information.

Following the measure of information seeking we defined in §5.1.2, we divide the 10.3 million pinning actions in **DATA-USER** into *uploaded pins*, *social repins* and *non-social repins* in Figure 5.4. In theory, each of these sources can be important to different extents in different kinds of information-seeking activities. Therefore, we also examine the first pin of a user in a *new category* or *new board*, which may have different information seeking patterns with subsequent pins.

From the figure, we find that uploads are fewer in number than repins, whether social or non-social. However, the striking result is that

Figure 5.5
Long-term Dynamics of Social vs. Non-social Repins.



showing the proportion of social and non-social repins vs. the "maturity" of user on Pinterest, as measured by the number of repins made since joining. Note that there are fewer and fewer users as the "age" in terms of repin steps increases.

across all kinds of information seeking, whether for the first pins in new boards/categories, or for subsequent pins, *non-social means of finding information from other users dominate over social repins.*

We then ask whether time matters: i.e., Do social repins become more important as the user matures and conducts more activities on Pinterest? Because the time between two pins may be widely different across users, we measure user age in terms of *repin steps*, the number of (re-)pins made since joining Pinterest. In Figure 5.5 we examine the entire history of activities of all users in **DATA-USER**, as they "age" in Pinterest by accumulating more activities. For each new repin activity of a user, we check whether it is a social or non-social repin. From the figure, we can see the proportion of social repins is much larger than non-social repins when users just join Pinterest. But the difference between the two proportions is reduced as users become more experienced in the platform. *This shows the growing importance of social repins for users with large numbers of repins, and for users just getting started on Pinterest (having very few repins), though a non-trivial portion of repin activities is still non-social.*

5.2.4 Discussion

In summary, our experiments in this section show that social networks are important in retaining users, but are not critical for users' information seeking. To understand these results, we turn to the concept of *social* *capital*. Social capital has been an influential concept in several fields such as economics [Knack and Keefer, 1997], development and policy [Woolcock and Narayan, 2000], organization theory [Nahapiet and Ghoshal, 1998; Tsai and Ghoshal, 1998], and, most relevant to us, modern sociology [Portes, 1998]. It is widely used to understand the value of "social structure to actors as resources to achieve their interests" [Coleman, 1988].

Various forms of social capital have been recognised. Putnam [1995], a key figure in the literature on social capital, distinguished between *bridging* social capital, which is derived from links that provide access to new information through social acquaintances, and *bonding* social capital, which arises from strong links that provide social grooming and support from within the community.

Consistent with the theory of bonding social capital, social links have a strong bearing on engagement and this has been demonstrated in Facebook Burke et al. [2009], Twitter Macskassy [2012], etc. Our findings in this section show that social users of Pinterest contribute the majority of activity, and have a higher probability of returning to the site, suggesting that bonding social capital is also important for in the context of Pinterest. Dividing social links into copied and native, we also observed different retention rate for users with involved in different types of social statuses, for example, bi-networked users are most likely to return soon. This drives us to ask, what is the difference between native and copied networks, and which type of network is better at promoting social interaction. We will explore these questions in §5.4 and §5.5.

Interestingly, Ellison et al. [2007] find that several forms of social capital, including *bridging* social capital are important in Facebook. In contrast, we find that social network is not critical for information seeking. This require more discussion. In §5.3, we will try to understanding these findings through mining user interest alignment and evolution.

5.3 Explaining Non-social Repins through Interest Alignment and Evolution

In this section[1], we attempt to explain why social network is not critical in information seeking. We first establish that Pinterest interactions are, as designed and expected by the platform, interest-based. Then we show how user interest evolution can explain non-social repinning.

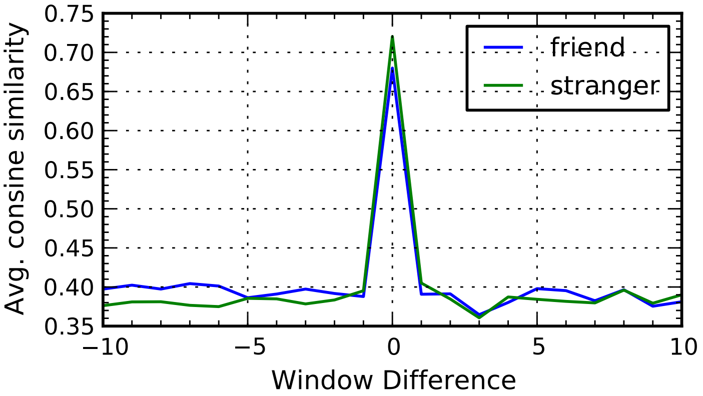
Core to our method is measuring how user interests are aligned before, during and after a repin activity. To do this, we make use of the implicit interest-based categorisation that happens when an image is pinned: Each pin is placed onto a so-called "pinboard", and involves an implicit categorisation of the image into one of the user's pinboards. As described in §3.1, most of pinboards belong to one of 32 global categories used on Pinterest, describing different interests (e.g., "DIY", "architecture", "fashion", etc.). Thus, the relative proportion of a user's pins in different interest categories can be used to construct a per-user *interest vector*. We use the cosine similarity between two users' interest vectors to measure *interest alignment* or *similarity* of the interests of the two users. In the following, we measure interest alignment using a time window of 1 day, but our results also hold true for weekly and monthly windows.

Figure 5.6 shows how the interest alignment evolves just before, at, and after a repin activity, both for social and non-social repins, averaged across all repins. The well-defined spike around the repin time strongly suggests that a temporary alignment of interests is the likely cause of repins. In other words, a user A is likely to repin another user B's pin, if A's current interest (at least temporarily) evolves to be similar to B's pin.

This naturally leads us to ask how user interests co-evolve over a time period. First, we check the long-term evolution of a single user's interests. Although they could be measured in terms of repin steps as in Figure 5.5, it is more meaningful to test how interests change in real time: a user who repins after several days might be more likely to repin something entirely different, as compared to a user who makes two repins within a few minutes or seconds of each other. Therefore, we again divide a user's timeline of repins into day-long windows, and compute interest vectors over each window. We then measure the self-similarity between each win-

[1] In this section, Figure 5.6, 5.7, and 5.8 are generated by Dr. Nicolas Kourtellis based on the **DATA-USER** data supplied by me, with myself participating in the experiment design.

Figure 5.6
Interest Alignment

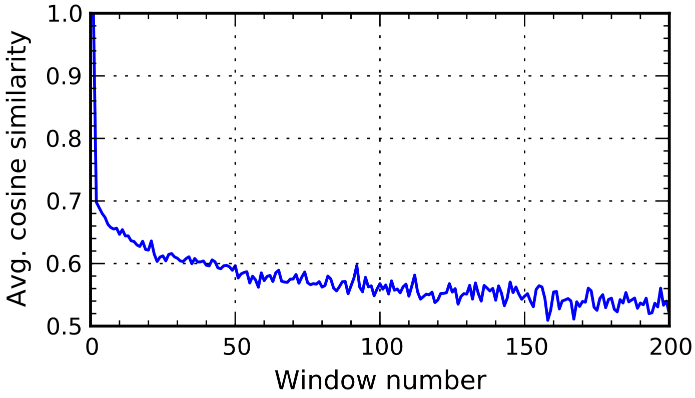


Interest alignment between pinner and repinner at different times before and after the repin action at time T. Repin time is set at t=0, and time windows before and after repin time are shown relative to repin time. A data point at a given window difference t for the blue (green) line shows the average similarity at time T = t across all social (non-social) repins. That is, the average similarity between a pinner and a social friend (non-social stranger) at time T = t relative to the time of the corresponding repin. Both social and non-social repins exhibit similar spikes in similarity just around the repin time.

dow when there is a repin and the previous window where the user had an activity. The results, plotted in Figure 5.7 , show that users' interests can change rapidly, although users are in general interested in the "same kind" of content, leading to a baseline cosine similarity of more than 0.5 even when there are 200 days between successive repins.

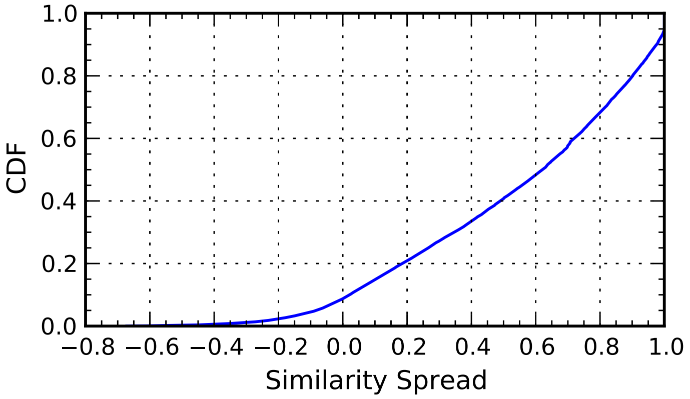
We therefore hypothesise that non-social repins may be a result of a user "drifting away" from her friends' interests. We test this at scale, by measuring the average *spread* or difference in similarities between two sets of users: a user and the stranger being repinned on the one hand, and a user and all her friends on the other. Note that this similarity spread is being computed at the time of repin. Figure 5.8 shows the cumulative distribution of similarity spread across all non-social repins, i.e., it shows the similarity spread distribution, by computing the spread each time a user repins a stranger. Interestingly, the majority of such interactions (> 80%) is with positive similarity spread, i.e., users have higher similarity with the stranger they repin from, than with their friends. We summarise the finding as: *Interests need to be matched when a user repins another user's pin. User interests evolve over time, and therefore, a stranger may be more*

Figure 5.7
Long-term Dynamics of Interest Self-similarity



This shows cosine similarity between interest vectors of the same user, between successive login days.

Figure 5.8
Users Drift away from Social Friends to the Strangers

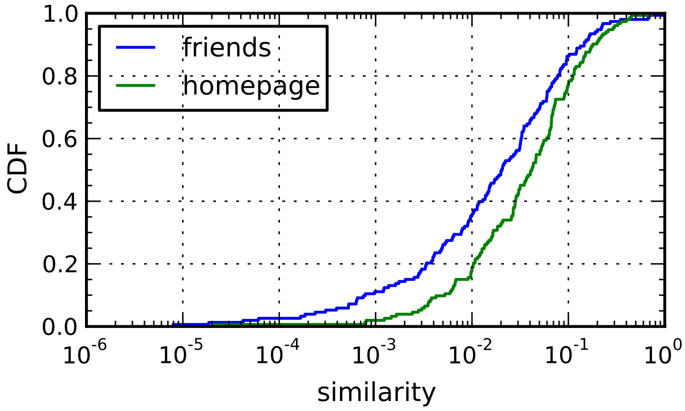


The CDF of the similarity spread or difference in similarities between (i) a user and the stranger being repinned, and (ii) the same user and all her friends, as measured at repin time, for all non-social repins. In a majority of cases, the spread is positive, indicating that the user is temporarily more similar to the stranger being repinned than all her friends.

similar in interests than a friend, leading to a non-social repin.

This result is consistent, regardless of time window size considered. Thus, as user interests evolve, there appear to be strangers who would be closer to their interests than friends accumulated over time. Such strangers and their pins are not hard to find: Pinterest highlights the most recent pins on the platform on its home page. Figure 5.9 shows that on average, the recent pins being highlighted on the home page at a given point

Figure 5.9
Homepage Pins vs. Friends' Pins



It shows that pins on homepage are more similar to a user's recent pins than to her friends' pins. For this study we use repin activities from ***DATA-IMG****. All of these pins were categorised using vectors constructed from 1000 objects detected on each pin, via Caffe [Jia et al., 2014] (refer §6.2.2). Cosine similarity was computed at random time points between vectors of pins of users and (1) pins of their friends and (2) pins of random users featured on the homepage at that time.*

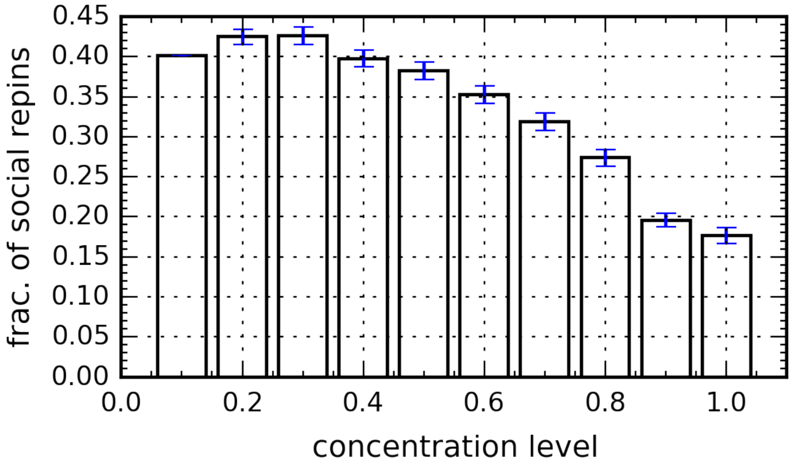
in time are likely to be more similar to user's current interests than the recent pins of the user's friends. Thus, given access to the homepage which proves to be a simple and easy to find source of interesting information, bridging social capital and social-based information seeking becomes less critical.

5.3.1 Social sessions are less goal-oriented

It thus appears that although social activities are extremely positive for bonding users, Yet, social assistance appears to be less efficient than the homepage for individual users to find pins suiting their interests and goals. To understand why, we turn to the recent finding of Linder et al. [2014] who identified two types of information seeking behaviour in Pinterest: (a) casual browsing, when they do not have a particular goal in mind, and (b) specific searching, when users only respond to (repin) a specific type of image. We ask whether the former is richer in social repins than the latter.

To distinguish different kinds of information seeking by the same user, we divide the user's pin timeline into "sessions", drawing session boundaries whenever there is a gap of T=6 hours or more between two consec-

Figure 5.10
Session Concentration vs. Social Repin Fraction



Users tend to have more social repins in sessions without a goal (lower concentration) than in sessions with higher concentration. In this figure, we first separate sessions into different bins according to their average concentration level. Next, for each bin, we random sample 1000 sessions and compute the average fraction of social repins. We repeat the experiment for 1000 times, and report the average result of all of 1000 experiments.

utive pins. Thus, a session is all pins which are "close" to each other in time; similar results are obtained for other small values of T (e.g., 1 or 3 hours), but are not reported due to space.

In each session, we define the major category as the category into which the user has repinned the most number of images, and define the *concentration level* of this session as the fraction of images of the session that have been repinned to the major category. We expect that the higher the concentration level, the more likely users are to be searching for a specific category of information, and less likely to be browsing casually, without a goal. For each browsing session, we also compute the fraction of repins which are social. In Figure 5.10, we compare the concentration level and social repin fraction. We notice that levels of less concentration are associated with higher levels of social repins, i.e., *when users do casual browsing without a specific information seeking need, they are more likely to be social, whereas when they have a specific information need, they are more goal oriented and find the information they need through the most efficient means (which may not be social).*

Collectively, findings in §5.2 and this section shows that users appear to be bonding and forming communities that are close, active and engaged through social activities. *In other words, the Pinterest social network, despite being intended for interest-based interactions, appears to show more evidence for bonding function than information seeking.*

5.4 Structural Benefits of Copying

Based on our study in previous sections, we know that although social network is not critical for information seeking, social users are more active and more likely to return soon, especially those who made social interactions. This drives us to ask, which type of social networks is better at promoting social interaction, the one copied by social bootstrapping or the one created natively. Thus, in this section, we turn to **DATA-SOC** and **DATA-ACT** datasets and empirically compare the social structure properties of the copied and native networks. We consider three properties that are *expected* to improve social interaction [Macskassy, 2012; Teng and Adamic, 2010]: reciprocity, clustering coefficient and giant component.

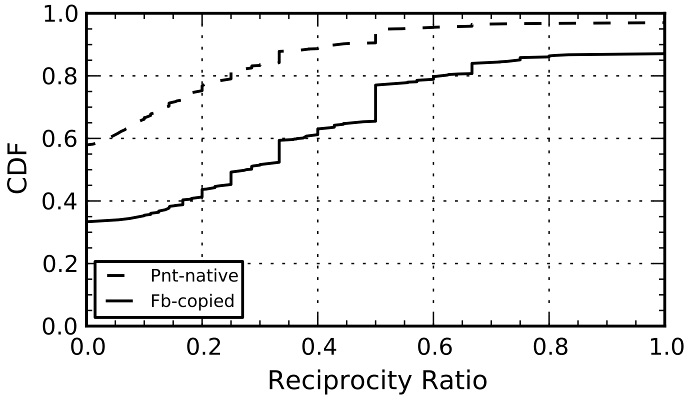
Copied network has higher reciprocity. Reciprocity is known to indicate positive bidirectional interaction between a pair of users, which is also known to increase user longevity in the system [Fehr and Gächter, 2000; Boguñá and Serrano, 2005; Zhu et al., 2014]. Here, we attempt to examine the effect of copying on creating structurally stronger bidirectional social ties, by defining *reciprocity ratio* as the fraction of social links that are *reciprocal*, or bidirectional. For a node in a network, let her follower (or following) set in the target network (e.g., Pinterest) be ind (or out) and her friend set copied from the source network (e.g., Facebook) be fr. Then the reciprocity ratios of that user in the entire target networks, and its partition into Fb-copied, and native networks are as follows:

$$R_{copied} = \frac{|fr \cap ind \cap out|}{|fr \cap (ind \cup out)|},$$

$$R_{native} = \frac{|(ind - fr) \cap (out - fr)|}{|(ind - fr) \cup (out - fr)|}.$$

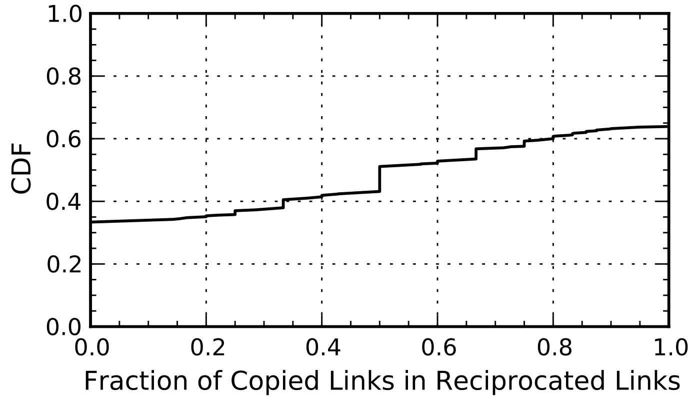
Figure 5.11 shows that in Pinterest, the reciprocity ratio is higher in links which are also found on Facebook, than on natively created links. Although in some cases, a link copied in one direction could be reciprocated by the other party merely in order to be "social" or "polite", the link creation creates an opportunity for social interaction on the target

Figure 5.11
Reciprocity of Copied and Native Subgraph



CDF of per-user fraction of links reciprocated in copied and natively created networks. More links are reciprocated in the copied network.

Figure 5.12
The Fractions of Copied Links among Reciprocated Links

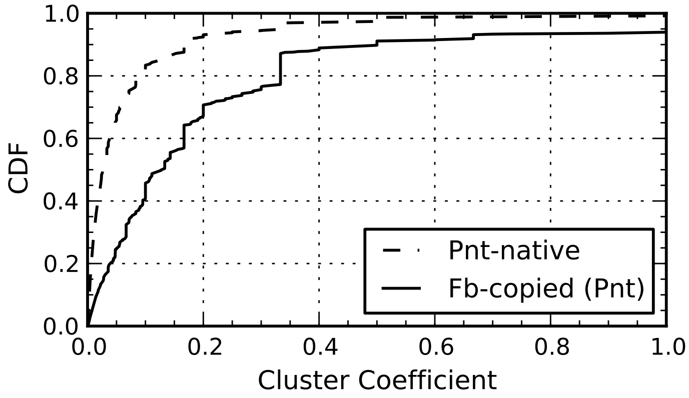


CDF of per-user fractions of Fb-copied links among reciprocated links in target networks. Many users have high proportions of Fb-copied links implying that copied links are important for establishing bidirectional or reciprocated relationships.

website, and reciprocity could promote positive bidirectional social interactions. Figure 5.12 shows that copying is extremely important for establishing reciprocal relationships, because a large proportion of users' reciprocal links are in fact those copied from Facebook.

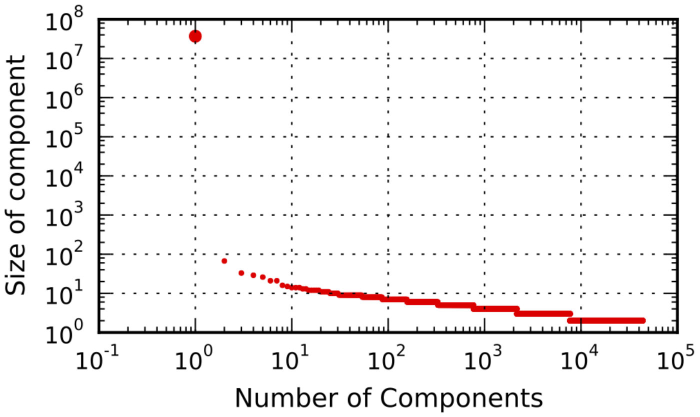
Copied network shows higher clustering. Next we explore the impact of copying on another popular measure of a strong social structure, clus-

Figure 5.13
Clustering of Copied subgraph



Per-user CDF of clustering coefficients in natively created and copied subgraphs of Pinterest (0 valued-points not shown). Clustering coefficients are higher in the copied network.

Figure 5.14
Connected Components of Copied subgraph



Distribution of the sizes of connected components on the FB-Copied network in the Pinterest datasets.

tering or the degree to which users share common friends. Figure 5.13 shows that in Pinterest, users have much higher clustering co-efficients on the copied network than on the network natively created on the website. Thus copying not only promotes reciprocal social interactions, but also creates a much denser social network structure in the target website.

Copying enhances connectivity. The increased clustering and reciprocity are properties relating to local structure around a node. Copied links are also crucial for connectivity, a global (network-wide) property. Figure 5.14 confirms that Pinterest copied network have a giant component. The largest component comprises 91% of all the connected nodes (i.e., nodes present on both source and target networks). Furthermore, this component encompasses 53% of all the nodes in the corresponding target network.

Link Bootstrapping Sampling model. Our empirical findings on social structure properties are also consistent with the findings from an analytical model named Linking Bootstrapping Sampling[2]. In the model, the process of social bootstrapping is modelled as a simplified random sampling process. It is a two-step model, which is a variation of the induced subgraph sampling process [Kolaczyk, 2009]. In the first step, users of the target network have to self-select to connect their accounts on the target network with the source network. In the second step, users have to select which of their friends from the source network to import onto the target network. Under this stylised model, we obtained expressions for the resulting degree distributions of the copied network and a condition for the emergence of a giant connected component in that network. It is shown that the social bootstrapping process tends to produce a giant connected component quickly and preserves properties such as reciprocity and clustering up to linear multiplicative factor.

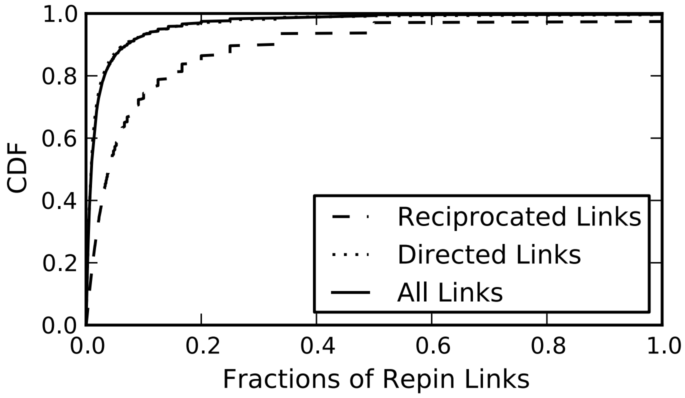
5.4.1 Copied network see more interaction

So far, we have shown that copying links results in a higher level of reciprocity and clustering, representing a stronger and denser social structure than its low-clustering and low-reciprocity native counterpart. Now, we ask whether the benefits of these structural properties are seen in the social interactions of the target network.

In order to determine the benefits of a close-knit structure, we define the *social repin network*, as the subgraph of links in the Pinterest network over which at least one social repin happens in our data. Then, we examine how the social repin network selectively samples the underlying network of Pinterest. First, we ask what proportion of a user's reciprocated and directed (unreciprocated) links have incurred repins. Figure 5.15 shows that repins happen more easily over reciprocated links. Next, in

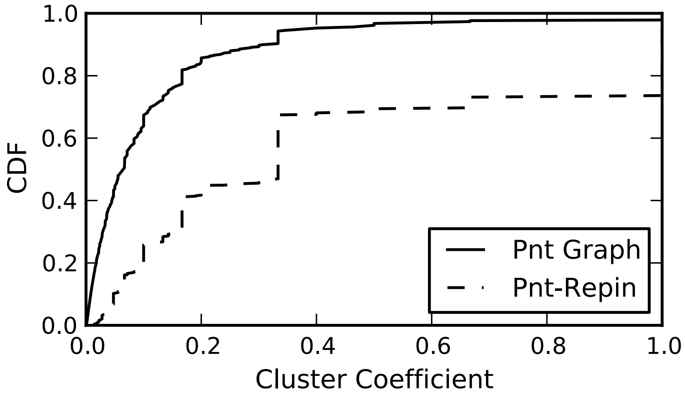
[2] The model is proposed in a collaborative effort and published in a WWW conference paper [Zhong et al., 2014] with me as the first author.

Figure 5.15
Repin Network Samples Reciprocal Links More



CDF of fraction of users' reciprocated and unreciprocated (directed) links, which are included in the repin network. A greater fraction of reciprocated links than directed links have repin activity.

Figure 5.16
Repin Network Shows Higher Clustering

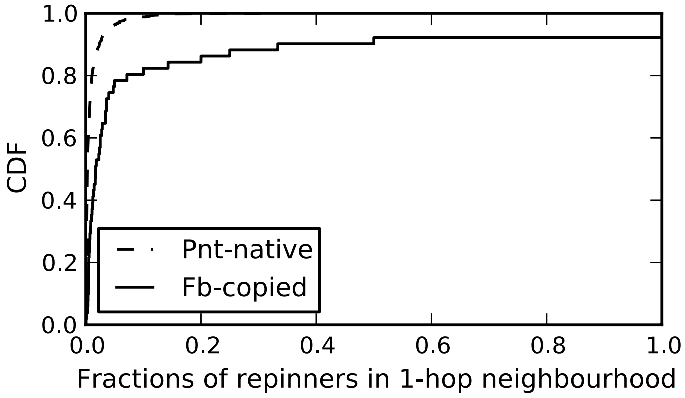


CDF of users' clustering coefficients in the Pinterest graph and the repin network. The repin network has higher clustering, indicating that users' social repins are directed more at closer friends.

Figure 5.16 we compare the clustering coefficient of users in the social repin network to the clustering coefficient of the underlying graph. Users have significantly higher clustering coefficient when we remove the links over which no repins happen. This suggests that social interactions tend to be directed towards the closer friends of a user, within highly clustered communities.

These results show that the social repin graph is richer in reciprocated links and is more highly clustered than the underlying network. Since

Figure 5.17
Repin Network Selects Copied Links More



CDF of the fractions of users' natively created and copied (Fb-copied) links which are sampled by the repin network. Copied links tend to have more repins.

reciprocal links and high clustering nodes will have more social repins, it is straightforward to infer that the copied network, which is higher in both reciprocity and clustering coefficient, should promote more social repins. This is proved by Figure 5.17, which shows that a larger fraction of social repinners tend to be from the copied network than from the natively created network.

5.5 Structural Advantages are *not* Critical for Advanced Users

We have shown copying links provides instant bootstrapping advantages by incurring a close-knit local structure (i.e., high reciprocity and clustering). Next, we take a long-term view of social bootstrapping to explore whether there is a limit to which a user can copy links from Facebook. Because beyond a certain point, a user may no longer find other Facebook friends to copy over. It is natural to ask whether this creates engagement bottlenecks for users as they become more prolific on the target network, or whether they find alternative solutions.

In this section, we describe a collective "weaning" process, through which users move away from their reliance on Facebook copied links to building new relationships natively on target websites. We find that users, as they become more active and influential on Pinterest, establish more native links within these services and copy less from Facebook. We discuss why users "go native" in this way and suggest a possible cause: through native links, users may find others similar to themselves on the target website.

To explore the relationship between users' attributes and their copying activities, we use a method similar to §4.3: Firstly, we separate users into bin based on the user's value of the attribute considered (e.g., the activity and influence level). Next for each bin, i.e., for each value of the attribute being considered, we random sample 1000 users and compute corresponding measures for copying activities. We repeat the experiment for 1000 times, and report the average result of all of 1000 experiments.

5.5.1 Active and influential users copy fewer links

In order to study levels of copying, we introduce a measure called the *copy ratio*. Denoting the set of all friends in the target network as all and the friend set copied from the source network (i.e., Facebook) as fr, the copy ratio in a undirected network, is defined as:

$$CR = \frac{|all \cap fr|}{|all|}$$

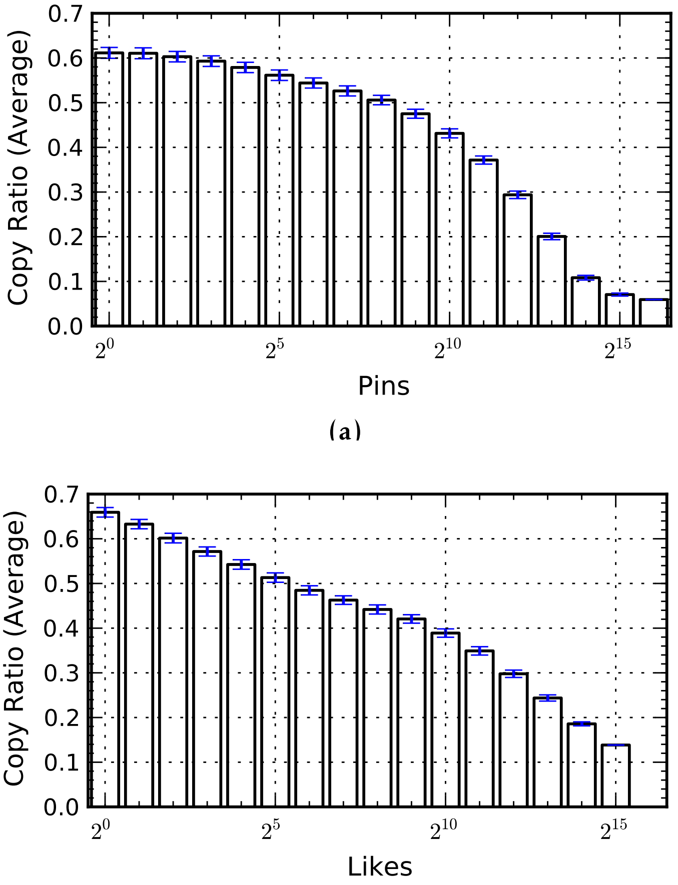
For a directed network, representing a node's follower (resp., following) set in the target network (i.e., Pinterest) by ind (resp., out), we define the *follower copy ratio* and *following copy ratio* as:

$$CR_{ind} = \frac{|ind \cap fr|}{|ind|}$$

$$CR_{out} = \frac{|out \cap fr|}{|out|}$$

we examine how the copy ratios change as activity levels increase in Figure 5.18, for the case of pins and likes in Pinterest. This demonstrates a clear inverse relationship between the activity levels and copy ratio, with users who pin a lot tending to have lower levels of copying—that is, higher activity levels are associated with lower copy ratio. Figure 5.19 shows that this result extends to measures of influence. We find that users who are influential, measured by repins, tend to have lower copy ratios. Overall, the results above indicate that as users settle down on the new service and become more active and influential, their investment in natively formed links increases proportionally.

Figure 5.18
Active Users Have Lower Copy Ratio



(a)

(b)

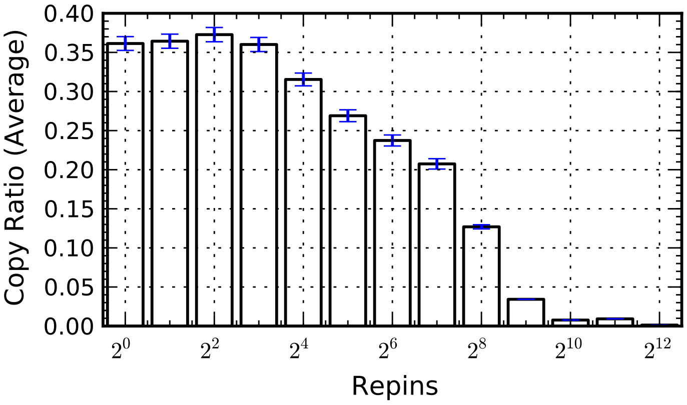
Higher activity levels measured by pins and likes are associated with lower following copy ratio in Pinterest

5.5.2 Influential and active users remain social, but with native rather than copied friends

Next we take a deeper look at the relationship between the increase in activity or influence level of a user and his level of social interaction. In order to quantify the level of social interaction, we again define the concept of a *social repin* as a repin where the user who is repinning follows the original pinner. We define a user's *social repin ratio* for activity (or influence) as the fraction of social repins made (or received) among all repins made (or received).

Figure 5.20 shows that users who are more active (or influential) tend to make (or receive) proportionally more social repins in relation to their

Figure 5.19
Influential Users Have Lower Copy Ratio



Users who are influential on Pinterest, as measured by repins, tend to have lower copy ratios.

activity (or influence) level, *confirming that social interaction continues to be increasingly essential* for active (or influential) users. This is consistent with the bonding effects of social networks we found in §5.2.

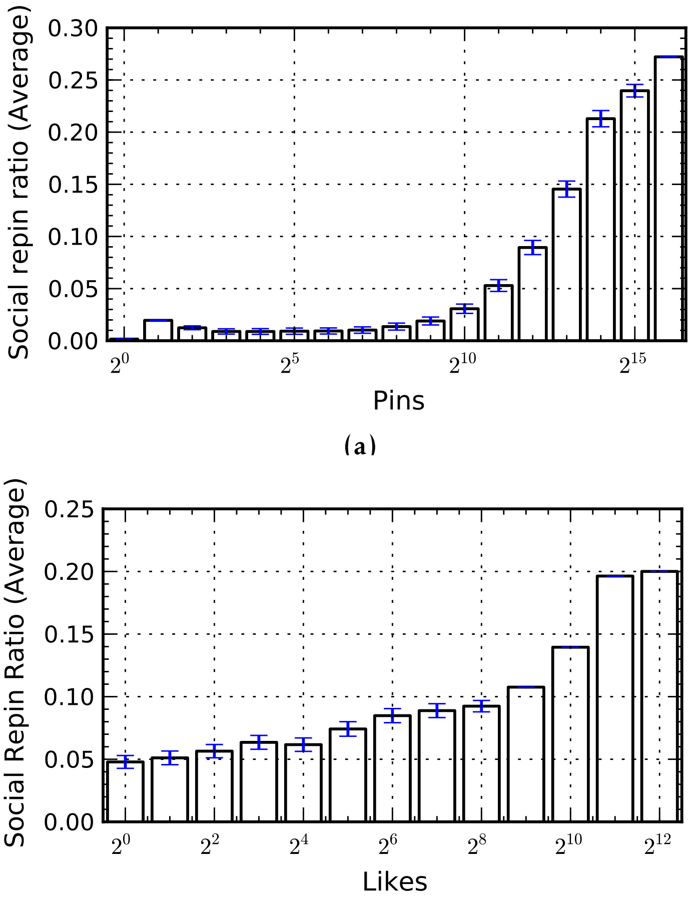
We also focus on social repins and ask whether copied links promote social repins. We define the *Facebook repin ratio* for activity (or influence) as the fraction of social repins made (or received) over Fb-copied links among all social repins made (or received). Figure 5.21 reveals that as activity (or influence) levels increase, social repins happening over copied links decrease.

5.5.3 Weaning, biases and community evolution

We conclude by asking how the nature of the target network community would evolve as users "wean" from copying to make more native links. To understand this, we study user preferences or biases in the kinds of links they copy and the links they make natively. We also seek to understand the role that copying plays in creating more native links.

User studies in §4.1 identified that Pinterest users most value the social aspect of the service that helps them find people with similar tastes in pictures. Therefore, we examine whether natively created links on Pinterest enables discovery of individuals with a more similar taste than those with copied links. Specifically, if I_1 is the set of user u_1's board categories,

Figure 5.20
Influential and Active Users Remain Social



(a)

(b)

(a) Social repins increase proportionally as activity level, measured by pins, increases. (b) The same increasing trend is shown for influence, measured by likes received.

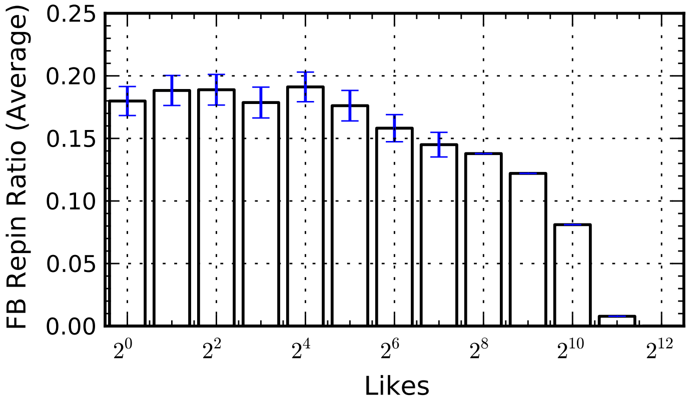
and I_2 is the set of u_2's, we define their similarity as

$$s = \frac{|I_1 \cap I_2|}{|I_1 \cup I_2|}.$$

Figure 5.22 confirms that according to this measure, users connected by native Pinterest links are more similar to each other than those connected by Facebook-copied links. Figure 5.22 also shows that there is no difference in similarity between users who are copied and users who are not copied over from Facebook. This implies that users are not selecting Facebook friends to copy based on similarity. This also align with our previous findings in §5.3.1.

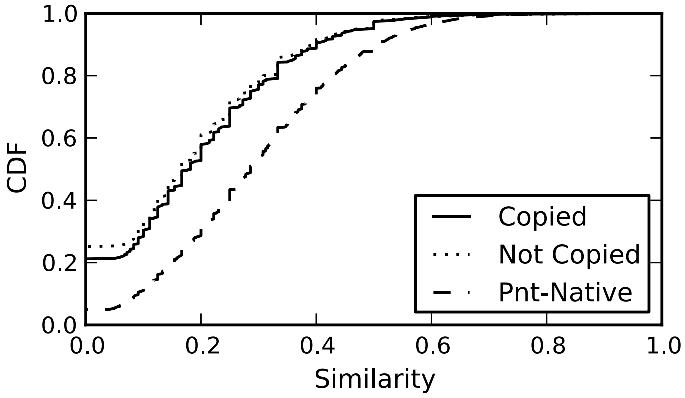
In Figure 5.23, we study whether closeness of friends has a role in deciding which friends to copy. In our analysis, we use the similarity of

Figure 5.21
Influential Users Interacted More with Native Friends



The fraction of social repins over links copied *from Facebook decreases proportionally as influence, measured by likes received, increases.*

Figure 5.22
Native Friends Have Similar Interests



CDF of similarity between users linked by copied, uncopied and pnt-native links, showing that native friends are more similar to a user than copied friends, but copied and uncopied friends do not differ significantly in tastes.

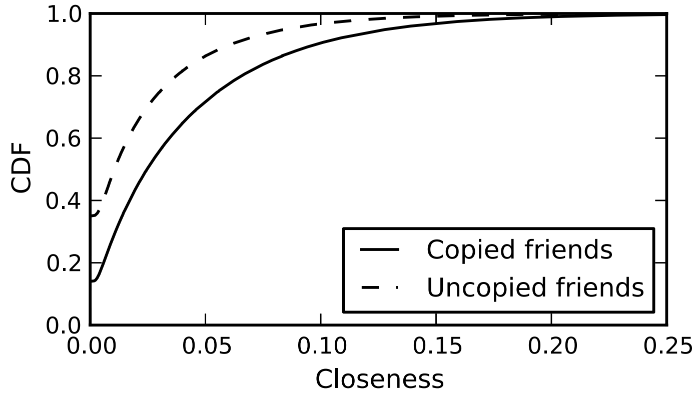
users' friend lists to show their closeness: if A's friend list is L_A and B's is L_B, we say their closeness is

$$c = \frac{L_A \cap L_B}{L_A \cup L_B}.$$

Figure 5.23 shows that closeness between copied friends is higher than between uncopied friends.

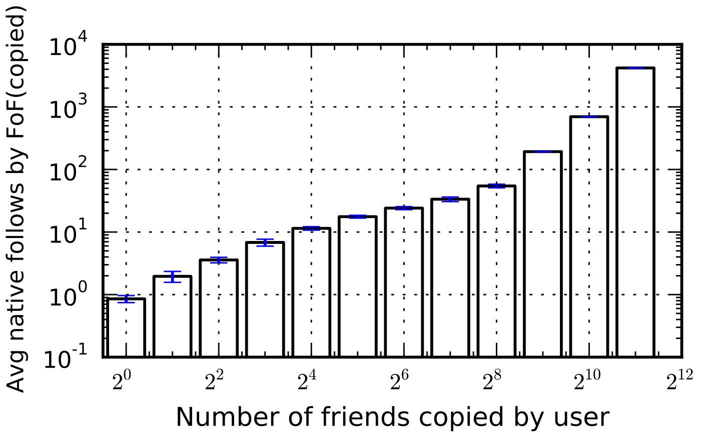
Together these results suggest that Pinterest users tend to use the "friend finder" tool to copy close friends they know from established source networks like Facebook, but when they discover new friends on the target

Figure 5.23
Close Friends Are Preferentially Copied



Per-user CDF of closeness between copied and uncopied friends. It shows that copied friends are closer than uncopied friends.

Figure 5.24
Native friends are FoFs of copied friends



The more friends a user copies and follows in the target network, the more follows she gets, from exposure to friends of the copied friends.

network, they tend to prefer users with similar tastes. Thus, as native links become more important and numerous than copied links, we expect the target network to become more interest-based.

However, copying continues to be important for the creation of native links over which interaction happens, even in networks like Pinterest, where copying appears to be governed by norms of social closeness – Figure 5.24 examines links over which social repin interactions happen over a sample representative day, and shows that users who have copied more of their friends from the source to target network, tend to have more native followers who are *friends of her friends* on the target network. i.e., copying creates the opportunity for users in the immediate social community of nodes in the copied sub-graph to discover and follow them, creating new native links, over which social interaction happens.

5.6 Conclusions and Discussion

In this chapter, we have used our large Pinterest **DATA-SOC** and **DATA-USER** datasets to unpack the role and utility of the social network on Pinterest. We find that social networks are not critical for information seeking, but social users are more active and are more likely to return soon. Interestingly, Ellison et al. [2007] find that social network in Facebook could provide access to new information to users through social acquaintances, which is exactly the opposite with our findings on information seeking and thus requires some discussion. In comparison with Pinterest, and other content-driven sites, Facebook does *not* have a homepage that highlights recent or interesting content from non friends. We conjecture that this gives a greater role for the social network and especially weak ties in providing new information on Facebook, leading to the importance of bridging social capital. In contrast, we find that a user's pins are much more similar with pins featured on the Pinterest homepage, than with pins of their social contacts (Fig. 5.9). Furthermore, users only match in interest space close to the time of interaction (Figs 5.6, 5.8). Thus, even if repinning a stranger, it may not make sense to befriend them for future information seeking. Therefore, in comparison with the Pinterest homepage, bridging social capital may not be as important on Pinterest as on Facebook, and users are just as likely to repin content from strangers as from friends.

However, consistent with the theory of social capital, we find that social users of Pinterest contribute the majority of activity, and have a higher probability of returning to the site. Social links have a strong bearing on engagement and this has been demonstrated in Facebook [Burke et al., 2009], Twitter [Macskassy, 2012], etc. The present work confirms this phenomenon on Pinterest as well, suggesting that bonding social capital is important for its functioning.

From the perspective of the platform operator and the collective community of users on the platform, the Pinterest social network is critical for healthy operation and to drive user activities. We may hypothesise that the social network could also be useful from an individual user's perspective, because of positive feelings of engagement and social interaction; however this requires further research (e.g., through user studies), to be confirmed.

In addition, this chapter also studied the impact of "social bootstrapping"—the act of copying one's social ties or links from a source network to a target network. This is a popular practice enabled by many new social network services and has implications on how a new online social network community can grow quickly. We gathered massive amounts of data from Facebook and Pinterest involving tens of millions of nodes and billions of links to understand this new phenomenon. Among a number of findings, we highlight that a "copying" process is useful to initiate social interaction in the target network, as one may expect. However, a "weaning" process, where a user moves away from copied social links and builds social relationships natively in the new network is essential for longer lasting user engagement. To the best of our knowledge, this chapter is the first this kind of study to utilise large-scale cross network data in understanding the interplay between heterogeneous services in terms of bootstrapping a network and engaging users to form a cohesive, interacting community.

After understanding the role of social networks in social curation platforms, in next chapter, we will dig deeper to understand users interactions. Since social relationships are less important for user interactions, we will try to train a machine learning model that incorporates features like content, user preference and interest matching to predict interactions.

Chapter 6

Predicting Content Curation

A crucial aspect of online content curation sites is that content curated by one user is also (by default) made available to the rest of the users to curate. For instance, users on Pinterest can copy images pinned by other users, and "repin" onto their own pinboards. Interestingly, such reappropriation and curation of content discovered by other users (termed as "repins") is by far the most common activity on Pinterest, constituting about 90% of user actions, as compared to directly discovering and pinning new images, which constitutes only 10% of actions, according to our **DATA-ACT**.

In this chapter, we consider such social curation process as distributed computation process, and attempt to use a machine learning approach to automate the process, thereby obtaining a mechanistic understanding of the end-to-end process of social curation on Pinterest. Our study can make it easier to re-appropriate and re-categorise content for personal use: *Given a pin (image) and a user, we wish to predict whether the user would be interested in repinning the pin. Moreover, we wish to predict the pinboard onto which they would repin, and automatically suggest these to the user.*

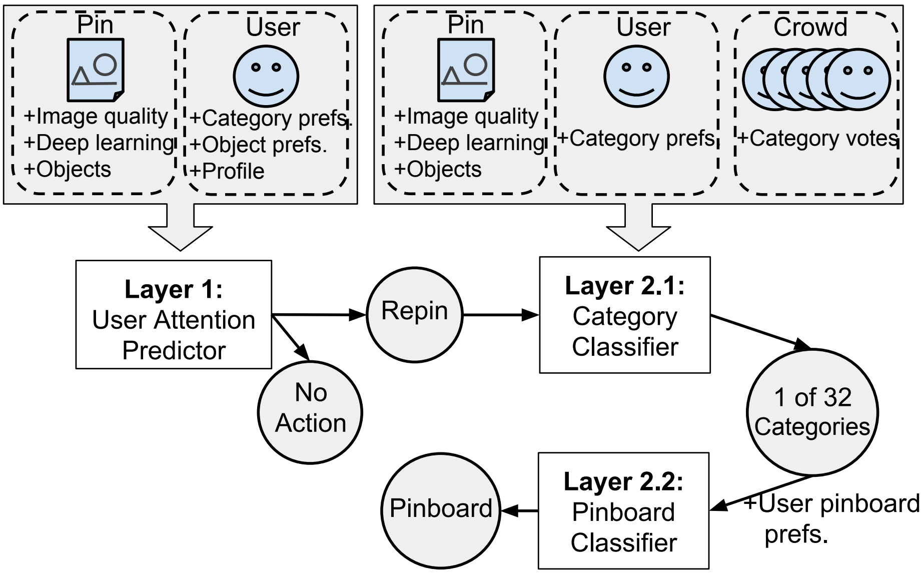
Summary of findings. We first visit the notion of agreement between curators and crowds in the context of Pinterest. Unlike traditional social bookmarking, pinning on Pinterest does not involve creating an explicit vocabulary of tags to describe the image. However, each pinboard may be associated to one of 32 categories defined globally for all users by Pinterest. Thus, even though each pinboard may be exclusive to a user, the act of pinning *implicitly* categorises the image. In other words, *Pinterest users can be seen as performing a massive distributed computation process, categorising images found on the Web onto an (extremely coarse-grained) global taxonomy of 32 categories.*

We find that this lower dimension approximation of Pinterest has several desirable properties: First, for a given image, a remarkable ≈75% of repins tend to agree on the majority category, although the level of agreement varies with the category. This enables us to robustly predict the category of an image. Second, users tend to specialise in a handful of categories; we use this to learn whether a user would be interested in an image given its category. Third, most users appear to have one or two boards per category. Thus, given the category of an image, and the user, it is often trivial to predict the board to which the user would pin the image. Based on these observations, we are able to build classifiers that, given that the user has curated an image, can predict the pinboard onto which the image would be repinned, and automatically suggest these to the user.

We augment this by showing that the *content* of the image can be used to predict whether the user would be interested in repinning the pin. To build this predictor, we derive several thousands of image-content features (Table 6.1), ranging from basic visual and aesthetic features to features extracted from the layer right before the final classification layer of the state-of-the-art deep convolutional network in Caffe [Jia et al., 2014], and by recognising objects in the image, using the same convolutional neural network. Using these features, we construct a supervised machine learning model that is able to assign an image to the majority category. We also learn user preferences for these image features, and predict whether the image would be repinned by the user.

We compose these classifiers in a pipeline of three layers (Figure 6.1). The first layer predicts whether the user will pay attention to a pin; the second predicts the category that the user will choose for the pin; and the third predicts the pinboard chosen given the category. Together this pipeline or cascade of classifiers is able to predict curation actions on Pinterest with an accuracy of 69% (Ground truth pinboard is in the top@5 predicted, with an accuracy@5 figure of 75%).

Figure 6.1
Prediction Cascade



Prediction cascade or pipeline to automate the manual curation of images on Pinterest

Chapter layout. The rest of this chapter is structured as follows. §6.1 demonstrates that Pinterest users are highly specialised in the categories they are interested in, and generally agree over category assignments. The rest of the chapter develops a pipeline of predictors: §6.2 sets the stage, discussing the cascaded structure of the predictors and the features used. §6.3 develops a classifier to predict whether a user will pay any attention to a pin. §6.4 then develops a two-stage multi-class classifier that predicts the board chosen by a user for a repin. §6.5 puts it all together, showing that repins can be predicted, both in terms of whether users would be interested in repinning a pin and which of their boards they would place it onto. §6.6 ends by discussing implications for the wider research agenda.

6.1 Predictability of repins

Curation on Pinterest is currently a highly manual procedure. Users select images that they like, and categorise it amongst one of several thematic collections or pinboards that they curate. Over 85% of respondents in a previous user study (§4.1) considered their pinning activity to be highly personal, akin to personal scrapbooking.

This paper, however, aims to automate this procedure, as much as possible. To this end, we examine the extent to which properties of the pin, or the user, can assist in suggesting an appropriate pinboard of the user for the pin.

We first take a pin-centric view, and ask whether repins of other users can help, and show that users tend to strongly agree on the *category* that they implicitly assign to a pin, via the category of the pinboard they choose. Next, we take a user-centric view, and show that users tend to be highly specialised in their choice of pinboards, focusing on a handful of categories, and also typically have very few boards within each category. We conclude by highlighting the implications of these findings, which we make use of in subsequent sections.

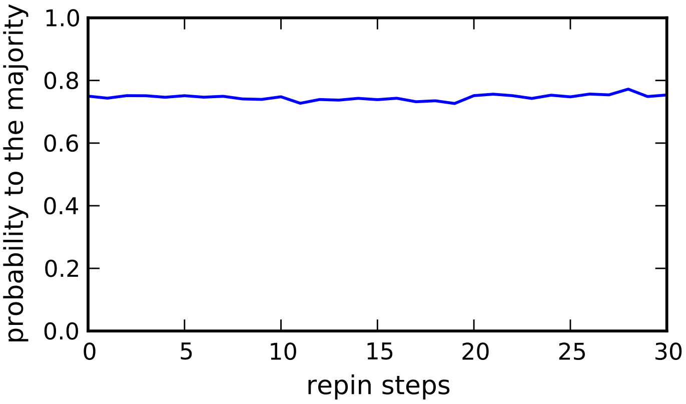
6.1.1 Pinterest users agree on image categories

Pinboards are personal to each user, and pinboards of different users typically share at most a handful of pins, if at all. However, each pinboard may be assigned to one of 32 categories which have been pre-determined by Pinterest. Therefore, we may regard a repin as implicitly assigning one-of-32 labels to an image, reminiscent of ESP [von Ahn and Dabbish, 2004], a human computation task which greatly improved label prediction for images. We ask whether users agree on the category assignment for images in the context of Pinterest.

Formally, each repin by user u of a pin p to a pinboard b whose category is c is interpreted as an assignment of the category c to pin p by user u; we denote this as $repin\_cat(p,u) = c$. After users $1..i$ have repinned a pin, one can define the count of the category assignments of c to p: $count_i(p,c) = |\{k | repin\_cat(p,k) = c, \forall 1 \le k \le i\}|$. We define the *majority category* of an image or a pin as the category chosen by the maximum number of repinners[1]. In other words, the majority category $\mathrm{maj}_i(p)$ after users $1..i$ have repinned a pin is the category with the maximum count: $\mathrm{maj}_i(p) = \mathrm{argmax}_c count_i(p,c)$. The final majority category $\mathrm{maj}_\infty(p)$ is the majority category after all r repins have been made. The consensus or agreement level after r repins can be computed as the fraction of pins in the final majority category after r repins: $agreement_r(p) = count_r(p, \mathrm{maj}_\infty(p))/r$.

[1] Note that we do not require >50% of pinners agree on a category, although this often happens.

Figure 6.2
Probability of Majority Pin



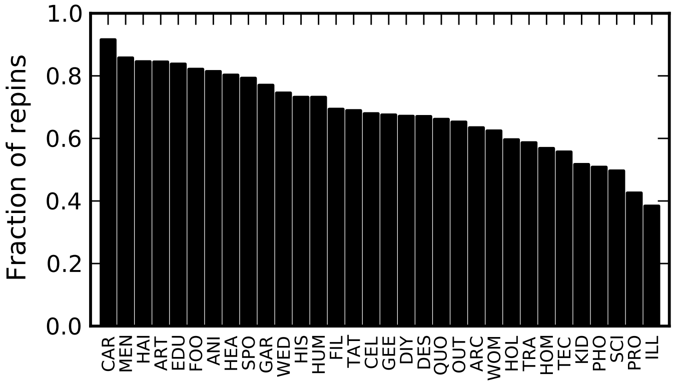
The category chosen by the ith pinner is independent of the category chosen by the previous $i-1$ pinners, and is the same as the category chosen by the majority of repinners with a remarkably high probability (≈0.75).

Whereas other curation systems (such as social tagging on del.icio.us) might push users towards agreement by suggesting tags [Golder and Huberman, 2006], in Pinterest, a pin does not come with a category of its own, and no category is suggested by the system or other users. Indeed, it is quite difficult on Pinterest to discover the category of a pinboard: Visitors to a pinboard's webpage can only determine its category through a meta-property in the HTML source code[2]. Even the owner of the board is only shown the category on the page for editing a board's details (not normally seen when the owner views her board). Because of this UI design decision in Pinterest, we expect that a user's choice of the Pinterest category to associate with a pin is made independently of other users or the system itself. Furthermore, the category choice is made only implicitly, as a result of an explicit choice made on which pinboard to place the image in. Thus, we expect this decision to be influenced by, for instance, whether the image being pinned fits thematically with the other images in the pinboard, and not by other users.

We first test our expectation that users individually decide on an image's category. We ask what is the probability $P[repin\_cat(p,i) = \mathrm{maj}_\infty(p)]$, that the ith repin of a pin agrees with the final majority category chosen for it. Confirming our intuition, Figure 6.2 shows that the ith repinner's

[2] Users often repin images from the homepage of Pinterest, and may not even visit the board of the original pin. Thus they may not know the category assigned to it by the original pinner, even if they can read HTML.

Figure 6.3
Levels of Agreement in Categories



The average fraction of pinners in the majority can vary across category, ranging from 91% (cars and motor cycles, or CAR), to 43% (Illustrations and Posters, or ILL). All except PRO (45%, Products) and ILL have a majority > 50%.

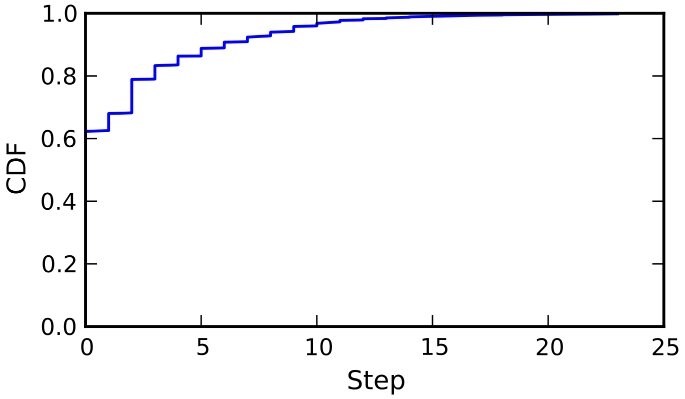
choice appears to be unaffected (either positively or negatively) by the choices of all the previous $i-1$ repinners. Furthermore, we see that there is a remarkably high chance (≈75%) that the category implicitly chosen by a user agrees with the majority. Figure 6.3 shows that the average levels of agreement can vary across pins of various categories, from 91% to 43%; and in all categories except Illustrations and Products, the final majority category has a clear majority of > 50% agreement.

Next, we ask how many pins it takes for the majority to emerge, *and stabilise*: Suppose we temporally order in ascending order the pinners of a pin p, starting with the first pinner as 1. We wish to know the number of repins required (smallest pinner number a) at which the majority category is the final majority category, and the consensus on the majority is unchanged by all subsequent pins. Formally, we want the smallest pin a such that $\text{maj}_k(p) = \text{maj}_\infty(p)$, $\forall k \geq a$. Figure 6.4 shows the cumulative distribution of the number of repins a required for stable agreement to emerge. In over 60% of images, this happens with the very first pin. After 5 repins, over 90% of images have reached consensus on the final majority category.

6.1.2 Pinterest users specialise in few categories

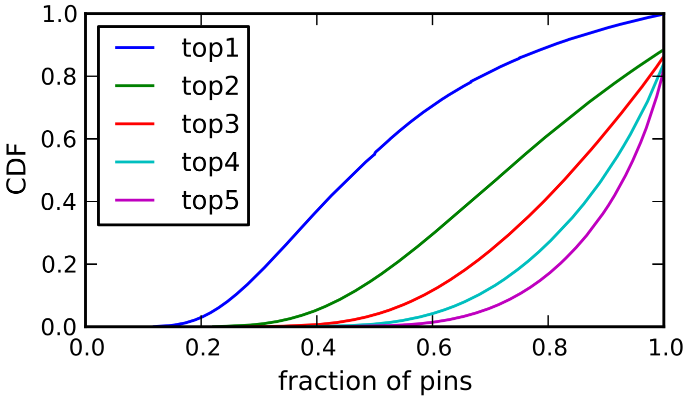
Having looked at a pin-centric view on predictability, we now look for user-specific patterns that can aid us in categorising their content auto-

Figure 6.4
Number Pins for Majority to Appear



Cumulative distribution function (CDF) of the probability that the majority category that emerges at the ith pin remains the majority category after all repins have occured. For over 60% of pins, the category of the very first pin determines the majority category; in over 90% of cases, a stable majority has emerged after just 5 repins (all pins have > 5 repins in our dataset).

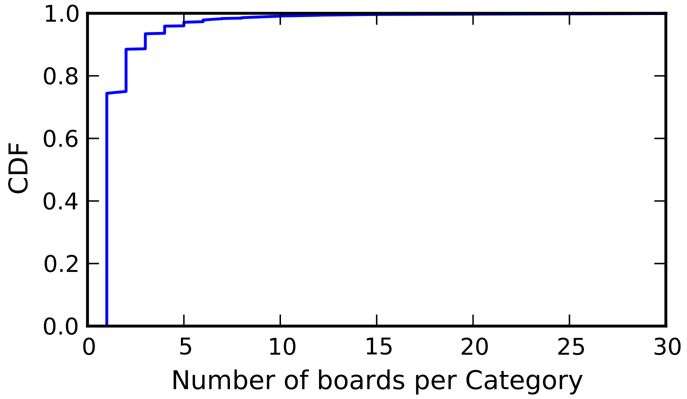
Figure 6.5
Category Concentration and User Specialisation



CDF of the fraction of users' pins in their top-k categories shows that each user specialises in a handful of categories.

matically. Again we focus on categories. We first look at the distribution of categories per user and find that users are highly specialised: Figure 6.5 considers the fraction of a user's pins which are in the top-*k* categories of the user. This shows, for example, that about half the users have nearly half their pins in pinboards belonging to their the top category, and 80% users have *all* their pins in pinboards belonging to their top-5 categories. This indicates a high degree of specialisation.

Figure 6.6
Users' Category Choices Can Predict Pinboard Choices



CDF of per-user number of boards per category shows that users tend not to have many boards in each category, implying that knowing the category assigned by a user, one can predict users' choice of pinboard for a pin.

We next consider how users choose to create pinboards. Figure 6.6 shows that most users have one or two pinboards in each category. Thus, it appears that users are mostly following the coarse-grained taxonomy developed by Pinterest, and are in fact simply categorising images in this space, rather than on highly personalised pinboards.

6.1.3 Implication: Board prediction is easy

The results of §6.1.1 strongly suggest that most repinners agree on the category to assign to a pin, and furthermore, this majority category can be deduced just by observing the first few (say 5) repins. Secondly, since users' pins in different categories are highly skewed (Figure 6.5), users' own personal favourite categories can be predicted as a choice for the category used. In examining the corpus of repins, we find that, consistent with Figure 6.2, ≈87% of repins after the first five repins are made to the majority category. A further 4.7% of repins are made not to the majority category, but to the category in which the user has most of her pins. Thus, we expect that predicting the category of a particular pin based on these powerful signals can be an easy problem, and exploit these in §6.4.1.

Further, §6.1.2 suggests that users tend to have very few boards per category. Thus, once the category is predicted, we expect to be able to predict the actual pinboard chosen by the user as well. Finally, for the few cases when the user's pins are not in the majority category, we propose to use the fact that users specialise in a few categories to predict the correct category and thereby the board used. We explore the above two strategies in §6.4.2.

We conjecture that the high levels of agreement seen for a pin's category may in fact be a result of the high degree of user specialisation within categories: Since users choose to repin in very few categories, the fact that the user has paid any attention to a pin and repinned it is a strong indicator that the pin belongs to the given category. This may help explain the result of Figure 6.2 that nearly 8 out of 10 repinners agree on the category for a pin.

6.2 Predicting Pinterest: an outline

In the rest of this chapter, we will use the notions of user category specialisation and agreement uncovered from the data, together with a number of user- and image-content related features to develop a model for predicting users' repins. Our ultimate goal, as stated earlier, is to automatically suggest, given a user and an image, whether the user will repin the image, and which pinboard it will be repinned to. In this section, we describe the features used, and our outline model for predicting content curation actions as a cascade of predictors. Later sections will use our **DATA-IMG** dataset to validate the different parts of the model.

6.2.1 Curation as a cascade of predictors

We model the process of curation as a cascade of predictors (Figure 6.1). A content curation action involves a user u who "repins" a pin onto one of her pinboards. The first prediction problem (§6.3) is to take a user u and a pin p, and predict an action $f_1 : (p,u) \rightarrow \{noaction, repin\}$. Next, the *repin* involves a further user-specific decision as to which pinboard the pin should be placed in (§6.4). We may formulate this problem as a classification task where a machine learning classifier f_2 is trained to recognize pinboard b_i to which user u is going to put repinned pin p, i.e., $f_2 : (p,u) \rightarrow \{b_1, b_2, ..., b_n\}$ where $\{b_1, b_2, ..., b_n\}$ is a set of user's u pinboards. However, taking cue from §6.1.1 and §6.1.2, we split this task into two. First, we predict the *category* that the user might implicitly choose for the image (§6.4.1), i.e., we train a classifier $f_{2.1}$ to recognise the category c_i in which user u is going to put the repinned pin p: Formally, we build a model to predict $f_{2.1} : (p,u) \rightarrow \{c_1, c_2, ..., c_n\}$ where $\{c_1, c_2, ..., c_n\}$ is a set of user's u categories. Then in §6.4.2, we train a classifier $f_{2.2}$ to predict the pinboard given the category as selected by the user. Formally, we develop the model $f_{2.2} : (c,u) \rightarrow \{b_1, b_2, ..., b_n\}$. As expected from Figure 6.6, this turns out to be an almost trivial problem.

6.2.2 Features

We tackle the above classification problems by incorporating both image- and user-related features (summarised in Table 6.1). In addition, we also use the consensus agreement (§6.1.1) around the category of the image, as measured from the first five repins as a feature derived from the "crowd". We do not incorporate social network information here, since we found that the information seeking activities in Pinterest rely on interest matching rather than social relationship in §5.

6.2.2.1 User-related features

We employ several user-related features which measure both the preferences as revealed by a user's repin history, and the user's extent of involvement with Pinterest, based on her profile.

User Image Preferences To describe users' preferences for particular types of content, we use two sets of features: *Category Preferences* and *Object Preferences*. The first set is a relatively coarse-grained approach wherein we measure how many images a user repins in each of the 32 different Pinterest categories. Object preferences are obtained by devising user-specific signatures from the visual features of those repins: We use the state-of-the-art approach in object recognition for images [Deng et al., 2009], deep convolutional networks [Krizhevsky et al., 2012], to extract a set of more fine-grained visual features. More specifically, we train a deep convolutional network using Caffe library [Jia et al., 2014] based on 1.3 million images annotated with 1000 ImageNet classes and apply it to classify Pinterest images. Then, for each user in our dataset we measure a vector of 1000 features which represents the centroid of the Image Objects recognised in their previous repins.

User Profile Features We also take into account the extent of the user's activity on Pinterest by measuring different statistics from their profiles:

number of followers, number of boards, number of repins, etc.

6.2.2.2 Image-related features

Image-related features are derived based on the content of the images. A more traditional approach is to use various metrics of the *quality* of the image. This is complemented by drawing a range of features using state-of-the-art object recognition for images.

Image Quality Features Firstly, we define 14 image quality features (*Image Quality-I and -II* in Table 6.1) to describe the content of an image. These include colour model-related *basic visual features* such as lightness, saturation, colorfulness, gray contrast, RMS contrast, naturalness, sharpness, sharp pixel proportion and left-right intensity balance. We also consider three *aesthetics-related features* that have been used previously in literature: simplicity, modified-simplicity and rule of thirds. Our goal is to assess how these image quality features can capture user preferences for a particular type of content.

We note that extraction of the majority of image quality features requires significant computational resources on the scale of 151K images. Therefore, we used a dataset of down-scaled images (i.e., with width equal to 200 pixels) to extract all image quality features, except for lightness, saturation, colorfulness and naturalness for which performance was not an issue. Our experiment on a random subset of 1,000 images showed that the Pearson's correlation coefficients between the features extracted from the original and rescaled images were over 0.90 across all features, suggesting that the error introduced by using the down-scaled images is reasonably small (an average absolute error of 0.01).

Object Recognition Features As described above, we train the Caffe library [Jia et al., 2014] using 1.3 million images annotated with 1000 ImageNet classes [Deng et al., 2009] and apply it to classify Pinterest images. Through this process, we extract two types of visual features: (1) *Deep neural network* features from the layer right before the final classification layer. These are known for a good performance in semantic clustering of images [Donahue et al., 2013]; and (2) *Recognised objects* in the image, from among the 1000 Image Object classes that the model is trained on.

Other pin-related features In addition, we use all user-specific features of the original pinner: Features such as the number of followers are indicative of the status and reputation of the user, and might have a bearing on whether/not an image is repinned. Similarly the taste and expertise of the original pinner may also indirectly influence the repinner, and are captured by the user's image object centroid, and activity levels in terms of number of boards, repins etc.

6.2.2.3 Crowd features

In addition to the above, driven by Figure 6.4, we extract a simple but powerful feature: the majority category as seen after the first five repins. We then use these to predict the user-specific categories for repins beyond the first five (§6.4).

Table 6.1

Prediction Features

	Features	Dim	Description
User features (U or P)			**User Image Preference Features**
	Category Preferences	32	Users preferences towards different Pinterest categories, described by the fraction of images users have (re)pinned into each category since they signed up on Pinterest.
	Object Preferences	1000	User preferences for object classes as recognised by the Deep Neural Network Caffe [Donahue et al., 2013] (see below for details) is computed as the centroid of the Caffe-generated object classes of all images (re)pinned by the user during a week-long training period in our dataset (3–9 Jan 2013).
			User Profile Features
	Pinboard count	1	Represents the number of personal pinboards of a user.
	Secret board count	1	Calculates the number of users' private pinboards, only accessible by the owner.
	Pin count	1	Represents the total number of pinned and repinned images.
	Like count	1	Measures the number of images a user has liked.
	Follower count	1	Accounts for the number of users who follow the user under consideration.
	Following count	1	Represents the number of users who are followed by the current user.
			Image Quality-I: Basic Visual Features
	Lightness	2	Derived directly from HSL color space. The average and standard deviation of the lightness values of all the pixels in the image are derived as lightness features.

	Features	Dim	Description
Image features (I)	Saturation	2	Derived directly from HSL color space. The average and standard deviation of all the pixels are used.
	Colourfulness [Hasler and Suesstrunk, 2003]	1	A measure of a image's difference against gray. It is calculated in RGB as [Hasler and Suesstrunk, 2003]: $\sqrt{\sigma_{rg}^2+\sigma_{yb}^2}+0.3\sqrt{\mu_{rg}^2+\mu_{yb}^2}$, where $rg = R-G$ and $yb = \frac{R+G}{2} - B$.
	Gray contrast [Cheng et al., 2012]	1	It measures relative variation of lightness across the image in HSL colour space. It is defined as the standard deviation of the normalised lightness $\frac{L(x,y)-L_{min}}{L_{max}-L_{min}}$ of all image pixels.
	RMS contrast [San Pedro and Siersdorfer, 2009; Webster and Miyahara, 1997]	1	Defined by the standard deviation of all the pixel intensities relative to the mean image intensity or $\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$.
	Naturalness [Huang et al., 2006]	1	A measure of the degree of correspondence between images and human perception of reality. It is described by grouping the pixels with $20 < L < 80$ and $S > 0.1$ in HSL color space according to their H (hue) values into three sets: Skin, Grass and Sky. The naturalness score NS_i, $i \in \{Skin, Grass, Sky\}$, and the proportion of pixels NP_i are derived from the image. Then the final naturalness score is: $NS = \sum_i NS_i \times NP_i$.
	Sharpness [San Pedro and Siersdorfer, 2009]	1	A measure of the clarity and level of detail of an image. Sharpness can be determined as a function of its Laplacian, normalized by the local average luminance in the surroundings of each pixel, i.e., $\sum_{x,y}\frac{L(x,y)}{\mu_{xy}}$, with $L(x,y) = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2}$, where $\mu_x y$ denotes the average luminance around pixel (x, y).

	Features	Dim	Description
Image features (I)	Sharp pixel proportion [Yeh et al., 2010]	1	Photographs that are out of focus are usually regarded as poor photographs, and blurriness can be considered as one of the most important features for determining the quality of the photographs. The photographs are transformed from spatial domain to frequency domain by a Fast Fourier Transform, and the pixels whose values surpass a threshold are considered as sharp pixels (t = 2). The sharp pixel proportion is the fraction of sharp pixels of total pixels.
	Intensity balance [Yeh et al., 2010]	1	It measures how different the intensity is on the left side of the image compared to the right. Two sets of histograms are produced for the left and right portions of the image. The histograms are later converted into chi-square distributions to evaluate the similarities between them, i.e., $\mid \sqrt{\sum_{i=1}^{k}(E_{left}-E_{right})} \mid$.
	Image Quality-II: Aesthetic Features		
	Simplicity-1 [Luo and Tang, 2008]	1	Simplicity in a photograph is a distinguishing factor in determining whether a photograph is professional or not. For a image, the RGB channels are quantized respectively into 16 different levels and the histogram (H) of 4096 bins are generated for the photographs. The simplicity feature is defined as: $(\parallel S \parallel / 4096) \times 100\%$, where $S = \{i \mid H(i) \geq \gamma h_{max}\}$ and $\gamma = 0.01$.
	Simplicity-2 [Yeh et al., 2010]	1	A modified version of Simplicity-1. Instead of evaluating the simplicity of the whole image, Simplicity-2 extracts the subject region of a photograph and what remains is the background region. The colour distribution of the background is used to evaluate simplicity as above.

	Features	Dim	Description
Image features (I)	Rule of Thirds [Yeh et al., 2010]	1	This is a well-known photograph composition guideline. The idea is to place main subjects at roughly one-third of the horizontal or vertical dimension of the photograph. It is measured by how close the main subjects are placed near these “power points”.
	Object Recognition Features		
	Deep Neural Network (DNN) [Donahue et al., 2013; Krizhevsky et al., 2012]	4096	The deep convolutional neural network from the ImageNet [Deng et al., 2009; Krizhevsky et al., 2012] image classification challenge. We use Caffe [Jia et al., 2014], an open-source implementation of deep convolutional networks to train an eight-layer convolutional network on 1.3 million images annotated with 1000 ImageNet classes. Then we extract 4096 features from the layer right before the final (following [Donahue et al., 2013]).
	Recognised Objects [Krizhevsky et al., 2012]	1000	We use the deep convolutional network described above to recognise object classes in Pinterest images and use them as 1000 Image Object features.

List of features used in the cascade of classifiers used to predict user curation actions on Pinterest. The dimension (Dim) column gives the number of scalar values in a feature. User-specific features are used both to describe the user who is repinning the image (U) as well as the original pinner (P) who introduced the image on Pinterest. We also use the majority category as computed by the crowd of (first 5) users pinning an image as a feature in §6.4. Image features (I) are based both on indicators of image quality, as well as object recognition using a Deep Neural Network. User preferences among the recognised object classes is also captured as the user-specific feature “Object Preferences”.

6.3 Predicting User Attention

Here we develop the first step of the cascade pipeline described in §6.2.1. We analyse the features which drive user attention towards a given pinned image and predict whether the user will take an action on it or not. Specifically, we consider two classes of signals: those of the pinned image and the user. The user u is described by the set of features U in Table 6.1, which depend on her category and object preferences as well as statistics of her user profile on Pinterest. The pin p is described by the set of image features I in Table 6.1, which may be attributed to the content of the pinned image. The image features are augmented by the user features of the pinner who published the image, as various characteristics of the original pinner, such as her "taste" of images, and how influential a user she is in Pinterest, may affect its repinnability. We formalise the problem of predicting user attention as a classification problem where we train a binary classifier $f_1 : (p,u) \to \{repin, noaction\}$ to predict a binary output $\{repin, noaction\}$ for a given input (p,u). For the purpose of this analysis we have chosen a Random Decision Forest classifier[3] known for a good prediction performance in image classification problems [Bosch et al., 2007].

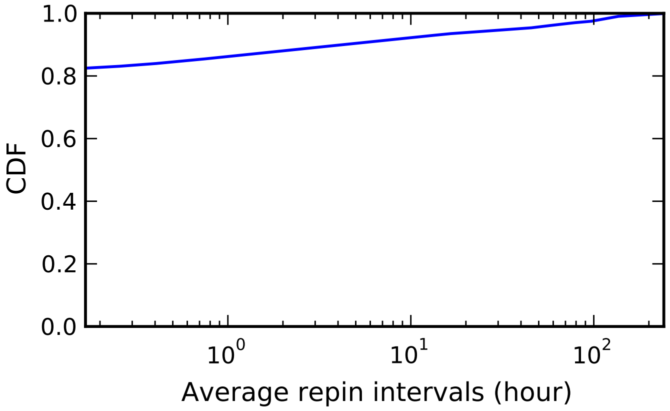
6.3.1 Generating negative samples

One of the challenges in training a model for the system with the absence of explicit negative feedback (as there is no "dislike" button in Pinterest) is to generate realistic negative training samples. The fact that a pin was not repinned by a user does not necessarily mean they would not have liked to repin it. It might have been the case that a repin did not happen simply because the user didn't have a chance to see the pin, and had she seen the pin, she might have taken an action on it. To account for this variance when generating negative samples, we assume that the pins which are published just before the time a user is taking an action are more likely to be noticed by the user. Thus, for a user u who took actions at times $\{t_1, t_2, \ldots, t_n\}$, we randomly select n negative samples[4] among pins

[3] We used the Random Forest implementation from the *SKLearn* package with $\sqrt{n_{features}}$ split and 500 estimators (other values from 10 to 1000 were also tested, but 500 showed the best tradeoff between speed and prediction performance).

[4] This means there are the same number of the positive and negative samples in our training set. We also evaluate our model in imbalanced cases (with positive/negative =

Figure 6.7
Repins Are Concentrated in Time



CDF of average time intervals between repins shows that successive repins for a pin tend to happen quickly.

that were published in the time intervals of *one hour before the time of the actions*[5] and which were not repinned by the user. Note, that this approach is justified by the fact that over 80% of repins happen in interval of one hour since previous repin of the same image (Figure 6.7), suggesting that pins which have not been curated in an hour long interval are likely to be replaced on the home page (or category boards) by more recent activities and therefore would be less likely noticed by a user.

6.3.2 Validation

To assess performance of the proposed model we split the dataset into three consecutive time intervals: We use all pins from the first interval to learn user's preferences; all repin activity from the second one to train the model and all repins from the third one to test the model. Further, we consider two different experimental settings: when only category preferences of a user are taken into account and when both category preferences and visual object preferences are considered together. This enables us to assess the extent to which Pinterest categories can capture specialisation of individual users. The results of the experiments are summarized in Table 6.2, and feature importances in Table 6.3.

10/90)and observe similar performance (accuracy and AUC).

[5] We tried time windows of other sizes, ranging up to six hours before the time of repin. The precise time window size does not appear to affect prediction performance.

Table 6.2
Performance of User Attention Prediction

	Without Obj. Prefs.	With Obj. Prefs.
Accuracy	0.66	0.77
Precision	0.70	0.83
Recall	0.64	0.69
F1-Score	0.66	0.75

Given an image, the task is to predict whether the user will pay it any attention, i.e., whether the user will repin it or not. Two different settings are considered: when only user preferences for categories are known, and when user preferences among the 1000 object classes recognisable by DeCAF [Donahue et al., 2013] are also taken into account.

Table 6.3
The Feature Importance in User Attention Prediction

Feature Type	**Without Obj. Prefs.**	**With Obj. Prefs.**
Object Preferences (U)	–	0.40
DNN (I)	0.37	0.32
Recognised Object (I)	0.21	0.26
Category Prefs (U)	0.32	0.005
Category Prefs (P)	0.005	0.005
Profile Features (U)	0.09	0.001
Profile Features (P)	0.001	0.001
Image Quality-I &-II (I)	0.001	0.001

The relative importance of different feature types is measured as expected fraction of the samples that a feature contributes to, in the Random Decision Forests constructed for the two scenarios of Table 6.2. Feature classes correspond to Table 6.1, and are ordered in descending order of importance when object preferences are used.

Firstly, we note that, consistent with §6.1.2, the prediction performance is high (Accuracy of 0.66 and Precision of 0.70) even when only category preferences of individual users are considered. From Table 6.3 we also note that Category Preferences of users along with the image-related DNN and Recognised Objects features are the most important to predict user attention in this scenario. However, when we add User Object Preferences, the performance of the prediction algorithm improves by 7-18% across all considered metrics (Accuracy, Precision, Recall, F1-Score) and, similarly, Category Preferences features are replaced by the Object Preferences in the feature importance rank. This suggests that the fine-grained object-related preferences are much more effective at capturing user preferences than the coarse-grained category preferences.

6.4 Category and Board Prediction

In this section we elaborate our model by introducing the pinboard classifier which aims to predict a user's choice of a board for a repined image. We recall that a Pinterest user may have several different pinboards each assigned to one of 32 globally defined categories. Each of the images in the datasets are thus implicitly labeled by users with one of the 32 categories. In this section, we first develop a model to predict which category a user will choose, given that she will repin the image. We then refine this and predict which pinboard is used, if the category chosen by the user is known. The latter problem is trivial, as users tend to have very few pinboards per category (Figure 6.6). The former problem is aided enormously by deriving a straightforward category indication from the actions of the crowd (§6.1.1).

6.4.1 Category prediction

We design a multi-class Random Forest classifier to learn which category a user will repin a given image into. Specifically, we consider three classes of signals:

User: Because of user specialisation, we expect that most of the repin activities of the user is restricted to only a few categories, and furthermore, even amongst these categories, there may be a skewed interest favouring certain categories over others. Thus, given that the user has repinned an image, we can expect her underlying skew of interest amongst different categories to be a reasonable guess for the category chosen for the repinned image. Thus the user category preference in Table 6.1 can be interpreted as the empirical probabilities $p_u(c_1)$, $p_u(c_2)$, ..., $p_u(c_{32})$ that a user u will repin an image into categories $c_1, c_2, ..., c_{32}$.

Image: The decision of the repinner on which category to assign can be modulated by the features of the image, the objects in the image, and how closely the objects in the image match the interests of the user.

Table 6.4
Performance of Repin Category Prediction

Features	**ACC**
User	0.42
Image+User	0.77(*)
Crowd+Image+User	0.88(*)
Crowd+User	0.85
Random	0.19

Given a user and an image repinned by her, the task is to predict which category she will repin the image into. The table shows performance in terms of Accuracy (ACC). Stars (*) indicate χ^2 feature selection is applied to select 200 most relevant features for the classification.

We therefore include all the Image features (I) (c.f. Table 6.1) as well as the matching object preferences of the user in this class of signals.

Crowd: There is also a good agreement over category assignment amongst different users (c.f. §6.1.1). Thus, beyond any preferences that the user may have, the consensus or crowd vote on the category, e.g., as seen after the first five repins for the image, is a heuristic for the category that might be assigned by a given repinner.

As before, to evaluate performance of the proposed model, we split the dataset into three consecutive time intervals: We use all pins from the first interval to learn users' object preferences (this is common to §6.3 and is not repeated). We train the model based on activities in the second interval, and all repins from the last one are used to test the model.

The results of the experiments are summarised in Table 6.4. Firstly, we note that the prediction performance is quite high even with only using users' skew in their category preferences (Accuracy=0.42, compared with the baseline accuracy=0.19 obtained by randomly selecting a category among user's categories). When we add deep learning-based image recognition and image quality features to modulate user preferences amongst different categories, we see a dramatic improvement in accuracy to 0.77. Further adding information about the crowd-indicated category gives us an extremely accurate model with an accuracy of 0.88.

Given that the image features we consider are based on a state-of-the-art deep learning library, it is interesting to compare the performance of image-related features with a similar signal derived from the crowd. Table 6.4 shows that even by just using the user preferences among categories together with crowd-derived category information, we can obtain an accuracy of 0.85 (compared with 0.77 for Image+User features), suggesting that crowdsourced image categorisation is more powerful than current image recognition and classification technology.

6.4.2 Pinboard prediction

Next, we look at the way users select pinboards under each category. From Figure 6.6 we observe that the vast majority of users (75%) has only one board under each category, suggesting that in the most of the cases the problem of choosing a pinboard for a repinned image is similar to that of choosing a category. Nevertheless, some users may have more than one pinboard per category. To account for this, we compute the empirical probabilities of a user choosing a given category for an image, and combine it with the empirical probability of putting any given image into that pinboard (computed as the fraction of the user's images in the given pinboard).

This gives us a prediction score for each pinboard, allowing us to compute a ranked list of pinboards. We evaluate prediction power of our method by calculating accuracy at a given cut-off K of the top predicted categories (Table 6.5). Formally, we define $Accuracy@K$ as a fraction of the experiments in which the ground truth pinboard was successfully predicted among the $top@K$ of the prediction list.

Comparing the ($Accuracy@1$) results of the Random benchmark between Table 6.5 and Table 6.4, we note that pinboard prediction is just a slightly more difficult problem than that of predicting categories. The accuracy results of board prediction with the user preference features alone, and with all features, reflect those of the category prediction with an average decrease of 10% in performance. We also note that prediction performance for the Top-5 pinboards goes over a mark of 94%, an observation which can be useful in the design of board recommendation applications.

6.5 End-to-End Performance

To test the end-to-end performance of the proposed methods, we devise a cascaded-predictor which sequentially combines individual classifiers

Table 6.5
Performance of Pinboard Prediction

A@k	Crowd+Image+User	User	Random
1	0.73	0.35	0.15
2	0.84	0.52	0.27
3	0.89	0.63	0.37
4	0.92	0.70	0.46
5	0.94	0.76	0.53

The performance is assessed by the Accuracy@k metric, which we defined as a fraction of the experiments in which the ground truth pinboard was successfully predicted among the top@k of the prediction list.

Table 6.6
End-to-end Performance for the Cascade of Predictors

	@1	@2	@3	@4	@5
Accuracy	0.69	0.71	0.73	0.74	0.75
Precision	0.60	0.70	0.72	0.76	0.77
Recall	0.50	0.58	0.62	0.63	0.64

introduced in the previous sections, i.e., the separately trained User Attention and Pinboard classifiers. We estimate the overall prediction performance of the system by calculating the accuracy, precision and recall of the proposed cascade predictor. These metrics are calculated as an outcome of an imaginary multi-class classifier $f : (u,p) \rightarrow \{noaction, b_1, b_2, ..., b_n\}$ where $b_1, b_2, ..., b_n$ denote users' pinboards. We also measure *Accuracy@K*, *Precison@K* and *Recall@K* at different cut-offs K of the *top@K* pinboards predictions. We note that the testing set for these experiments is sampled such that the fraction of non-action and repin cases is set to 1:1, assuring that the number of positive and negative cases in attention prediction experiments are equal.

The results of the experiments presented in Table 6.6 suggest that the end-to-end performance remains on a high level of $Accuracy@1 = 0.69$ for the $Top@1$ pinboard prediction and further increases to $Accuracy@5 = 0.75$ for the $Top@5$ users' pinboards. Since we need to predict among multiple users' boards, we define precision and recall by distinguishing between correct or incorrect classification of a user's board (defined as true/false positives) and correct or incorrect prediction of no action (defined as true/false negatives). From Table 6.6, we report that the end-to-end precision remains on a level of 0.60, and reaches 0.77 for predicting among the *Top*@5 users' pinboards, suggesting an overall high level of predictability of individual curation actions on Pinterest.

6.6 Discussion and Conclusions

Social bookmarking and curation is becoming increasingly important to the Web as a whole: Pinterest for instance has become an important source of referral traffic, second only to Facebook amongst all the major social networks [Rose, 2014]. Furthermore, Pinterest referral traffic is valuable for e-commerce sites, being 10% more likely to result in sales, with each sale being on average $80, double the value of sales from Facebook referrals [Hayes, 2012]. Therefore, understanding Pinterest can result both in fundamental insights into the nature of content curation, as well as commercially valuable applications such as recommending items that users are willing to buy, and optimising marketing campaigns for brand awareness and recall. Understanding what makes users curate an image could also help other applications such as improving the relevance of image search results.

This chapter takes first steps towards this research agenda by showing that although Pinterest users are curating highly personalised collections of content, they are effectively participating in a crowdsourced categorisation of images from across the web, by their choices of pinboards. By exploiting the fact that user pinboards can have an associated category, we reinterpret the act of pinning as a distributed computation process that categorises images from across the Web into the 32 categories recognised on Pinterest. When viewed through this perspective, it becomes readily apparent that there is overwhelming agreement among users on how to categorise images. Additionally, we see that users tend to specialise in a handful of categories, and tend not to have several boards in the same category. Furthermore, even within their favourite categories, their attention is skewed towards the top 1-5 categories.

Based on these observations, we developed a cascade of predictors, that, given a pin and a user, is able to predict whether the user would repin it, and if so, to which of her pinboards. The three layers of the cascade could be conceived as providing a possible mechanistic understanding of content curation on Pinterest. Although one could possibly conceive of alternate mechanistic views, the behaviour and performance of the predictors we built serve to illuminate some of the various factors involved in content curation: As can be expected, the first decision of whether the user repins the pin at all, depends to a large extent on the visual features of the image. In particular, object features extracted using a state-of-the-art deep neural network yielded up to 18% improvement across all considered metrics (Accuracy, Precision, Recall, F1-score) highlighting that object recognition may play a central role in understanding what a user is interested in. The next layer in the cascade, predicting what category a user is likely to assign to a pin, is dominated by one factor: that most users agree on the category. Indeed, by looking at the first five repins, we are able to predict other repins with $\approx$ 85% accuracy, and although several other types of features, ranging from visual features of the image to features of the user were examined, none were able to improve over this single feature. The final layer of predicting the board given the category turned out to be an almost trivial problem, suggesting that users, rather than showing complicated behaviours, appear to be "operating" just in the 32-dimension approximation of Pinterest global categories.

Because of the collective efforts of large numbers of Pinterest users, we have amassed an extensive annotated set of images (over 1.2 million category annotations for 214K pins). Although there is a great deal of agreement in general, individual users may have slightly different views about categorisation, and similarly, the categories may not be mutually exclusive (e.g., one user might classify an image of the Eiffel tower into an "Art" pinboard while another might choose "Architecture", both of which are standard Pinterest categories). We find that by incorporating user preferences, we are able to further improve the performance, evidence that users are "personalising" these categories by reinterpreting and reaggregating them through pinboards. Thus, this arrangement of allowing users the freedom to create pinboards suiting their own purposes, whilst at the same time associating the pinboards to a small number of standard categories appears to strike a good balance between the rigid taxonomies of an ontology and the free-for-all nature of so-called folksonomies [Mathes, 2004], enabling a meaningful global categorisation (with enough power to predict image categories based on their visual features), whilst at the same time allowing user flexibility.

Chapter 7

Reflections and outlook

This dissertation has been looking on a buzz topic in online content marketing, social curation. Rather than creating new content, social curation allows its users to categorise and organise online content they find online, and thus created their personal Web taxonomies. Despite the large number of discussions about the idea, there has been a lack of datasets of appropriate scale to allow an extensive validation of theories discussed. At the same time, some basic questions of social curation have not been answered yet.

In this dissertation, we have collected a large social curation dataset with millions of users and images from Pinterest. The dataset covers interactions between curators and three important factors: content, friends and crowds. Using th dataset, we implemented some large-scale empirical studies of social curation as well as machine learning based automation in previous chapters. In this chapter, we conclude the dissertation by summarising our contributions and describing some potential directions for future work.

7.1 Summary of Contributions

In the preceding chapters, we have presented and discussed the results of our research, in relation to the four research questions we proposed in

§1. These results, and the contributions they represent to build on the existing body of work outlined in Chapter 2, are summarised below.

First, **Chapter 4** used a quantitative analysis of several weeks of curation actions on Pinterest to characterise the phenomenon of social curation. We showed that curation tends to focus on items that may not rank highly in popularity and search rankings. This is consistent with the theory that "curation comes up when search stops working" [Shirky, 2010]. Then, we discovered that collectively, the user base of Pinterest focused most of its curation actions on a small number of items, resulting in an extremely skewed distribution of curation activity. We then examined how people curate, and proposed a distinction between structured curation, which highlights an item and organises it and unstructured curation, which simply highlights an item by liking or loving it. We found that the former is more prevalent for popularly curated items. Likes, however, are initially accumulated at a faster pace. Finally, we studied what will affect the social value of curators. Our data pointed to at least three factors: consistent and regular curation actions, diversity of interests, and a preference for structured curation.

This drove us to explore the social networks in social curation services. In **Chapter 5**, we examined social bootstrapping, that allows users to copy friends from established social networks to third-party websites, and the role of social networks. We found that social users are more active and are more likely to return soon. This indicates a bonding effect enabled by social networks. The next question we asked is which type of social networks are more useful for user interaction, the network copied by social bootstrapping or the one created natively. Thus we compare the social structure and interaction of both networks. We found that copied network shows an more important role in promoting social interaction, as it initiates a strong and dense social structure comparing with native networks.

However, social networks also have limitations. We found that social networks are not critical for information seeking in social curation services. Because a non-trivial number of users' content are curated from strangers. To explain the observation, we explored the motivation of a curation action, "repin", in Pinterest and found that it results from interest matching. But this does not mean that users are less social, as we found users tend to have more and more social interaction, as they become more active and influential. We think this is due to the bonding effects of social networks. But when we divide users' friends into copied and native friends, we found they tend to wean from copied friends to have more interactions with interest-based native friends.

Knowing that most curation actions are for personal reasons and based on interest matching, we explored the interaction between curators and crowds. More specifically, in **Chapter 6**, we asked to what extent we could reproduce the content curation. We first visited the notion of agreement in the context of Pinterest. Unlike traditional social bookmarking, pinning on Pinterest does not involve creating an explicit vocabulary of tags to describe the image. However, each pinboard may be associated to one of 32 categories defined globally for all users by Pinterest. Thus, even though each pinboard may be exclusive to a user, the act of pinning *implicitly* categorises the image. In other words, interest users can be seen as performing a massive *distributed computation process*, categorising images found on the Web onto an (extremely coarse-grained) global taxonomy of 32 categories. This lower dimension approximation of Pinterest enable use to build classifiers that, given that the user has curated an image, can predict the pinboard onto which the image would be repinned, and automatically suggest these to the user.

We augmented this by showing that the *content* of the image can be used to predict whether the user would be interested in repinning the pin. To build this predictor, we derived several thousands of image-content features (Table 6.1), ranging from basic visual and aesthetic features to features extracted from the layer right before the final classification layer of the state-of-the-art deep convolutional network in Caffe [Jia et al., 2014], and by recognising objects in the image, using the same convolutional neural network. Using these features, we constructed a supervised machine learning model that is able to assign an image to the majority category. We also learned user preferences for these image features, and predicted whether the image would be repinned by the user.

We composed these classifiers in a pipeline of three layers. The first layer predicted whether the user will pay attention to a pin; the second predicted the category that the user will choose for the pin; and the third predicted the pinboard chosen given the category. Together this pipeline or cascade of classifiers was able to predict curation actions on Pinterest with an accuracy of 69% (accuracy@5=75%).

7.2 Future Directions

Our research on social curation provided a first look to this new trend in World Wide Web. But there are still a lot of potential studies can be implemented in the area. Here, we list some of them:

Interaction between social curation and content provider . As we have discussed, rather than creating new content, curators discover and select content from content providers. Thus, an important question that we need to explore is the interaction between social curation platforms and content provider websites and the effect of such curation activity for providers. On the one hand, social curation platforms makes it easier for users to find interesting content, and draw a lot of attention. This may affect the traffic of the content provider. On the other hand, social curation platforms can also provide referral traffic for those content provider. Pinterest for instance has become an important source of referral traffic, second only to Facebook amongst all the major social networks [Rose, 2014]. Furthermore, Pinterest referral traffic is valuable for e-commerce sites, being 10% more likely to result in sales, with each sale being on average $80, double the value of sales from Facebook referrals [Hayes, 2012]. Therefore, it would be interesting to study whether providers can be benefits from the social curation systems.

Applications of social curation. Social curation is thought to be a solution for information overload [Shirky, 2008; Grineva and Grinev, 2012; Anderson, 2015], but we are not yet clear how we could achieve that. Some straightforward ways are, similar with what we have done in §6, to simulate a social curation system and then do recommendation based on prediction results. This would be extremely useful for multimeida content, because even the-state-of-art computer vision algorithms are still not comparable with human recognition yet, as suggested by our category prediction result Table 6.4. Thus, crowd-based curation method can potentially used to improve the performance of exiting recommendation and search algorithms for images and videos. But more attention is required to find more advanced methods to utilise social curation data.

At the same time, social curation is also an important opportunity for fields like computer vision, as it provides a large number of human labelled multimedia items. For example, in Pinterest, there are millions of curated images with high quality of labels. Those images can be used to train new computer vision algorithms. For example, Yang et al. [2015] start to explore how to profile user preference using Pinterest data.

Social bootstrapping. We have proposed the definition of social bootstrapping, and done a empirical study for it. Followed by our research in §5, Miller et al. [2015] have implemented an user study for social bootstrapping, and found that social bootstrapping will affect the impression of users for new website. But more studies for the topic are still required. There are a lot of questions are not yet solved, due to limitation of data or other factors. For example, in §5, we have studied the effect of social bootstrapping for target networks, but we still not clear what are the benefit and limitation of social bootstrapping for *source network*. Also, the process of bootstrapping from multiple source networks is another unsolved problem need more future study.

Bibliography

S. W. Anderson. *Content Curation*. How to Avoid Information Overload. Corwin Press, 2015. Cited on pages 3 and 111.

D. E. Baird and M. Fisher. Neomillennial user experience design strategies: Utilizing social networking media to support “always on” learning styles. *Journal of educational technology systems*, 34:5–32, 2005. Cited on pages 7, 26, and 50.

S. Bakhshi and E. Gilbert. Red, purple and pink: The colors of diffusion on pinterest. *PLoS ONE*, 10(2):1–20, 2015. Cited on page 24.

N. Beagrie. Digital Curation for Science, Digital Libraries, and Individuals. *International Journal of Digital Curation*, 1(1):3–16, 2008. Cited on pages 3, 16, and 17.

R. Bhargava. Manifesto for the content curator: The next big social media job of the future ?, 2009a. Online blog, available from http://www.rohitbhargava.com/2009/09/manifesto-for-the-content-curator-the-next-big-social-media-job-of-the-future-.html, last accessed 14 May 2016. Cited on pages 4, 6, and 46.

R. Bhargava. How to use curation to make your blog better: Lessons from postsecret, 2009b. Online blog, http://www.rohitbhargava.com/dev/2009/01/how-to-use-cura.html, last accessed 23 April 2016. Cited on page 18.

M. Boguñá and M. A. Serrano. Generalized percolation in random directed networks. *Physical Review E*, 72(1):016106, 2005. Cited on page 69.

A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *International Conference on Computer Vision*, ICCV, pages 1–8. IEEE, 2007. Cited on page 99.

S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464 (7291):1025–1028, 2010. Cited on page 27.

Bureau of Labor Statistics. Archivists, curators, and museum workers. *Occupational Outlook Handbook (2016 - 17 Edition), U.S. Department of Labor*, 2015. Available at http://www.bls.gov/ooh/education-training-and-library/curators-museum-technicians-and-conservators.htm, last accessed 23 April 2016. Cited on page 16.

M. Burke, C. Marlow, and T. Lento. Feed me: motivating newcomer contribution in social network sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 945–954, New York, NY, USA, 2009. ACM. Cited on pages 25, 27, 55, 63, and 81.

M. Burke, C. Marlow, and T. Lento. Social network activity and social well-being. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1909–1912, New York, NY, USA, 2010. ACM. Cited on page 25.

M. Burke, R. Kraut, and C. Marlow. Social capital on facebook: Differentiating uses and users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 571–580, New York, NY, USA, 2011. ACM. Cited on page 25.

D. Carr. A code of conduct for content aggregators. The New York Times, 2012. Available from http://www.nytimes.com/2012/03/12/business/media/guidelines-proposed-for-content-aggregation.-online.html, last accessed 23 March 2013. Cited on page 22.

D. Chaffey. Global social media research summary 2016, 2016. Online blog, available from http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/, last accessed 14 May 2016. Cited on page 31.

E. A. Chambers. Defining the role of the curator. *Museum Studies: Perspectives and Innovations*, 2006. Cited on page 16.

S. Chang, V. Kumar, E. Gilbert, and L. G. Terveen. Specialization, homophily, and gender in a social curation site: findings from pinterest. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 674–686, New York, NY, USA, 2014. ACM. Cited on pages 27 and 28.

H. Cheng, R. v. Zwol, J. Azimi, E. Manavoglu, R. Zhang, Y. Zhou, and V. Navalpakkam. Multimedia features for click prediction of new ads in display advertising. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 777–785, New York, NY, USA, 2012. ACM. Cited on page 96.

C. M. Cheung, P.-Y. Chiu, and M. K. Lee. Online social networks: why do students use facebook? *Computers in Human Behavior*, 27(4):1337–1343, 2011. Cited on page 25.

D. Choi and J. Kim. Why people continue to play online games: In search of critical design factors to increase customer loyalty to online contents. *CyberPsychology & behavior*, pages 11–24, 2004. Cited on pages 7, 26, and 51.

J. S. Coleman. Social capital in the creation of human capital. *American Journal of Sociology*, 94:S95–S120, 1988. Cited on page 63.

G. Conole and J. Culver. The design of cloudworks: Applying social networking practice to foster the exchange of learning and teaching ideas and designs. *Computers & Education*, 54:679–692, 2010. Cited on pages 7 and 50.

Delicious.com. The introduction page of delicious, 2016. http://delicious.com/about, last accessed 23 April 2016. Cited on pages 3 and 22.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR'09, pages 248–255, 2009. Cited on pages 92, 93, and 98.

J. DiMicco, D. R. Millen, W. Geyer, C. Dugan, B. Brownholtz, and M. Muller. Motivations for social networking at work. In *Proceedings of the 2008 ACM Conference on Computer Supported & Cooperative Work*, CSCW '08, pages 711–720, New York, NY, USA, 2008. ACM. Cited on pages 7, 27, and 50.

J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. *arXiv:1310.1531 [cs]*, 2013. Cited on pages 93, 95, 98, and 101.

R. Dube. Here are the fastest growing social networks you need to join, 2015. Online blog, available from http://www.makeuseof.com/tag/7-fastest-growing-social-networks-according-google-trends/, last accessed 14 May 2016. Cited on page 31.

N. Ducheneaut and R. J. Moore. The social side of gaming: A study of interaction patterns in a massively multiplayer online game. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, CSCW '04, pages 360–369, New York, NY, USA, 2004. ACM. Cited on pages 7, 26, and 51.

K. Duh, T. Hirao, A. Kimura, K. Ishiguro, T. Iwata, and C. M. A. Yeung. Creating stories: Social curation of twitter messages. In *Sixth International AAAI Conference on Weblogs and Social Media*, ICWSM. AAAI, 2012. Cited on page 23.

N. Ellison, C. Steinfield, and C. Lampe. Spatially bounded online social networks and social capital. *International Communication Association*, 36(1–37), 2006. Cited on page 25.

N. B. Ellison, C. Steinfield, and C. Lampe. The benefits of facebook "friends:" social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, 2007. Cited on pages 25, 55, 63, and 81.

G. Eysenbach. Medicine 2.0: Social networking, collaboration, participation, apomediation, and openness. *Journal of Medical Internet Research*, 10, 2008. Cited on pages 7, 26, and 51.

Facebook.com. The company information of facebook, 2016a. The stats, http://newsroom.fb.com/company-info/, last accessed on 10 May 2016. Cited on page 53.

Facebook.com. Facebook developers' webpage on open graph, 2016b. Available at https://developers.facebook.com/docs/sharing/opengraph, last accessed on 10 May 2016. Cited on page 8.

E. Fehr and S. Gächter. Fairness and Retaliation: The Economics of Reciprocity. *The Journal of Economic Perspectives*, 14(3):159–181, 2000. Cited on page 69.

A. García-Silva, J.-H. Kang, K. Lerman, and O. Corcho. Characterising emergent semantics in twitter lists. In *Proceedings of the 9th international conference on The Semantic Web: research and applications*, ESWC'12, pages 530–544, Berlin, Heidelberg, 2012. Springer-Verlag. Cited on page 24.

B. Gelley and A. John. Do i need to follow you?: Examining the utility of the pinterest follow mechanism. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1751–1762, New York, NY, USA, 2015. ACM. Cited on pages 7, 26, and 50.

E. Gilbert, S. Bakhshi, S. Chang, and L. Terveen. "I need to try this?:" a statistical overview of pinterest. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2427–2436, New York, NY, USA, 2013. ACM. Cited on pages 4, 24, 27, and 28.

R. Givhan. Citizen curators' two cents: Worth every penny. *The Washington Post*, 2007. Available at http://www.washingtonpost.com/wp-dyn/content/article/2007/12/21/AR2007122100717_pf.html, last accessed 23 April 2016. Cited on page 19.

S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of information science*, 32(2):198–208, 2006. Cited on pages 9 and 87.

S. Gómez, A. D. Guilera, G. G. nes, C. J. P. Vicente, Y. Moreno, and A. Arenas. Diffusion Dynamics on Multiplex Networks. *Physical Review Letters*, 110(2):028701, 2013. Cited on page 27.

J. Gómez-Gardeñes, I. Reinares, A. Arenas, and L. M. M. Floría. Evolution of cooperation in multiplex networks. *Scientific reports*, 2, 2012. Cited on page 27.

D. Greene, F. Reid, G. Sheridan, and P. Cunningham. Supporting the curation of twitter user lists. *arXiv preprint arXiv:1110.1349*, 2011. Cited on page 24.

D. Greene, D. O'Callaghan, and P. Cunningham. Identifying topical twitter communities via user list aggregation. *arXiv:1207.0017*, 2012. Cited on page 24.

M. Grineva and M. Grinev. Information Overload in Social Media Streams and The Approaches To Solve It. In *Proceedings of the 21st international conference on World Wide Web*, WWW, 2012. Web Science Track. Cited on pages 3 and 111.

F. Guerrini. Newsroom curators and independent storytellers: Content curation as a new form of journalism. *Reuters Institute Fellowship Paper*, pages 1–62, 2013. Cited on pages 16, 18, 19, 20, 21, and 22.

C. Hall and M. Zarro. Social curation on the website pinterest.com. *proceedings of the American Society for Information Science and Technology*, 49(1):1–9, 2012. Cited on pages 23 and 27.

J. Han, D. Choi, B.-G. Chun, T. Kwon, H.-c. Kim, and Y. Choi. Collecting, organizing, and sharing pins in pinterest: Interest-driven or social-driven? In *The 2014 ACM International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '14, pages 15–27, New York, NY, USA, 2014. ACM. Cited on page 28.

D. Hasler and S. E. Suesstrunk. Measuring colorfulness in natural images. In *Electronic Imaging*, volume 5007, pages 87–95, 2003. Cited on page 96.

M. Hayes. How pinterest drives ecommerce sales, 2012. Available from http://www.shopify.com/blog/6058268-how-pinterest-drives-ecommerce-sales, last accessed 16 June 2016. Cited on pages 106 and 111.

G. Heiberger and R. Harper. Have you facebooked astin lately? using technology to increase student involvement. *New Directions for Student Services*, 2008(124):19–35, 2008. Cited on pages 7 and 50.

D. Hendricks. Are interest-based networks the way of the future?, 2014. Available at http://www.forbes.com/sites/drewhendricks/2014/

10/16/are-interest-based-networks-the-way-of-the-future/, last accessed 8 May 2016. Cited on pages 7 and 50.

K.-Q. Huang, Q. Wang, and Z.-Y. Wu. Natural color image enhancement and evaluation algorithm based on human visual system. *Computer Vision and Image Understanding*, 103(1):52–63, 2006. Cited on page 96.

K. Ishiguro, A. Kimura, and K. Takeuchi. Towards automatic image understanding and mining via social curation. In *2012 IEEE 12th International Conference on Data Mining*, ICDM, pages 906–911. IEEE, 2012. Cited on page 24.

J. Jamison. Beyond facebook: The rise of interest-based social networks. *TechCrunch*, 2012. Avaival at http://techcrunch.com/2012/02/18/beyond-facebook-the-rise-of-interest-based-social-networks/, last accessed 8 May 2016. Cited on pages 7 and 50.

A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, August 2007. ACM. Cited on page 55.

Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. Cited on pages 67, 84, 92, 93, 98, and 110.

K. Y. Kamath, A.-M. Popescu, and J. Caverlee. Board recommendation in pinterest. In *UMAP Workshops*, 2013. Cited on page 28.

D. Kim, Y. Jo, I.-C. Moon, and A. Oh. Analysis of twitter lists as a potential source for discovering latent characteristics of users. In *ACM CHI Workshop on Microblogging*, 2010. Cited on page 24.

S. Knack and P. Keefer. Does social capital have an economic payoff? a cross-country investigation. *The Quarterly journal of economics*, 112(4): 1251–1288, 1997. Cited on pages 25 and 63.

E. D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer Publishing Company, Incorporated, 1st edition, 2009. Cited on page 72.

J. Korr. A 21st century newswire—curating the web with links, 2008. Online blog, http://niemanreports.org/articles/a-21st-century-newswire-curating-the-web-with-links/, last naccessed 23 April 2016. Cited on page 22.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105. Curran Associates, Inc., 2012. Cited on pages 92 and 98.

T. Y. Lee, C. Dugan, W. Geyer, T. Ratchford, J. Rasmussen, N. S. Shami, and S. Lupushor. Experiments on motivational feedback for crowdsourced workers. In *Seventh International AAAI Conference on Weblogs and Social Media*, ICWSM. AAAI, 2013. Cited on pages 7, 26, and 50.

X. Li, L. Guo, and Y. E. Zhao. Tag-based social interest discovery. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 675–684, New York, NY, USA, 2008. ACM. Cited on page 24.

R. Linder, C. Snodgrass, and A. Kerne. Everyday ideation: All of my ideas are on pinterest. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 2411–2420. ACM, 2014. Cited on pages 20, 23, 40, 48, and 67.

S. B. Liu. The rise of curated crisis content. In *Proceedings of the Information Systems for Crisis Response and Management Conference*, ISCRAM 2010, 2010. Cited on page 23.

S. B. Liu. *Grassroots Heritage: a Multi-method Investigation of How Social Media Sustain the Living Heritage of Historic Crises*. PhD thesis, University of Colorado, US, 2011. Available at http://sophiabliu.com/sophiabliu-dissertation.pdf, last accessed 23 Apri 2016. Cited on pages 19 and 22.

Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *Computer Vision – ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part III*, pages 386–399, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. Cited on page 97.

S. A. Macskassy. On the study of social interactions in twitter. In *International AAAI Conference on Weblogs and Social Media*, ICWSM. AAAI, 2012. Cited on pages 25, 27, 63, 69, and 81.

M. Magnani and L. Rossi. Formation of multiple networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, 2013. Cited on page 26.

A. Mathes. Folksonomies-cooperative classification and communication through shared metadata. *Computer Mediated Communication*, 2004. Cited on page 107.

H. J. Miller, S. Chang, and L. G. Terveen. "i love this site!" vs. "it's a little girly": Perceptions of and initial user experience with pinterest. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1728–1740, New York, NY, USA, 2015. ACM. Cited on page 112.

K. Musial and N. Sastry. Social media: Are they underpinned by social or interest-based interactions? In *Proceedings of the Fourth Annual Workshop on Simplifying Complex Networks for Practitioners*, SIMPLEX '12, pages 1–6, New York, NY, USA, 2012. ACM. Cited on pages 7, 26, and 50.

J. Nahapiet and S. Ghoshal. Social capital, intellectual capital, and the organizational advantage. *Academy of management review*, 23(2):242–266, 1998. Cited on pages 25 and 63.

V. Nicosia, G. Bianconi, V. Latora, and M. Barthelemy. Growing multiplex networks. *Physical review letters*, 111(5):058701, 2013. Cited on page 27.

S. Noël and R. Beale. Sharing vocabularies: tag usage in citeulike. In *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction-Volume 2*. British Computer Society, 2008. Cited on page 9.

R. Ottoni, J. P. Pesce, D. B. Las Casas, G. Franciscani Jr, W. Meira Jr, P. Kumaraguru, and V. Almeida. Ladies first: Analyzing gender roles and behaviors in pinterest. In *International AAAI Conference on Weblogs and Social Media*, ICWSM. AAAI, 2013. Cited on pages 4, 24, 27, and 28.

R. Ottoni, D. Las Casas, J. P. Pesce, W. Meira Jr, C. Wilson, A. Mislove, and V. Almeida. Of pins and tweets: Investigating how users behave across image-and text-based social networks. In *International AAAI Conference on Weblogs and Social Media*, ICWSM. AAAI, 2014. Cited on pages 24, 27, and 28.

S. Ovadia. Digital content curation and why it matters to librarians. *Behavioral & Social Sciences Librarian*, 32(1):58–62, 2013. Cited on page 23.

J. Pearsall. *The New Oxford Dictionary of English*. Oxford University Press, 1998. Cited on pages 3 and 16.

Pinterest.com. The statistics of pinterest, 2015. From the official Twitter account of Pinterest: https://twitter.com/Pinterest/status/582960872093556736, last accessed 25 April 2016. Cited on page 3.

B. Podobnik, D. Horvatić, M. Dickison, and H. E. Stanley. Preferential attachment in the interaction between dynamically generated interdependent networks. *EPL (Europhysics Letters)*, 100(5):50004, 2012. Cited on page 27.

A. Portes. Social capital: Its origins and applications in modern sociology. *Annual Review of Sociology*, 24(1):1–24, 1998. Cited on pages 25 and 63.

R. D. Putnam. Bowling alone: America's declining social capital. *Journal of democracy*, 6(1):65–78, 1995. Cited on pages 55, 61, and 63.

E. Rader and R. Wash. Influences on tag choices in del.icio.us. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, pages 239–248, New York, NY, USA, 2008. ACM. Cited on page 9.

K. Rose. Pinterest is sneaking up on twitter, facebook, and google, 2014. New York Magazine, Available from http://nymag.com/daily/intelligencer/2014/05/pinterest-is-sneaking-up-on-twitter-and-facebook.html, last accessed 16 June 2016. Cited on pages 106 and 111.

D. Rotman, K. Procita, D. Hansen, C. Sims Parr, and J. Preece. Supporting content curation communities: The case of the encyclopedia of life. *Journal of the American Society for Information Science and Technology*, 63 (6):1092–1107, 2012. Cited on page 23.

H. M. SalahEldeen and M. L. Nelson. Losing my revolution: How many resources shared on social media have been lost? In *Theory and Practice of Digital Libraries*, TPDL '12, pages 125–137, Berlin, Heidelberg, September 2012. Springer Berlin Heidelberg. Cited on page 20.

J. San Pedro and S. Siersdorfer. Ranking and classifying attractiveness of photos in folksonomies. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 771–780, New York, NY, USA, 2009. ACM. Cited on page 96.

N. Sastry. How to tell head from tail in user-generated content corpora. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, ICWSM. AAAI, 2012. Cited on page 24.

M. Shatzkin. Aggregation and curation: two concepts that explain a lot about digital change, 2009. Online blog, http://www.idealog.com/blog/aggregation-and-curation-two-concepts-that-explain-a-lot-about-digital-change/ last accessed 23 April 2016. Cited on page 19.

C. Shirky. It is not information overload. It is filter failure. In *Web 2.0 Expo*, 2008. Cited on pages 3 and 111.

C. Shirky. Talk about curation. Published as on online interview with Steve Rosenbaum, 2010. Available from http://curationnationvideo.magnify.net/video/Clay-Shirky-6, last accessed 15 Feb 2013. Cited on pages 4, 5, 37, 39, and 109.

C. Steinfield, N. B. Ellison, and C. Lampe. Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of Applied Developmental Psychology*, 29(6):434–445, 2008. Cited on page 25.

G. Swamynathan, C. Wilson, B. Boe, K. Almeroth, and B. Y. Zhao. Do social networks improve e-commerce?: A study on social marketplaces. In *Proceedings of the First Workshop on Online Social Networks*, WOSN '08, pages 1–6, New York, NY, USA, 2008. ACM. Cited on pages 7, 26, and 50.

M. Szell, R. Lambiotte, and S. Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107:13636–13641, 2010. Cited on page 27.

C.-Y. Teng and L. A. Adamic. Longevity in second life. In *International AAAI Conference on Weblogs and Social Media*, ICWSM. AAAI, 2010. Cited on pages 25, 27, and 69.

N. Thurman and A. Walters. Live blogging–digital journalism's pivotal platform? a case study of the production, consumption, and form of live blogs at Guardian.co.uk. *Digital Journalism*, 1(1):82–101, 2013. Cited on page 21.

W. Tsai and S. Ghoshal. Social capital and value creation: The role of intrafirm networks. *Academy of management Journal*, 41(4):464–476, August 1998. Cited on pages 25 and 63.

Twitter.com. Using twitter lists, 2016. Twitter help center, https://support.twitter.com/articles/76460, last accessed 14 May 2016. Cited on page 24.

S. Valenzuela, N. Park, and K. F. Kee. Is there social capital in a social network site?: Facebook use and college students' life satisfaction, trust, and participation. *Journal of Computer-Mediated Communication*, 14(4): 875–901, 2009. Cited on page 25.

M. Villi. 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, 2012. Cited on page 23.

L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 319–326, New York, NY, USA, 2004. ACM. Cited on page 86.

M. Webster and E. Miyahara. Contrast adaptation and the spatial structure of natural images. *Journal of the Optical Society of America. A, Optics, image science, and vision*, 14(9):2355, 1997. Cited on page 96.

B. Wellman, A. Q. Haase, J. Witte, and K. Hampton. Does the internet increase, decrease, or supplement social capital? social networks, participation, and community commitment. *American behavioral scientist*, 45(3):436–455, 2001. Cited on page 25.

M. Woolcock and D. Narayan. Social capital: Implications for development theory, research, and policy. *The world bank research observer*, 15 (2):225–249, August 2000. Cited on pages 25 and 63.

Y. Yamaguchi, T. Amagasa, and H. Kitagawa. Tag-based user topic discovery using twitter lists. In *2011 International Conference on Advances in Social Networks Analysis and Mining*, ASONAM, pages 13–20, 2011. Cited on page 24.

L. Yang, C.-K. Hsieh, and D. Estrin. Beyond classification: Latent user interests profiling from visual contents analysis. *arXiv preprint arXiv:1512.06785*, 2015. Cited on pages 28 and 112.

C.-H. Yeh, Y.-C. Ho, B. A. Barsky, and M. Ouhyoung. Personalized photograph ranking and selection system. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 211–220, New York, NY, USA, 2010. ACM. Cited on pages 97 and 98.

M. Zarro, C. Hall, and A. Forte. Wedding dresses and wanted criminals: Pinterest. com as an infrastructure for repository building. In *International AAAI Conference on Weblogs and Social Media*, ICWSM. AAAI, 2013. Cited on pages 23, 27, and 28.

C. Zhong, S. Shah, K. Sundaravadivelan, and N. Sastry. Sharing the Loves: Understanding the How and Why of Online Content Curation. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, ICWSM, 2013. Cited on pages 20, 39, 40, 41, 42, 43, 46, 47, and 48.

C. Zhong, M. Salehi, S. Shah, M. Cobzarenco, N. Sastry, and M. Cha. Social bootstrapping: How pinterest and last.fm social communities benefit by borrowing links from facebook. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 305–314, New York, NY, USA, 2014. ACM. Cited on page 72.

Y.-X. Zhu, X.-G. Zhang, G.-Q. Sun, M. Tang, T. Zhou, and Z.-K. Zhang. Influence of reciprocal links in social networks. Technical Report 7, 2014. Cited on page 69.

(114 references in total)