ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed
Computing

Explore the world of Bookcrossing

Master's Thesis

Yu Li
yul@student.ethz.ch

Distributed Computing Group
Computer Engineering and Networks Laboratory
ETH Zurich

Supervisors:
Raphael Eidenbenz
Prof. Dr. Roger Wattenhofer

October 3, 2011

Acknowledgements

I would like to acknowledge the great support from my supervisor, Raphael Eidenbenz. He spent a lot time having regular meetings, rehearsing my presentations, reviewing my report. He also provided lots of great ideas.

I also want to thank Prof. Dr. Roger Wattenhofer for the several meetings where he helped me set the final goal and gave me many good suggestions.

At last, I would appreciate Jasmin Smula for her invitation of the birthday party and the delicious cheese cake. And many thanks to the whole group for recommendations during my first presentation.

Abstract

Bookcrossing is a unique website where books are shared among members and travel around the world. In this paper, we study several features of Bookcrossing regarding book flows, book genres and social networking. We first examine the direction, density, length and speed of book flows and find that the US is the biggest source of book flows, while Finland and Australia are two largest sinks and that length and speed of the book journey are influenced by the crossing zone, the publishing region and the category. Second, we compare book copy distributions and customer reviews in Bookcrossing and Amazon and propose hypotheses to explain the differences. Third, we study the social networking function in Bookcrossing concluding that interactions of Bookcrossing members through books increase the friend probability.

Keywords: Bookcrossing, book flow, book journey

CR Categories: H.2.8 [**Database Applications**]: Data Mining

Contents

Chapter 1

Introduction

Bookcrossing is a website that makes books travel and lets them connect people. The idea of Bookcrossing[1] developed from a website, called Where's George[2], which tracks the circulation of dollar bills. After being launched on April 21, 2001, it soon attracted the attention. In 2004, the word "Bookcrossing" was included in the the Concise Oxford Dictionary [3]. In 2007, Singapore became the first official Bookcrossing country and built over 2,000 "hot spots" [7]. Now it already became a community for book lovers with 964,315 Bookcrossers and 8,220,835 books travelling throughout 132 countries [1].

The journey of a book is like a relay race. Users can register their own books on Bookcrossing and then get a unique identification number for each instance of a book. Upon finishing reading a book, this member can label it with the ID and release it at any place, such as a bench in a park or in front of her home. Optimally, when these information published on Bookcrossing by this member, another Bookcrosser might find this book interesting and go pick it up. She can then post a status on the homepage of this book announcing that she has just found the book. After reading this book, she can release it again just like what her predecessor did, which keeps the book travel.

Several reasons motivate us to study traveling books in Bookcrossing. First, Bookcrossing is the only website that enables books to travel geographically, which is also the biggest feature of Bookcrossing. Second, Bookcrossing makes the world more like a library. People find more free books in more places, such as hotels, restaurant etc. Third, traveling books make more fans of the same book or author become friends offline. Fourth, faster and longer book journeys also mean more active user participation, which brings profits to Bookcrossing.

In this paper, we visualize book flows between different countries and states, analyze the traffic patterns in individual countries or continents. we also study the factors that might influence speed and length of the journey. Those factors include book genres, release places, language and ratings.

[1] www.bookcrossing.com

[2] www.wheresgeorge.com

What's more, we study the social network function in Bookcrossing and find that books passing along different members make them more likely to become friends.

At last, we compare Bookcrossing with other websites and its history data and find some interesting facts.

The remainder of the paper proceeds as follows. In the next section, we review related work. In Section 3 we describe our data set. In Section 4 we study the book flows on country and state levels and then analyze factors that might influence the speed and length of the book journey. Section 5 compares Bookcrossing with Amazon. In Section 6 we focus on the social network side of Bookcrossing. Finally, we summarize our findings in Section 7.

Chapter 2

Related Work

Bookcrossing has gained more and more popularity since this website was launched over 10 year ago. It is, to the best of our knowledge, the only website featuring book flowing, book rating and social networking. Thus, we review related work in the above three fields in this section.

Book Flowing

As far as we know, no one has researched the flow of books in Bookcrossing. However the flow of other objects, such as bank notes, has been studied before. Brockman et al.[12] studied the circulation of over 1 million bank notes in US and concluded that the massive item flows caused by human beings could be modeled by a random walk process with scale free jumps and long waiting time between movements. Our study differs from the previous work by shifting attention from modeling item flows to studying characteristics of book flows, such as direction, speed and length.

Book Rating

Similar to other online book websites, such as Amazon[1] and Goodreads[2], Bookcrossing also has the function of rating books. Ziegler et al.[17] used 1,157,112 ratings they crawled from Bookcrossing to evaluate their topic diversification recommendation method. However our study focuses on the difference between ratings in Bookcrossing and that in Amazon.

Social Networking

There have been extensive studies on social networks. Some of them focused on the causes that make people build relationships. Liben-Nowell et al.[15] studied the relationship between geography and the probability of befriending and

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

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

found that the befriending probability is related to the number of closer candidates. Java et al.[13] found that micro-blogging users with same intentions connect among each other. Thus they recommended that Twitter categorizes users. In this paper, we study the relationship between the interaction of users and probability of becoming friends.

Other work measured network properties of social networks, such as Youtube [10, 16], Twitter [14], Myspace [9], Orkut [16, 9], Cyworld [9], Brightkite [14], Flickr [16], LiveJournal [16]. In the same line of the research, we compare network properties of two networks(interaction and friendship graphs) in Bookcrossing with the properties of selected social networks.

Chapter 3

Data Description

3.1 Entities

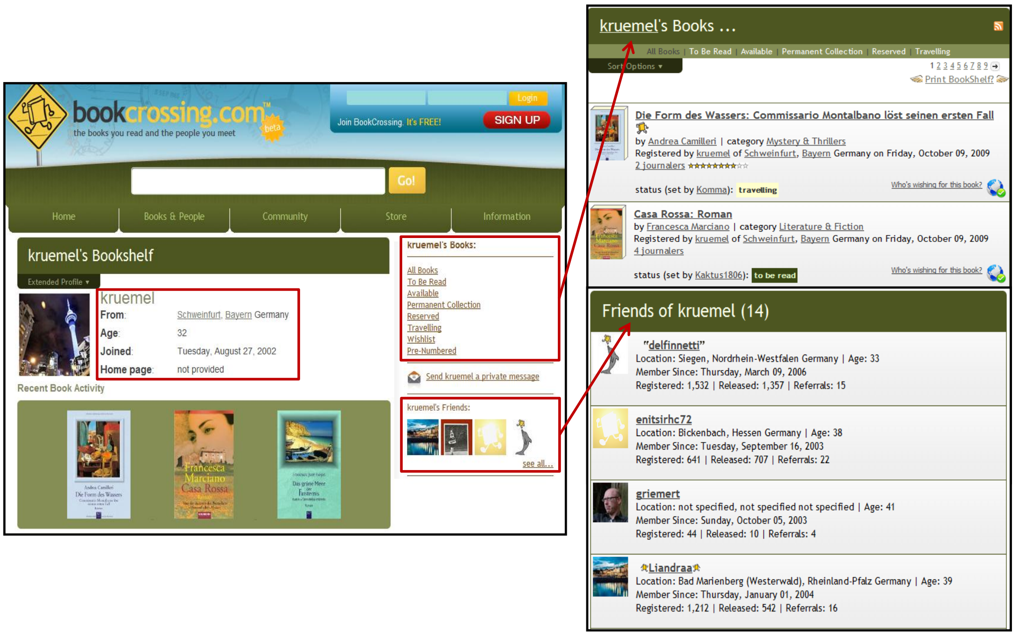
In Bookcrossing, there are several important basic entities:

*Member*s in Bookcrossing have public profile homepages. We can find their ages, hometowns, joining dates, even their friends and books on this page. See Figure 3.1.

Crossing Zone is any place where a book has been released. It can be a park bench, a doctor's office, or a phone booth. There are two kinds of zones. *Virtual zone* indicates the way how books are sent, such as "by mail" or "in person", which are not real places. *Actual zone* means actual places, such as restaurants, bus stops, and places of interests. There are already over 500,000 zones in Bookcrossing. A few actual zones are shown in Figure 3.2.

Book and *Book Copy* are two different concepts in Bookcrossing. A book can be associated with multiple book copies registered by different members. Different copies might also have their individual journeys. *BookCrossing ID*(*BCID*) is identification number given to each book copy registered and unique to that copy of the book only. Figure 3.3 shows the homepage of a book copy.

Journal Entry is a status posted by a certain member on the homepage of a book copy. It might contain information, such as reviews, time stamp, release information etc.



kruemel's Bookshelf

kruemel
From: Schweinfurt, Bayern Germany
Age: 32
Joined: Tuesday, August 27, 2002
Home page: not provided

kruemel's Books:
All Books
To Be Read
Available
Permanent Collection
Reserved
Travelling
Wishlist
Pre-Numbered

Send kruemel a private message

kruemel's Friends:

kruemel's Books ...

All Books | To Be Read | Available | Permanent Collection | Reserved | Travelling

Die Form des Wassers: Commissario Montalbano löst seinen ersten Fall
by Andrea Camilleri | category Mystery & Thrillers
Registered by kruemel of Schweinfurt, Bayern Germany on Friday, October 09, 2009
2 journalers
status (set by Komma): travelling

Casa Rossa: Roman
by Francesca Marciano | category Literature & Fiction
Registered by kruemel of Schweinfurt, Bayern Germany on Friday, October 09, 2009
4 journalers
status (set by Kaktus1806): to be read

Friends of kruemel (14)

"delfinnetti"
Location: Siegen, Nordrhein-Westfalen Germany | Age: 33
Member Since: Thursday, March 09, 2006
Registered: 1,532 | Released: 1,357 | Referrals: 15

enitsirhc72
Location: Bickenbach, Hessen Germany | Age: 38
Member Since: Tuesday, September 16, 2003
Registered: 641 | Released: 707 | Referrals: 22

griemert
Location: not specified, not specified not specified | Age: 41
Member Since: Sunday, October 05, 2003
Registered: 44 | Released: 10 | Referrals: 4

Liandraa
Location: Bad Marienberg (Westerwald), Rheinland-Pfalz Germany | Age: 39
Member Since: Thursday, January 01, 2004
Registered: 1,212 | Released: 542 | Referrals: 16

Figure 3.1: The homepage of a Bookcrossing member.

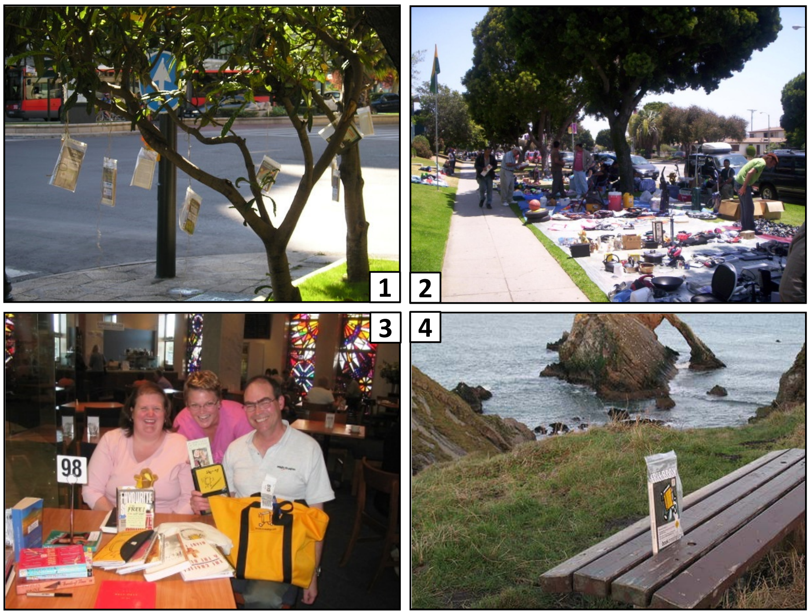
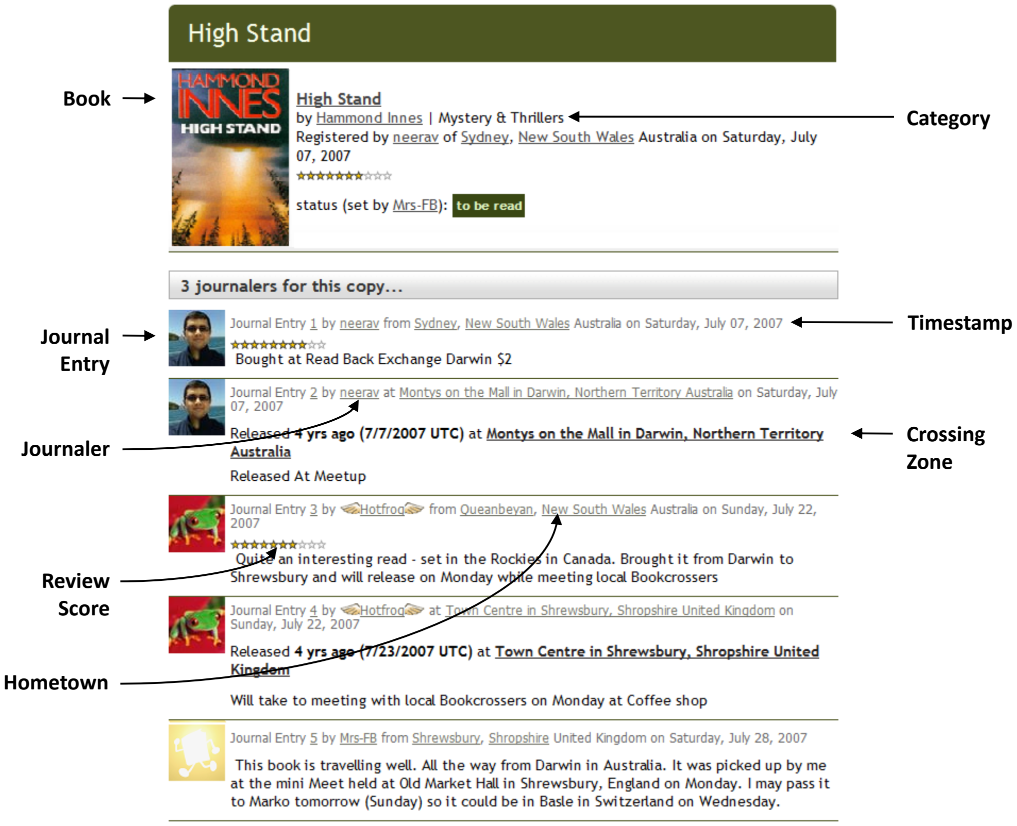


Figure 3.2: Four interesting crossing zones. Photo 1 shows released books hanging on trees. Photo 2 is the yard sale used by Bookcrossers to exchange books. Photo 3 is a meetup of Bookcrossing fellows in a restaurant. Photo 4 is a very romantic release.



High Stand

High Stand
by Hammond Innes | Mystery & Thrillers
Registered by neerav of Sydney, New South Wales Australia on Saturday, July 07, 2007

status (set by Mrs-FB): to be read

3 journalers for this copy...

Journal Entry 1 by neerav from Sydney, New South Wales Australia on Saturday, July 07, 2007

Bought at Read Back Exchange Darwin $2

Journal Entry 2 by neerav at Montys on the Mall in Darwin, Northern Territory Australia on Saturday, July 07, 2007

Released 4 yrs ago (7/7/2007 UTC) at Montys on the Mall in Darwin, Northern Territory Australia

Released At Meetup

Journal Entry 3 by Hotfrog from Queanbeyan, New South Wales Australia on Sunday, July 22, 2007

Quite an interesting read - set in the Rockies in Canada. Brought it from Darwin to Shrewsbury and will release on Monday while meeting local Bookcrossers

Journal Entry 4 by Hotfrog at Town Centre in Shrewsbury, Shropshire United Kingdom on Sunday, July 22, 2007

Released 4 yrs ago (7/23/2007 UTC) at Town Centre in Shrewsbury, Shropshire United Kingdom

Will take to meeting with local Bookcrossers on Monday at Coffee shop

Journal Entry 5 by Mrs-FB from Shrewsbury, Shropshire United Kingdom on Saturday, July 28, 2007

This book is travelling well. All the way from Darwin in Australia. It was picked up by me at the mini Meet held at Old Market Hall in Shrewsbury, England on Monday. I may pass it to Marko tomorrow (Sunday) so it could be in Basle in Switzerland on Wednesday.

Figure 3.3: The homepage of a book copy.

3.2 Entity Relationship Model

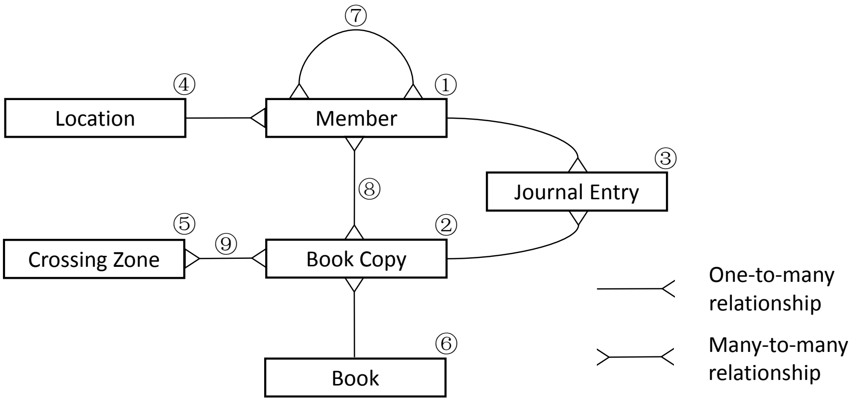


Figure 3.4: Entity relationship model

Figure 3.4 shows the six basic entities in Bookcrossing and their relationships. Some relationships are one-to-many relationships, such as member-to-journal entry, because members can post multiple journal entries. Others are many-to-many relationships, such as member-to-book copy, because a member can be associated to multiple copies and a book copy might pass along several members. We will explain the numbers in the next section.

3.3 Database Schema

We spent two months crawling Bookcrossing and collected information about the six entities and their relationships. Table 3.5 shows 9 tables in our database. The 6 basic entities are also stored in the first 6 tables. The one-to-many relationships are implicitly stored in these 6 tables. For example, journal entry-to-book copy exists in the table "Journal Entry", because the attribute "BCID" links to the table "Book Copy". The many-to-many relationships, such as member-to-book copy, are recorded in three separate tables. What's more, underlined attributes are primary keys. Italic attributes are foreign keys. The table numbers correspond to the numbers in Figure 3.4.

Table 3.5: Database schema.

No.	Table Name	Attributes
1	Member	Name, Age, JoinDate, Country, State, City
2	Book Copy	BCID, ISBN, Journalers, Name
3	Journal Entry	EntryID, Content, *BCID*, *Member*, Timestamp
4	Book	ISBN, Title, Genre, NoOfReviewInBC, AvgReviewInBC, NoOfReviewInAmazon, AvgReviewInAmazon, RankAmazon
5	Location	LocID, City, State, Country, Latitude, Longitude
6	Crossing Zone	ZoneID, Name, City, State, Country Latitude, Longitude
7	Friendship	Member1, Member2
8	Member_Bookcopy	Member, BCID
9	Bookcopy_Crossingzone	BCID, ZoneID

Chapter 4

Book Flows

4.1 The Journey of A Book Copy

In this section, we study the journey of a book copy. We first explain the book copy journey model and then show an example of a book copy's journey. At last, we show a tool developed by ourselves to retrieve and visualize information about the journey.

4.1.1 Book Copy Journey Model

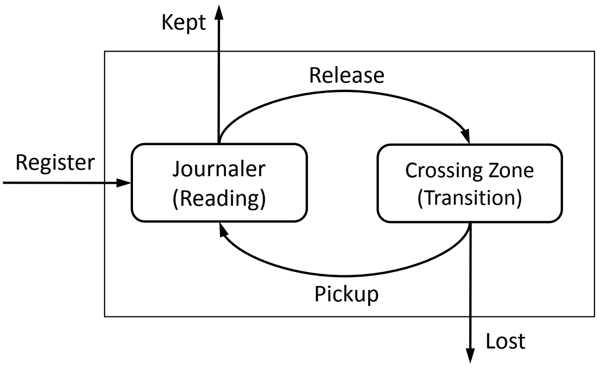
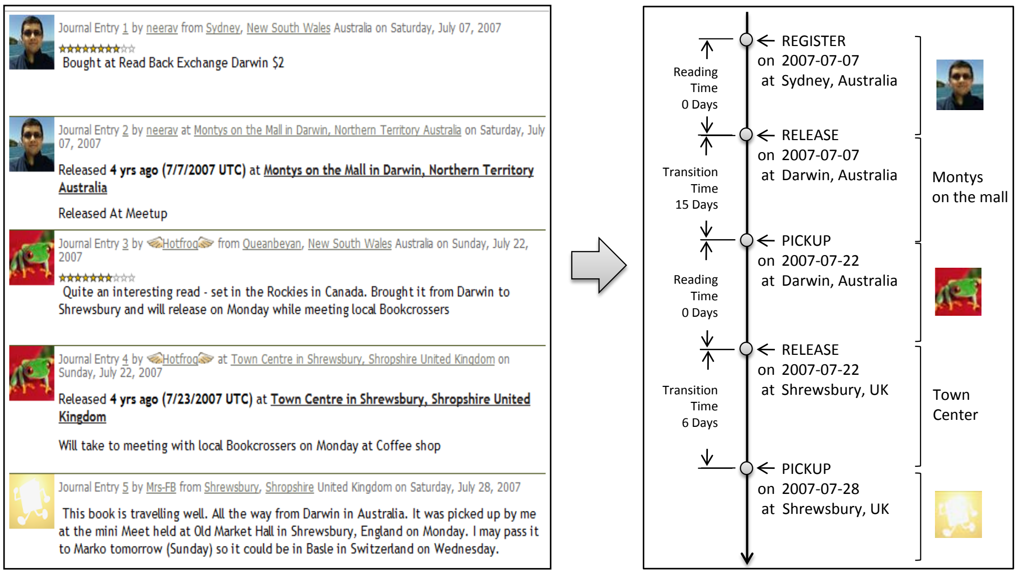


Figure 4.1: Book Copy Journey Model

Figure 4.1 shows the model of a book copy's journey. A book copy's journey starts as it is registered on Bookcrossing. A book copy stays by a journaler and waits in a crossing zone alternatively. It can be released to a crossing zone by a journaler and then picked up by another Bookcrosser. We define *reading time* as the period from the time when a book copy is picked up to the time when it is released and *transition time* as the period when it is in a crossing zone. At last, a book copy's journey terminates, when it is permanently kept by a journaler or lost in a crossing zone.

4.1.2 Book Copy Journey Example



Journal Entry 1 by neerav from Sydney, New South Wales Australia on Saturday, July 07, 2007

Bought at Read Back Exchange Darwin $2

Journal Entry 2 by neerav at Montys on the Mall in Darwin, Northern Territory Australia on Saturday, July 07, 2007

Released **4 yrs ago (7/7/2007 UTC)** at **Montys on the Mall in Darwin, Northern Territory Australia**

Released At Meetup

Journal Entry 3 by Hotfrog from Queanbeyan, New South Wales Australia on Sunday, July 22, 2007

Quite an interesting read - set in the Rockies in Canada. Brought it from Darwin to Shrewsbury and will release on Monday while meeting local Bookcrossers

Journal Entry 4 by Hotfrog at Town Centre in Shrewsbury, Shropshire United Kingdom on Sunday, July 22, 2007

Released **4 yrs ago (7/23/2007 UTC)** at **Town Centre in Shrewsbury, Shropshire United Kingdom**

Will take to meeting with local Bookcrossers on Monday at Coffee shop

Journal Entry 5 by Mrs-FB from Shrewsbury, Shropshire United Kingdom on Saturday, July 28, 2007

This book is travelling well. All the way from Darwin in Australia. It was picked up by me at the mini Meet held at Old Market Hall in Shrewsbury, England on Monday. I may pass it to Marko tomorrow (Sunday) so it could be in Basle in Switzerland on Wednesday.

Reading Time 0 Days
REGISTER on 2007-07-07 at Sydney, Australia
Transition Time 15 Days
RELEASE on 2007-07-07 at Darwin, Australia
Montys on the mall
Reading Time 0 Days
PICKUP on 2007-07-22 at Darwin, Australia
Transition Time 6 Days
RELEASE on 2007-07-22 at Shrewsbury, UK
Town Center
PICKUP on 2007-07-28 at Shrewsbury, UK

Figure 4.2: An example of the reconstruction from journal entries to a book copy journey

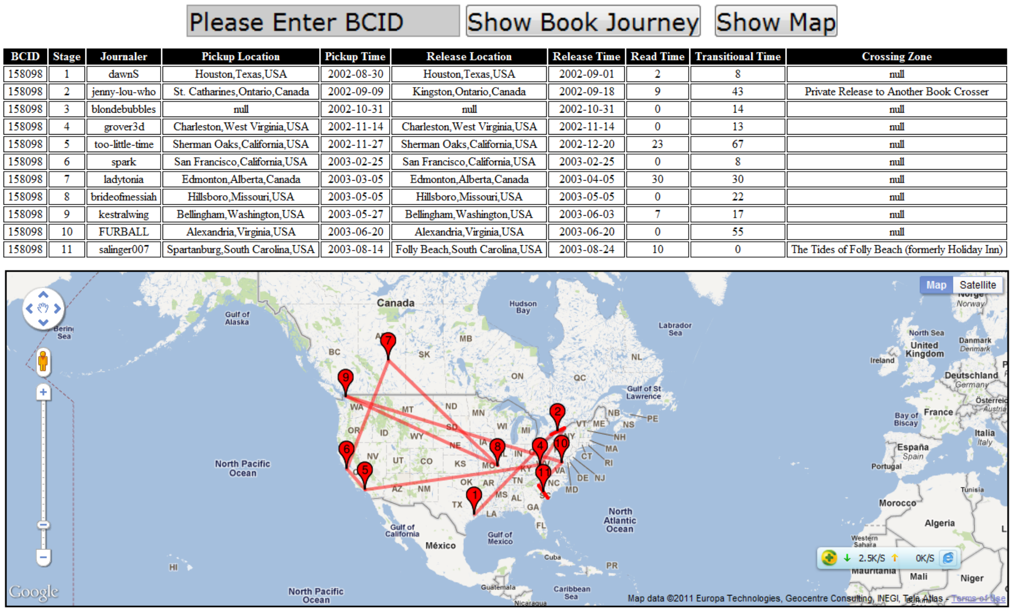
In order to analyze traveling books, we need to reconstruct the book copy journey from its journal entries. Figure 4.2 shows an example of conversion from journal entries to the book copy journey. The homepage of this book copy is on the left, and the book copy journey can be found on the right. Events, such as registrations, pickups and releases, are organized chronologically. Additionally, journalers, crossing zones and their corresponding reading time and transition time are shown on the both sides of events.

4.1.3 Book Copy Journey Visualizer

Book copy journey visualizer can extract and reorganize the journey information from the homepage of a book copy and show them in a table and then visualize the book journey on Google Maps. Figure 4.3 shows the interface of this tool and the journey of a book copy (BCID: 158098). The table in the middle organizes the information of the journey. The tour on Google Maps shows that this book has been to western and eastern coasts of US and Canada.

4.2 Where Do Books Travel?

In this section, we study the book flows on the world map. Book flows are analyzed on two different levels, country level and state level. These two book



BCID	Stage	Journaler	Pickup Location	Pickup Time	Release Location	Release Time	Read Time	Transitional Time	Crossing Zone
158098	1	dawnS	Houston,Texas,USA	2002-08-30	Houston,Texas,USA	2002-09-01	2	8	null
158098	2	jenny-lou-who	St. Catharines,Ontario,Canada	2002-09-09	Kingston,Ontario,Canada	2002-09-18	9	43	Private Release to Another Book Crosser
158098	3	blondebubbles	null	2002-10-31	null	2002-10-31	0	14	null
158098	4	grover3d	Charleston,West Virginia,USA	2002-11-14	Charleston,West Virginia,USA	2002-11-14	0	13	null
158098	5	too-little-time	Sherman Oaks,California,USA	2002-11-27	Sherman Oaks,California,USA	2002-12-20	23	67	null
158098	6	spark	San Francisco,California,USA	2003-02-25	San Francisco,California,USA	2003-02-25	0	8	null
158098	7	ladytonia	Edmonton,Alberta,Canada	2003-03-05	Edmonton,Alberta,Canada	2003-04-05	30	30	null
158098	8	brideofmessiah	Hillsboro,Missouri,USA	2003-05-05	Hillsboro,Missouri,USA	2003-05-05	0	22	null
158098	9	kestralwing	Bellingham,Washington,USA	2003-05-27	Bellingham,Washington,USA	2003-06-03	7	17	null
158098	10	FURBALL	Alexandria,Virginia,USA	2003-06-20	Alexandria,Virginia,USA	2003-06-20	0	55	null
158098	11	salinger007	Spartanburg,South Carolina,USA	2003-08-14	Folly Beach,South Carolina,USA	2003-08-24	10	0	The Tides of Folly Beach (formerly Holiday Inn)

Figure 4.3: Book copy journey visualizer

flows maps can show the direction and density of book flows.

4.2.1 Country-level Analysis

We analyze the book flows on the country level. Some books in Bookcrossing travel from one country to another. We define *source* as the place where a book copy's journey starts and *destination* as the place where its journey ends. We consider here only the books whose source and destination are different.

Figure 4.4 shows country-level book flows. We calculate the difference of inflow and outflow for each country and color-code them accordingly. Red represents that a country has larger inflow than outflow, in other words, acquiring more books than sharing. Blue, on the contrary, means less inflow than outflow. White represents "neutral" countries wherein inflow and outflow are balanced. Numbers on the color bar show the difference of incoming and outgoing books.

As Figure 4.4 shows, Australia and Finland are two largest sinks, while the US is the biggest source. The US "Generously" shared the most books, while a large amount of books ended up staying at Finland and Australia. Most countries in Middle America, Asia and Africa are well balanced. Our data shows that those countries also have few books flowing across the border.

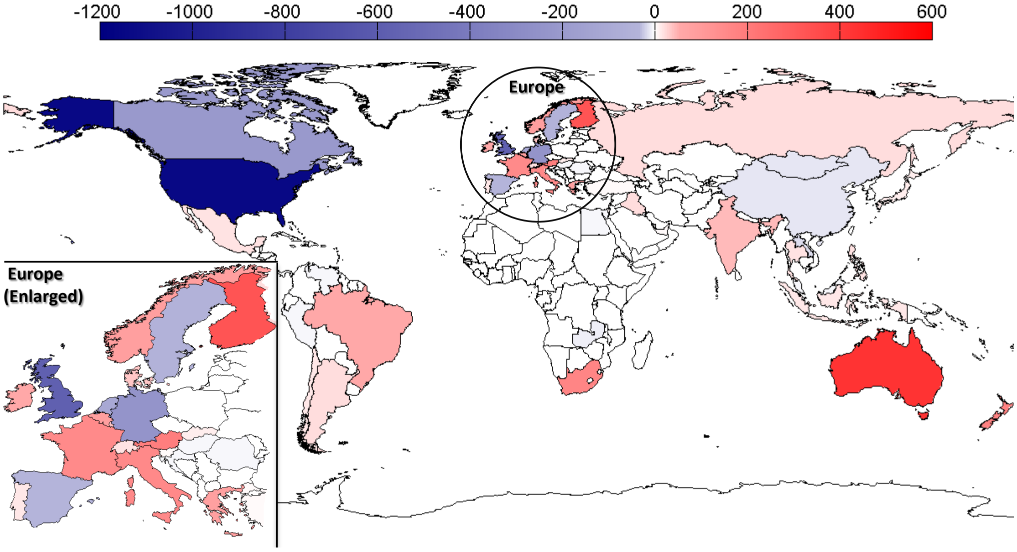


Figure 4.4: Country-level book flows.

4.2.2 State-level Analysis

We also analyze the book flows on the state level. In this section, books that start and terminate in different states or provinces are concerned. We visualize book flows in Figure 4.5 by using jFlowMap [11]. Lines on the map connect two arbitrary states that have had book flows between them. The width of the line depends on the sum of bidirectional flows.

Figure 4.6 shows enlarged views of three heavy traffic countries or continents, US, Australia and Europe. We made two adjustments. First, we notice that most domestic flows are much heavier than international ones. In order to highlight traffic inside concerned countries and continents, we set a threshold of 32 books, which filters out most of long-distance book exchanges. Second, we label some states by the name of the capital city for convenience of readers.

We conclude, from Figure 4.5 and 4.6, that different territories are clustered by languages. There are tight connections between the US, Australia and Europe, especially the UK and Scandinavian countries. People in these countries know English very well. Also, German speaking countries (Germany, Switzerland and Australia) are clustered. What's more, the traffic inside the US and Australia also forms a well-connected network.

Other interesting observations are as follows.

In North America, California is the busiest hub exchanging books with users from other states in US. The reason could possibly be that Bookcrossing was launched in California and therefore has an extensive user base. In addition, compared to western coast, traffic is more uniformly spread among eastern cities.

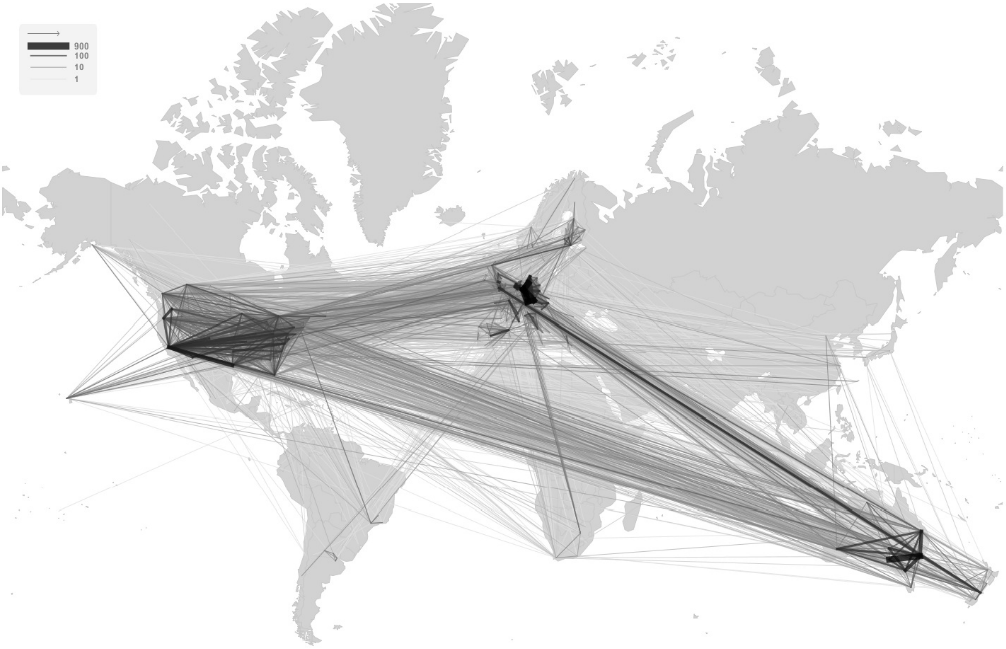


Figure 4.5: State-level book flows (Global).

In Australia, book flows can be found both in western and eastern coasts. Five of all six states in Australia actively participate in crossing books. Most of high-volume connections are linked to New South Wales, whose capital is Sydney.

In Europe, extensive inter-state book flows can be noticed in Germany, which nearly forms a comprehensive book flow network. The heaviest flows, the triangle among Munich, Stuttgart and Dusseldorf, are also in Germany. In France, similar to its population distribution, most flows are concentrated in Paris. In Italy, flows are relatively unbalanced. Most activities exist in North Italy. What's more, high-volume flows can be seen between two biggest Spanish cities, Barcelona and Madrid. At last, there are many Bookcrossing events happening in Netherlands and the UK as well.

4.3 What Makes Books Travel Far?

In this section, we investigate the factors which influence how far books travel. We find three indicators reflecting how far books travel: number of journalers, pickup probability and long-journey book ratio. We focus on four factors: crossing zones, publishing regions, categories and reviews.

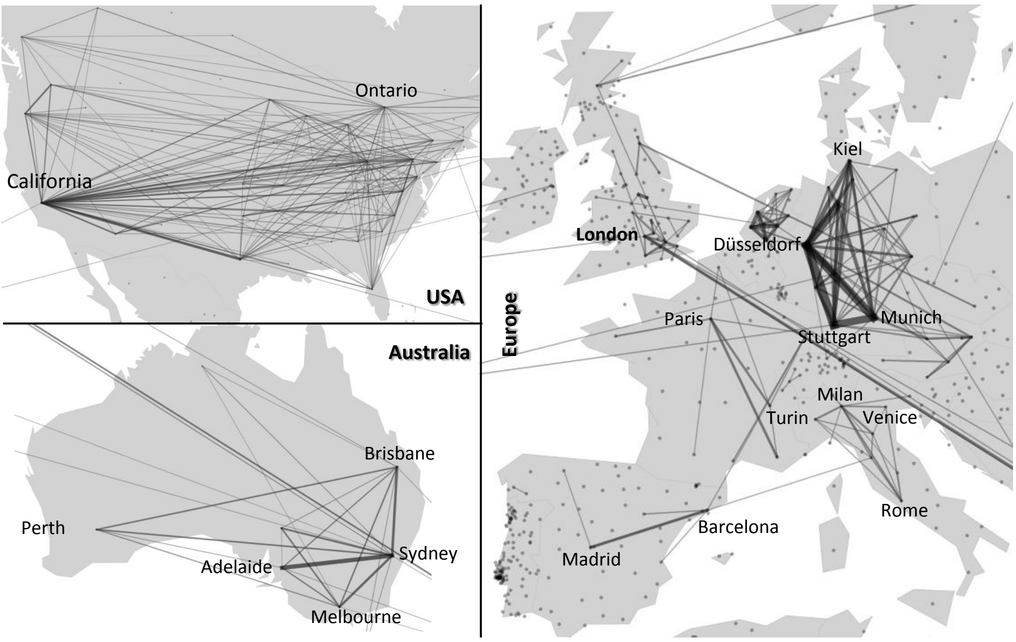


Figure 4.6: State-level book flows (the US, Europe and Australia).

4.3.1 Crossing Zone

The crossing zone is an important factor which influences the journey length, because it is an indispensable part of the book journey. If a book is put in a crossing zone and nobody picks it up, the journey of this book will then stop.

We define 12 kinds of crossing zones, as shown in Table 4.7. Three of them are *virtual zones*, "official Bookcrossing zones", "by mail" and "in person". Official Bookcrossing zones are events organized by Bookcrossing, such as book relay where the sequence of journalers is set before passing a book. Other zones are *actual zones*. For example, Geocache[1] indicates a treasure box where many bookcrossers hide books as a treasure to let others look for them. Apart from Geocache, restaurants and shops etc also belong to actual zones. We have successfully classified 75.01% of all 548,067 crossing zones by using the keywords we have collected. The uncategorized crossing zones usually have ambiguous names or in languages other than English and German.

After categorizing different crossing zones, we calculate *pickup probability* for each kind of crossing zones, as can be formally defined as

$$P(pickup) = \frac{N_{pickup}}{N_{release}} \tag{4.1}$$

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

Table 4.7: 12 kinds of crossingzones and selected key words used to categorize crossing zones.

Crossing zones	**Selected key words**
Official BC Zone	Book Box, Book Relay, Book Ring, Ruhr Crossing, RABCK, OCZ, OBCZ, Book Festival, Unconvention
By mail	Post, mail, post office, send to, postal
In Person	To my acquaintances, trade, BCer, swap, BCX, hand, in person, meet up, Bookcrossing member
Geocache	Geocache
Restaurant	Cafe, bar, club, pub, inn, restaurant, Starbucks
Shop	store, cinema, shopping center, Casino, supermarket Ikea
School	University, institute, school, college, campus
Street and Square	Square, St., center, downtown
Park	Zoo, forest, riverside, park, garden, bench, fountain
Hotel	Hotel, hostel, motel
Place of Interests	Church, library, office building, museum, castle, hospital, town hall
Public Transportation	Airport, bus stop, train station, flight, tram

Table 4.8: Distribution of released book copies by crossing zones and pickup probability of different crossing zones.

Crossing Zones	**Percentage**	**Pickup Prob.**
Official BC Zone	3.49%	42.93%
By mail	4.27%	38.27%
In person	1.99%	30.78%
Geocache	0.14%	12.27%
Restaurant	15.09%	11.94%
Shop	2.49%	6.34%
School	9.70%	6.17%
Street and Square	26.53%	5.84%
Park	2.23%	4.91%
Hotel	2.63%	4.84%
Place of Interests	1.71%	4.37%
Public Transportation	7.33%	3.17%
Others	22.38%	5.45%

where $P(pickup)$ is pickup probability; N_{pickup} means the number of books released in all crossing zones of a certain category; $N_{release}$ means the number of books picked up from all crossing zones of a certain category; *Pickup Probability* indicates the percentage of books that pass though crossing zones.

Table 4.8 shows the distribution of released book copies by crossing zones and the pickup probability for each kind oof crossing zone. We point two interesting observations. First, "Official BC zone", "by mail" and "in person" have highest pickup probability, however they only account for less than 10% released books. If Bookcrossers can make better use of these zones, books might enjoy longer journey. Second, over one third of books are released through "Street and Square" and "Public Transportation", which are among unsafest crossing zones. So it is highly recommended that Bookcrossing encourages members to send books by mail or in person and organizes more official events.

4.3.2 Publishing Region

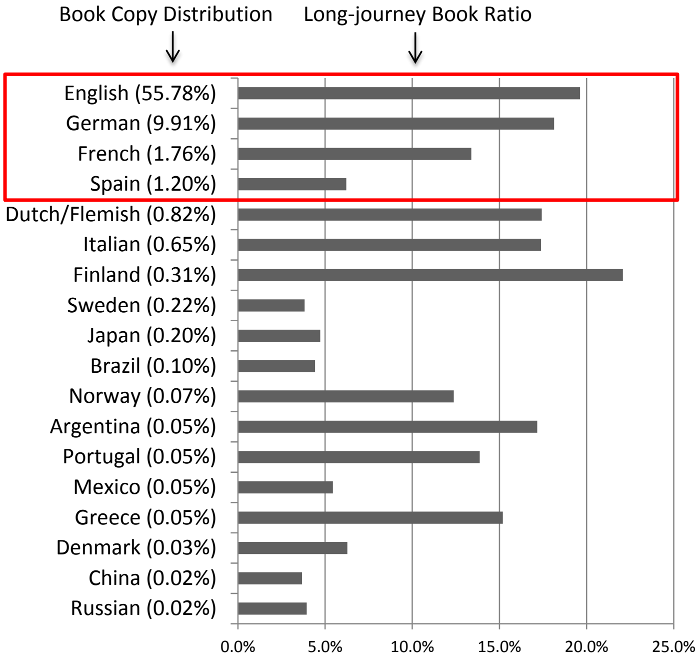


Figure 4.9: Relationship between the book copy distribution and the long-journey book ratio. Bars show the long-journey book ratio of each language. The number beside languages show the proportion of books of this language.

We now examine the publishing region, the second factor that could possibly influence journey length. Since most of registered book copies (76.17%) in Bookcrossing have ISBN, and publishing region information is encoded in ISBN, we could then know the language for each book copy. Figure 4.9 shows the long-journey book ratio for each publishing region and the distribution of book copies by languages. Names of some publishing regions are replaced their languages names. For example, “English” mean English speaking countries. We define *long-journey book copies* as book copies that have been kept by more than one member. We choose this indicator because it makes the differences of languages more obvious.

If we focus on the books in four major language zones in Bookcrossing, (book copies in these four languages account for over 60% registered copies), we can conclude that the long-journey book percentage is proportional to registered copies share of languages. The larger proportion the books in a certain language account for, the longer their journeys are.

4.3.3 Category

Table 4.10: Top(Bottom) 5 categories with longest(shortest) journey length in terms of long-journey book ratio.

	Categories	**Long-journey Book Ratio**
Top 5	Audiobooks	25.44%
	Gay & Lesbian	21.65%
	Graphic Novels	21.59%
	Humor	21.29%
	Women’s Fiction	20.12%
Bottom 5	Education	12.65%
	Religion & Spirituality	12.32%
	Business & Investing	9.35%
	Professional & Technical	7.53%
	Computers & Internet	7.38%

Here we study the category, the third factor. Table 4.10 shows the long-journey book ratio for top 5 and bottom 5 categories. we observe that categories that have high percentage of long-journey books are diversified and that in professional and religious categories, the proportion of long-journey books is small.

4.3.4 Review

Now we study the relationship between customer reviews and how far books travel. For each book copy in Bookcrossing, we get the average review from all its journalers in Bookcrossing and the average review from Amazon. Because users

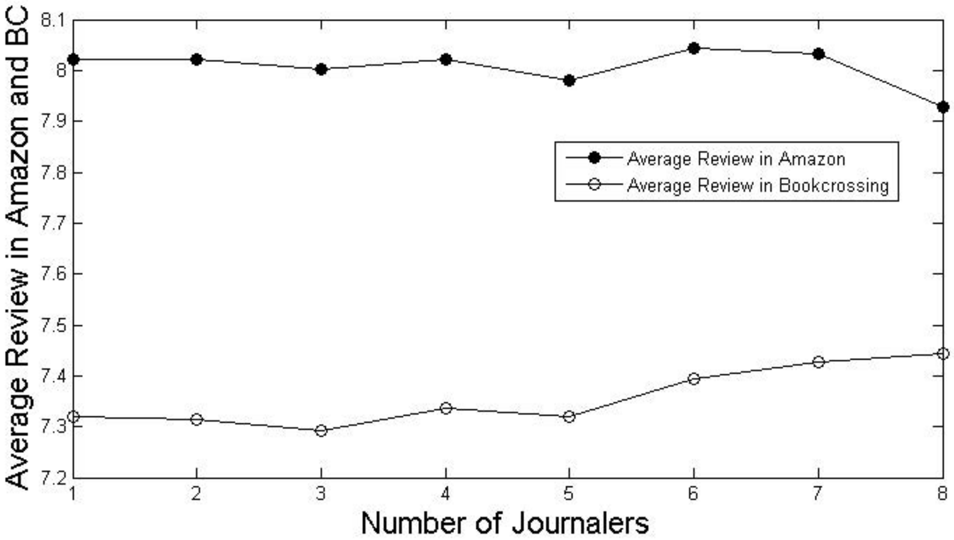


Figure 4.11: Relationship of the average review and number of journalers

in Bookcrossing can rate books from 0 to 10 stars, while Amazon employs a 5-star system, we double the Amazon average review for convenience in comparison.

Figure 4.11 shows the correlation between the number of journalers and the average review in Bookcrossing and Amazon, respectively. We observe that Bookcrossing members rate book copies that have more journalers higher than those having fewer journalers, while no obvious relationship can be noticed between jounnalers and Amazon reviews.

4.4 What Makes Books Travel Fast?

Recall that the period of a book copy journey is consisted of the reading time and the transition time. Reading and transition time can be influenced by many factors, such as books, readers, and crossing zones. In this section, we study the influence of book genres on reading time and the relationship between crossing zones and transition time.

4.4.1 Reading Time

Recall that *reading time* indicates the period from the time when a book is picked up to the time when it is released. We show the top 5 and the bottom 5 categories in terms of reading time in Table 4.12. We observe that "Gay & Lesbian" needs the longest reading time, while "Westerns" has the shortest reading time.

Table 4.12: Top/Bottom 5 categories in terms of reading time

	Categories	**Reading Time (Days)**
Top 5	Gay & Lesbian	150
	Professional & Technical	144
	Home & Garden	137
	Reference	136
	Education	134
Bottom 5	Poetry	105
	Entertainment	104
	Computers & Internet	104
	Children's Books	101
	Westerns	95

4.4.2 Transition Time

The staying time at a certain crossing zone, or *transition time*, also influences the speed of book traveling. We categorize crossing zones in the same way as in section 4.3.1 and calculate the average transition time for each kind of crossing zones, as shown in Table 4.13. The transition time of varying crossing zones ranges from one to four months. Among all, "by mail" and "in person" are two fastest releasing methods, while Geocache has the longest transition time.

Table 4.13: Average transition time by crossing zones

Categories	**Transition Time (Days)**
By mail	37
In Person	42
Restaurant	60
Official Bookcrossing Zone	61
Street and Square	62
Shop	68
Public Transportation	75
Place of Interests	80
School	80
Hotel	100
Geocache	121
Others	60

Chapter 5

Book Genre

5.1 Book Copy Distributions

In this section, we compare book copy distributions in Bookcrossing and in Amazon. Book copy distribution in Bookcrossing means the distribution of registered copies by genres. Book copy distribution in Amazon is the distribution of sold copies by genres.

Three issues related to comparison need to be explained. First, the reason of choosing Amazon is because it represents the new book market, while Bookcrossing can reflect the second-hand market. By comparison, we could find out which category second-hand book lovers prefer reading. Second, Our comparison is based on 40 categories which exist in both book websites. Third, we already have the number of registered copies for each book in Bookcrossing, however the amount of sold copies for each book in Amazon is kept secret [2]. Thus we predict the sold units of each book in Amazon [6] by using the bestseller rank for each book. In Amazon, each book has its own *bestseller rank*, a indicator based on historical sales of every item sold [8],

Based on the data from two websites, we show book copy distributions in Bookcrossing and Amazon in Table 5.1. To compare two websites, we calculate the proportion difference between Amazon and Bookcrossing for each genre and show categories with difference larger than 1% in Figure 5.2. There are 8 genres *underrepresented* in Bookcrossing, which means that they have smaller share in Bookcrossing than in Amazon. On the contrary, 4 other genres are *overrepresented.*

We conclude four characteristics for the underrepresented categories.

1. Expensive. Valuable books, such as business and textbooks, are not suitable for sharing.

2. Multiple read. People read religion books constantly. These books are more suitable for collection.

Table 5.1: Distribution of registered(sold) copies in Bookcrossing(Amazon) by genres

Category	**Sold Units in Amazon**	**Registered Units in Bookcrossing**
Arts & Photography	0.55%	0.41%
Audio Downloads	0.00%	0.01%
Audiobooks	0.72%	0.25%
Biographies & Memoirs	3.55%	3.46%
Business & Investing	1.79%	0.36%
Children's Books	9.67%	7.34%
Computers & Internet	0.19%	0.20%
Cooking, Food & Wine	0.79%	0.71%
e-Books	0.03%	0.03%
Education	1.86%	0.65%
Entertainment	1.79%	1.12%
Gay & Lesbian	0.06%	0.12%
Graphic Novels	0.33%	0.42%
Health, Mind & Body	4.79%	1.42%
History	2.39%	1.69%
Home & Garden	0.22%	0.31%
Horror	0.62%	1.42%
Humor	2.16%	2.24%
Journals	0.14%	0.22%
Literature & Fiction	21.16%	30.26%
Mystery & Thrillers	14.79%	17.72%
Nonfiction	5.39%	2.88%
Outdoors & Nature	0.25%	0.31%
Parenting & Families	1.75%	0.37%
Pets & Animals	0.22%	0.40%
Philosophy	1.43%	0.52%
Plays & Scripts	0.49%	0.43%
Poetry	0.35%	0.81%
Professional & Technical	0.39%	0.18%
Reference	0.60%	0.44%
Religion & Spirituality	2.86%	1.55%
Romance	3.25%	7.13%
Science	0.66%	0.40%
Science Fiction & Fantasy	4.96%	4.92%
Sports	0.39%	0.30%
Teens	3.44%	2.00%
Travel	0.57%	1.15%
Westerns	0.12%	0.23%
Women's Fiction	1.57%	2.72%
Others	3.69%	2.91%

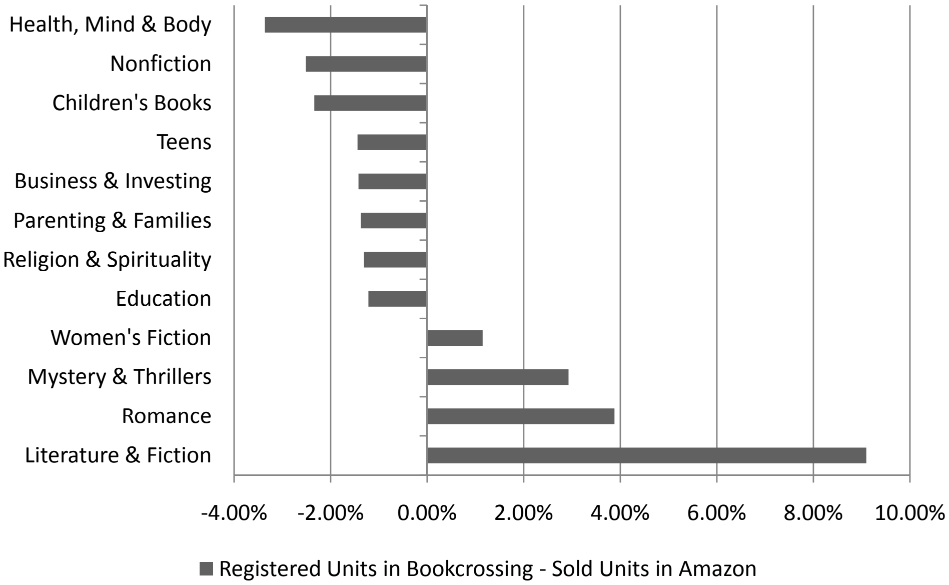


Figure 5.2: 12 selected book genres whose shares in Bookcrossing and Amazon have more than one percent difference.

3. Privacy exposure. Books, like "Health, Mind and Body", expose privacy of book readers, because mental and physical illnesses can be easily guessed from books they read.
4. Narrow audience. Some categories, such as "Business & Investing" and "Parenting & Families", only aim at a specific group of readers.

However, *overrepresented* genres, mainly fictions and novels, are relatively cheap, usually read-once, in most cases privacy protective, and enjoy wide audience, which makes them more welcomed in Bookcrossing.

5.2 Customer Reviews

Here we study customer reviews in Amazon and Bookcrossing. In Bookcrossing, a book can be rated from one to ten stars, while in Amazon books can be rated from one to five stars. So we crawled the average customer review in Amazon for each book and then doubled the Amazon review for comparison. Table 5.3 shows the average review in Bookcrossing and Amazon for each genre.

There are several interesting observations. First, we notice that all genres are rated higher by Amazon users than by Bookcrossing members. Our hypothesis is that this could possibly be caused by the difference of 5-star and 10-star

Table 5.3: Average review scores on the same set of books by customers in Amazon and Bookcrossing grouped by categories

Category	**Review in Bookcrossing**	**Review in Amazon**	**Amazon - Bookcrossing**
Health, Mind & Body	7.0	8.6	1.6
Business & Investing	6.8	8.3	1.5
Cooking, Food & Wine	7.2	8.7	1.5
Audio Downloads	7.4	8.8	1.4
Humor	7.1	8.4	1.3
Graphic Novels	7.1	8.4	1.3
Philosophy	7.2	8.5	1.3
Gay & Lesbian	6.9	8.2	1.3
Sports	7.1	8.4	1.3
Entertainment	6.9	8.2	1.3
Religion & Spirituality	7.4	8.7	1.3
Home & Garden	7.3	8.6	1.3
Poetry	7.7	8.9	1.2
Teens	7.3	8.5	1.2
Westerns	7.4	8.6	1.2
Professional & Technical	7.0	8.1	1.2
Biographies & Memoirs	7.3	8.4	1.1
Romance	7.0	8.0	1.0
Children's Books	7.8	8.8	1.0
Pets & Animals	7.5	8.5	1.0
Reference	7.3	8.3	1.0
Literature & Fiction	7.1	8.1	1.0
Education	7.4	8.3	1.0
Nonfiction	7.3	8.2	1.0
Outdoors & Nature	7.7	8.7	0.9
Science	7.4	8.3	0.9
Plays & Scripts	7.7	8.6	0.9
Women's Fiction	6.9	7.8	0.9
Computers & Internet	7.1	8.0	0.9
Science Fiction & Fantasy	7.2	8.1	0.9
Arts & Photography	7.6	8.5	0.9
History	7.4	8.3	0.9
Travel	7.3	8.2	0.9
Parenting & Families	7.6	8.5	0.9
Horror	7.1	8.0	0.9
Journals	7.5	8.3	0.8
e-Books	7.5	8.3	0.8
Audiobooks	7.3	8.0	0.8
Mystery & Thrillers	7.2	7.9	0.7
Others	7.0	8.2	1.2

rating systems or that most users are satisfied with Amazon services, such as delivery. Second, “Health, Mind & Body” and ”Business & Investing” are two genres with biggest differences in reviews. Also shares of these two genres have obvious disparities between Amazon and Bookcrossing. In other words, these two genres are underrepresented and badly rated in Bookcrossing. Third, “Mystery & Thriller” has the tiniest review gap with Amazon and is also overrepresented in Bookcrossing.

5.2.1 Case Study

We now study three books with largest average review differences between Bookcrossing and Amazon. We select these three books based on two conditions.

- First, the largest average review difference in two websites (two books better rated in Bookcrossing, the other better reviewed in Amazon).
- Second, enough reviews received in both websites (500+ reviews in Amazon and 10+ reviews in Bookcrossing).

Table 5.4: Three books in the case study and their number of reviews in Amazon and Bookcrossing and their review difference.

Book	**# Reviews in BX**	**# Reviews in Amazon**	**Review Diff.**
Endlich Nichtraucher (The easy way to stop smoking)	17	555	2.4
Eat right 4 your type	11	619	2.4
Blow fly	56	756	-6

We set the limit of 500+ in Amazon and 10+ in Bookcrossing, because “Harry Potter Paperback Boxed Set (Books 1-4)”, which can be believed to have received enough reviews, has 423 reviews in Amazon. Since only 1.4% books in Bookcrossing have more than 10 reviews, books with more than 10 reviews can be considered well reviewed. See Table 5.4 for details.

Figure 5.5 shows the covers for three books. The left and middle ones have the largest Amazon-Bookcrossing differences (well rated in Amazon, badly reviewed in Bookcrossing). The right one has the largest Bookcrossing-Amazon difference. See also Table 5.4.

“Endlich Nichtraucher” is a advisory book for people who want to quit smoking. The book received splendid reviews on Amazon, however is not well rated in Bookcrossing where most of bad reviews come from those smokers who failed to quick smoking.

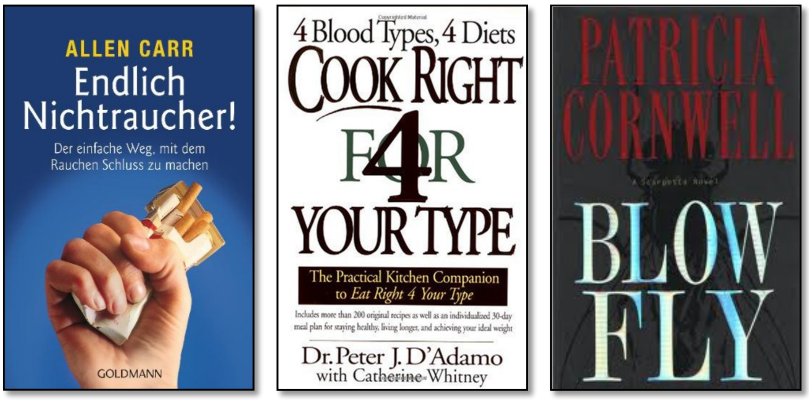

ALLEN CARR
Endlich Nichtraucher!
Der einfache Weg, mit dem Rauchen Schluss zu machen
GOLDMANN
4 Blood Types, 4 Diets
COOK RIGHT FOR 4 YOUR TYPE
The Practical Kitchen Companion to Eat Right 4 Your Type
Includes more than 200 original recipes as well as an individualized 30-day meal plan for staying healthy, living longer, and achieving your ideal weight
Dr. Peter J. D'Adamo
with Catherine Whitney
PATRICIA CORNWELL
A Scarpetta Novel
BLOW FLY

Figure 5.5: Three books with biggest differences in Amazon and Bookcrossing.

"Each Right 4 Your Type" is a also recommendation book which makes an argument that people should have individualized diets according to their blood types. The majority of buyers on Amazon like it. But many Bookcrossers disagree with the argument.

"Blow Fly" is a novel from a famous author Patricia Cornwell. It is rated only two stars in Amazon, because many readers think it disappointing. However, it is rated much higher in Bookcrossing, because bookcrossers think that fans of this author must have a thorough collection.

5.3 Historical Changes

At last, we study historical changes of the book copy distribution. Since we know the register time for each book copy, we could calculate each year's *cumulative book copy distribution* by genres, which can be defined as the distribution of all books registered before the specific year by genres.

We focus on the proportion variation for 40 genres and choose 4 out of them which have the largest peak/bottom difference, as shown in Figure 5.6. We observe that shares of "Children's Books" and "Women's Fiction" shrink over the past ten years, while "Sci. Fiction & Fantasy" and "Religion & Spirituality" grow.

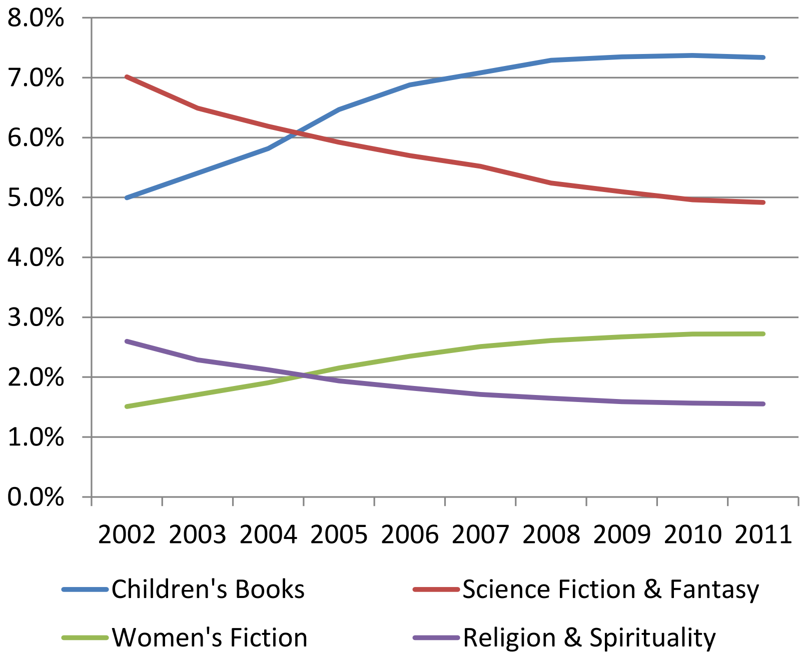


Figure 5.6: Horizontal analysis of genre distribution in Bookcrossing

Chapter 6

Social Networking

6.1 Friend Probability

In Bookcrossing, books make interactions between users happen. For example, if a book copy is passed on to five people one after another, five members can see journal entries posted by others on the homepage of this copy. The interactions caused by books connect members implicitly. However, members can add other Bookcrossers as friends like people do in Facebook, which establishes explicit connection. In this section, we study the conversion from implicit interactions to explicit friendships.

Definition 1 (Interaction)**.** *Two members have interaction, if a specific book copy has been in possession of both members.*

Definition 2 (Friendship)**.** *Two members are officially friends in Bookcrossing.*

Definition 3 (Interaction Graph)**.** *We use $G_1(V, E)$ to describe the interaction between members. V is the set of all Bookcrossing members. E is a set of interaction relationship, if $(u, v) \in E$, members u and v once had the same book copy. And we define $w(u, v)$ as the number of book copies both u and v once had.*

Definition 4 (Friendship Graph)**.** *We use $G_2(V, F)$ to describe the friendship between members. V is the set of all Bookcrossing members. F is a set of friendships, if $(u, v) \in F$, members u and v are friends.*

Definition 5 (Friend Probability)**.** *Friend probability shows the conversion from the interaction to friendship. We define friend probability on all members as,*

$$\Pr = \Pr\left[(u,v) \in F | (u,v) \in E'\right] = \frac{|E' \cap F|}{|E'|} \qquad \forall (u,v) \in E' \tag{6.1}$$

where E' is a subset of E. $|E'|$ is the number of all interaction edges in E'

Bookcrossing members can interact in different degrees, which depends on how many books pass along both users. The more common books they have, the more they interact. We now study the relationship between common books and the conversion from the interactions to the friendship. This relationship can be formally expressed as follows

$$\Pr(c) = \Pr\left[(u,v) \in F | (u,v) \in E \ \& \ w(u,v) \geq c\right] \qquad \forall (u,v) \in E, c = 1, 2... \quad (6.2)$$

where s is the number of common book copies.

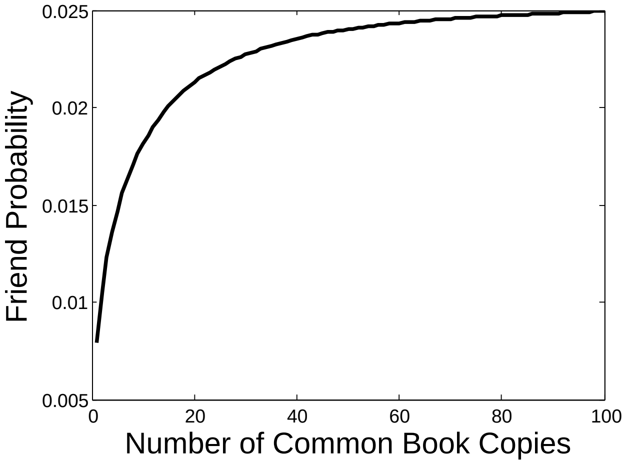


Figure 6.1: Relationship between friend probability and number of common book copies.

Figure 6.1 shows the result. The calculation is based on a data set of 705,686 members, 1,663,880 interactions and 112,364 friendships. we cut off the part with more than 100 common book copies because of very limited amount of data available. From Figure 6.1, we conclude that the more common books two members have, more likely they become friends, which means Bookcrossing is community that help members make new friends. It also tells Bookcrossing users that the rewards of actively participating in crossing books are more friends.

6.2 Network Property

Many previous studies have focused on the relationships between users. In Bookcrossing, both interaction graph and friendship graphs represent the user relationships. So we selected two popular social networks (Myspace and Orkut) and compare the network properties of them with that of two graphs in Bookcorssing.

Table 6.2 shows some basic network properties of these four above mentioned graphs. For interaction graph and friendship graph, we filter out all the isolated

Table 6.2: Basic network properties of interaction and friendship graphs in Bookcrossing, relationship graphs in Myspace and Orkut

Properties	**Interaction**	**Friendship**	**Myspace**	**Orkut**
Sampling Ratio	17.68%	3.9%	0.08%	0.3%
# Nodes	124,751	27,574	100,000	100,000
# Links	1,663,880	112,364	6,854,231	1,511,117
Average Degree	13.338	4.247	137.1	30.2
Clustering Coefficient	0.262	0.125	0.26	0.31
# Components	10,743	3,634	1	1

notes, which means that we wipe out those members who have no interactions with others in interaction graph and no-friend members are also deleted. Sampling ratio means the proportion of the rest users. Average degree represents the connections of a certain a member to other users. Clustering coefficient is an indicator for network connectivity showing how well the direct neighbors of a certain member are connected. Components are isolated clusters.

From the number of nodes and links, we can see that interaction is much more comprehensive in Bookcrossing than friendship. Different users are more easily involved in interactions. Average degree and clustering coefficient show that the interaction graph is better connected. Not only has each user more extensive connections to others on average, but their direct neighbors are more tightly connected as well. Both facts tell that those potential users who have been involved in interactions but not yet in friendship are a huge group. We could recommend those users, which would be much easier for people to trust than strangers.

What's more, we also list the network properties of two famous social networks (Myspace and Orkut) [9]. Myspace is, as of September 2007, the largest social network in the world. While Orkut was launched by Google and is very extensively adopted by India. Ahn Y.-Y. et al collected a sample 100,000 members from both websites who are all directly or indirectly connected to a randomly selected seed user. So in Table 6.2, their component number is 1. And 100,000 users account for 0.08% and 0.3% respectively in Myspace and Orkut.

From the comparison, we notice that the size of Bookcrossing is much smaller than that of Orkut and Myspace. Because the sample sizes of four graphs are similar, however the sampling rates of the latter two is much smaller. Additionally, Myspace and Orkut, as two very mature social networks, have much better connectivity than Bookcrossing. The reason could be that Bookcrossing, as book-based social network, mostly attracts book lovers. While Orkut and Myspace have wider audience. which results in tighter and larger network structure.

Chapter 7

Conclusions and Outlook

7.1 Conclusions

This paper presents our findings from three perspectives. First, we have studied the book flows on country and state levels by visualizing book journeys on the world map and found that there are heavy traffic among and inside three areas, North America, Europe and Australia and that the US is the biggest source, while Finland and Australia are two largest sinks. Second, We have also compared book copy distributions and customer review in Bookcrossing and Amazon and tried to explain the differences between Bookcrossing and Amazon. Third, our result has shown that the friend probability between two members grows, as their interactions increase.

7.2 Outlook

We propose future work in the following four directions.

First, in criminology, the amount of unreported or undiscovered crime, also called *dark figure*, is an important research topic [4]. Similarly, the group of people who share books but are not Bookcrossing users yet should also acquire the attention. Because it helps Bookcrossing be able to target and develop potential users. The study of this issue can be done by interviewing the sample group of subjects or performing questionnaires.

Second, a cell phone application can be developed to show released free books in the neighborhood on Google Maps. It can also have a "fast registration" function. Bookcrossing members can register books by scanning the barcode on the back of books, which would be much faster than registering a book on Bookcrossing.com.

Third, Goodreads is another online community for book lovers. It encourages members to organize their book shelves, rate and review books and discuss them with other members [5]. Since both Goodreads and Bookcrossing have functions, such as rating and social networking, we can compare these two websites.

Fourth, recall that we classify 548,067 crossing zones, in order to study journey length. However, 24.9% crossing zones can not be categorized, because some zones, like “Montys on the mall” (A restaurant), do not indicate their category literally. Location-based social networks, such as foursquare[1], provide mappings between location names and their categories. We can make use of APIs from those websites to categorize more crossing zones.

[1] http://foursquare.com/

Bibliography

[1] About bookcrossing. `http://www.bookcrossing.com/about`.

[2] The book industry's best-seller lists. `http://www.slate.com/id/3504/`.

[3] Bookcrossing wikipedia. `http://en.wikipedia.org/wiki/BookCrossing`.

[4] Dark figure. `http://en.wikipedia.org/wiki/Dark_number`.

[5] Need advice on what to read? ask the internet. `http://goo.gl/UiH3n`.

[6] Power law converting amazon sales ranks to units sold. `http://goo.gl/TFgNe`.

[7] Singapore is first bookcrossing country in the world. `http://goo.gl/wjwhg`.

[8] What amazon bestsellers rank means. `http://www.amazon.com/gp/help/customer/display.html?nodeId=525376`.

[9] Ahn, Y.-Y., Han, S., Kwak, H., Moon, S., and Jeong, H. Analysis of topological characteristics of huge online social networking services. *Proceedings of the 16th international conference on World Wide Web - WWW '07* (2007), 835.

[10] Benevenuto, F., Duarte, F., Rodrigues, T., a.F. Almeida, V., Almeida, J. M., and Ross, K. W. Understanding video interactions in youtube. *Proceeding of the 16th ACM international conference on Multimedia - MM '08* (2008), 761.

[11] Boyandin, I., Bertini, E., and Lalanne, D. Using flow maps to explore migrations over time. In *Geospatial Visual Analytics Workshop in conjunction with The 13th AGILE International Conference on Geographic Information Science* (2010).

[12] Brockmann, D., Hufnagel, L., and Geisel, T. The scaling laws of human travel. *Nature 439*, 7075 (Jan. 2006), 462–5.

[13] Java, A., Song, X., Finin, T., and Tseng, B. Why We Twitter : Understanding Microblogging. *Network pages*, ACM Press (2007), 56–65.

[14] Li, N., and Chen, G. Analysis of a Location-Based Social Network. *2009 International Conference on Computational Science and Engineering* (2009), 263–270.

[15] Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., and Tomkins, A. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America 102*, 33 (Aug. 2005), 11623–8.

[16] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. Measurement and analysis of online social networks. *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement - IMC '07* (2007), 29.

[17] Ziegler, C., McNee, S., Konstan, J., and Lausen, G. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web* (2005), ACM, pp. 22–32.