A NETWORK APPROACH TO TOPIC SUMMARY AND KNOWLEDGE DISCOVERY
IN SOCIAL TAGGING

BY

XIN XIANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Professor Linda C. Smith, Chair
Associate Professor P. Bryan Heidorn, University of Arizona, Director of Research
Professor John M. Unsworth
Associate Professor ChengXiang Zhai
Abstract

As evidenced by the growing popularity of collaborative tagging sites like librarything, last.fm and del.icio.us, social tagging has provided a social and information organizing platform that warrants public attention and academic investigation alike. This doctoral research focuses on studying the semantic relations between social tags, items and content creators through co-occurrence analysis, social network analysis and information visualization, thus revealing the role played by social tags in representing and classifying contents and creators, and implications they might have for facilitating information seeking practice, particularly knowledge discovery and information summary, and as a result, helping the design of information retrieval and browsing interface. User-oriented studies are conducted to evaluate the advantage of visual and presentational features based on tagging analysis over existing constructs such as tag clouds in performing high-level information seeking tasks.

The social tagging paradigm is widely considered an extension beyond keyword-based indexing and hierarchical classification schemes. The new massive manual indexing method characterized by social tagging differs from automatic indexing that lays the foundation of modern information retrieval in that its manual nature obviates the common pitfalls of computer-based automatic indexing. It also complements traditional manual indexing since tag word distribution reflects the opinions of a large number of people with various background and knowledge instead of a limited number of domain experts who are dominant in the classification and cataloging undertakings.

Parallel to the observation that an individual’s social identity is defined by the collectivities to which the individual belongs, the topical, temporal, geographic, and stylistic features
of an information item (book, song, etc.) are represented by the tags that are applied to it in a social tagging context. Employing similarity analysis, bipartite social network theory and small-world network model, this study analyzes the patterns and trends of networks formed through co-occurrences of tags, and clusters and community structures found in the networks that convey topical or stylistic cues of the underlying items.

The abundant tagging data available at public tagging sites makes it possible to reveal the relations between tags, items and creators from a social network perspective. A small-world network, characterized by low average path length and high clustering coefficient compared to a randomly generated network with a similar number of nodes and edges, is a typical form of social network frequently found in real world social networks and physical networks. The study demonstrates the small-world network property of networks of tags/authors in the book tagging site librarything, presenting the network of tags/authors as groups of highly related items which can be detected by community detection methods and convey semantic meanings.

The big picture of the tagging universe aside, several network reduction techniques are used to contract the original tag/author networks containing thousands of nodes to a smaller network of around 150 nodes for better visualization and presentation of details, especially of those most salient tags and creators. The exact number of nodes in the contracted network is fine-tuned based on weights of edges and topological characteristics of the network. Networks of different scales seem to present different levels of information about the tagging world.

A user study is conducted to evaluate the effectiveness of visual constructions based on similarity and network analysis for several tasks that they can be used to support. Participants are presented with both visualizations based on tagging analysis and an existing interface such as tag cloud, and asked to determine which one is more instructive and helpful in performing different high-level information seeking tasks, such as topic summary, grouping and navigation.

This dissertation work aims at studying tagging data from a social network perspective,
providing clues as to how analysis of tagging data offers new angles to interpret semantic
relations and facilitate content presentation and discovery. It seeks to strengthen the func-
tion of social tagging as a point of connection between personal content management and
serendipitous knowledge discovery in a social context.
# Table of Contents

List of Tables . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . viii

List of Figures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . x

List of Symbols . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . xii

Chapter 1 Introduction .................................................. 1
  1.1 Background ......................................................... 2
    1.1.1 Hierarchical Classification Schemes .......................... 2
    1.1.2 OPAC and Exploratory Information Seeking ................... 3
    1.1.3 The Social Tagging Phenomenon ............................ 5
    1.1.4 Tag Clouds as Interface to the Tagging World ............ 9
  1.2 Purpose of the Study ............................................. 11
  1.3 Research Questions ............................................. 17

Chapter 2 Literature Review .......................................... 19
  2.1 Keyword-based Indexing and Hierarchical Classification .......... 19
  2.2 Information Seeking and Visualization .......................... 20
  2.3 Social Network Analysis and Visualization ....................... 24
    2.3.1 Affiliation Networks and One-Mode Networks ................ 25
    2.3.2 Small-World and Scale-Free Network ........................ 26
    2.3.3 Social Network Visualization ................................ 29
  2.4 Classification and Cluster Analysis ............................ 30

Chapter 3 Research Design and Methods ............................ 32
  3.1 Data Collection .................................................. 33
    3.1.1 Web Crawling ................................................. 33
    3.1.2 Database Store and Schema Mapping ........................ 35
  3.2 Data Modeling .................................................... 36
    3.2.1 The Mathematical Definition of Folksonomy ................. 36
    3.2.2 Affiliation Networks and Bipartite Graphs ................. 37
    3.2.3 Representation: The Bag-of-Words Approach ............... 38
    3.2.4 Term Weighting Schemes .................................... 39
  3.3 Data Analysis .................................................... 42
    3.3.1 Co-occurrence Analysis ................................. 42
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3.2</td>
<td>Similarity Measures</td>
<td>43</td>
</tr>
<tr>
<td>3.3.3</td>
<td>From Similarity Matrix to Network</td>
<td>47</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Detection of Community Structure</td>
<td>52</td>
</tr>
<tr>
<td>3.4</td>
<td>Visualization and User Study</td>
<td>54</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Visualization of Network Data</td>
<td>54</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Experimental Design</td>
<td>56</td>
</tr>
<tr>
<td>3.5</td>
<td>Summary</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td><strong>Chapter 4</strong></td>
<td><strong>Tag Distribution and Small-World Properties</strong></td>
</tr>
<tr>
<td>4.1</td>
<td>Tag Distribution in Two-Mode Network</td>
<td>72</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Degree Distribution</td>
<td>72</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Occurrence vs Rank</td>
<td>74</td>
</tr>
<tr>
<td>4.2</td>
<td>Small-World Properties of One-Mode Networks</td>
<td>76</td>
</tr>
<tr>
<td>4.3</td>
<td>Dynamics of Small-World Properties</td>
<td>78</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Distribution of Similarity Values</td>
<td>79</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Clustering Coefficient</td>
<td>80</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Average Degree</td>
<td>81</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Average Path Length</td>
<td>81</td>
</tr>
<tr>
<td>4.4</td>
<td>Degree Distribution of One-Mode Networks</td>
<td>85</td>
</tr>
<tr>
<td>4.5</td>
<td>Summary</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td><strong>Chapter 5</strong></td>
<td><strong>Identifying Groups in One-Mode Networks</strong></td>
</tr>
<tr>
<td>5.1</td>
<td>Fine Tuning the Parameters</td>
<td>90</td>
</tr>
<tr>
<td>5.1.1</td>
<td>Term Weighting Schemes</td>
<td>91</td>
</tr>
<tr>
<td>5.1.2</td>
<td>Similarity Calculation Methods</td>
<td>95</td>
</tr>
<tr>
<td>5.1.3</td>
<td>Network Construction Methods</td>
<td>96</td>
</tr>
<tr>
<td>5.2</td>
<td>Closer Examination of Results from Community Detection</td>
<td>98</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Cliques and Average Clustering Coefficient</td>
<td>98</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Two Author Network Examples</td>
<td>99</td>
</tr>
<tr>
<td>5.3</td>
<td>Summary</td>
<td>110</td>
</tr>
<tr>
<td></td>
<td><strong>Chapter 6</strong></td>
<td><strong>User Study</strong></td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>113</td>
</tr>
<tr>
<td>6.2</td>
<td>Apparatus</td>
<td>114</td>
</tr>
<tr>
<td>6.3</td>
<td>Number of Groups Identified</td>
<td>115</td>
</tr>
<tr>
<td>6.4</td>
<td>Number of Authors Identified</td>
<td>117</td>
</tr>
<tr>
<td>6.5</td>
<td>Closer Examination of Results from the Grouping Task</td>
<td>119</td>
</tr>
<tr>
<td>6.6</td>
<td>Rating of Groups</td>
<td>122</td>
</tr>
<tr>
<td>6.7</td>
<td>Overall Impression</td>
<td>125</td>
</tr>
<tr>
<td>6.8</td>
<td>Summary</td>
<td>129</td>
</tr>
<tr>
<td></td>
<td><strong>Chapter 7</strong></td>
<td><strong>Conclusions and Future Research</strong></td>
</tr>
<tr>
<td>7.1</td>
<td>Conclusions</td>
<td>132</td>
</tr>
<tr>
<td>7.2</td>
<td>Future Research</td>
<td>135</td>
</tr>
</tbody>
</table>
References ................................................................. 138

Appendix A  Core Questions for the Survey ....................... 147

Appendix B  Visualizations Viewed by All the Subjects .......... 149
List of Tables

1. Social tagging vs. manual indexing .................................................. 8

3.1 Schema for the table {author} ......................................................... 35
3.2 Schema for the table {book} .......................................................... 35
3.3 Schema for the table {tag} ............................................................. 36
3.4 Schema for the table {author tag} .................................................. 36
3.5 Schema for the table {book tag} .................................................... 37
3.6 10 most frequent tags in librarything data set .................................. 50
3.7 Configuration of stimuli ............................................................... 62

4. Small-world property of network of authors in librarything ................. 78

5. Characteristics of networks generated using different term weighting schemes (C: clustering coefficient, l: average path length) ................. 91
5.2 Characteristics of networks generated using different term weighting schemes ......................................................... 94
5.3 Characteristics of networks generated using different similarity calculation (C: clustering coefficient, l: average path length) .................. 95
5.4 Characteristics of networks generated using different network construction (C: clustering coefficient, l: average path length) .................. 96
5.5 Characteristics of KNN network and ϵ-neighborhood network (C: clustering coefficient, l: average path length) .................................. 97
5.6 Characteristics of networks generated using the “classic” author data set (C: clustering coefficient, l: average path length) .................. 100
5.7 Output of the community detection algorithm on “classic” authors .......... 101
5.8 Characteristics of networks generated using the “fiction” author data set (C: clustering coefficient, l: average path length) .................. 109
5.9 Output of the community detection algorithm on “fiction” authors ........ 110

6. Rating of familiarity with literature .................................................. 113
6.2 Configuration of apparatus ............................................................ 114
6.3 Number of groups identified .......................................................... 116
6.4 Number of groups identified (ϵ-neighborhood) ................................ 116
6.5 Number of groups identified (KNN) ................................................ 116
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.6</td>
<td>Two-way ANOVA with repeated measures. A: network construction method</td>
<td>117</td>
</tr>
<tr>
<td></td>
<td>((\varepsilon\text{-neighborhood vs KNN}), \text{B: visualization method (cloud \text{vs network})})</td>
<td></td>
</tr>
<tr>
<td>6.7</td>
<td>Number of authors identified</td>
<td>118</td>
</tr>
<tr>
<td>6.8</td>
<td>Number of authors identified ((\varepsilon\text{-neighborhood}))</td>
<td>118</td>
</tr>
<tr>
<td>6.9</td>
<td>Number of authors identified (\text{(KNN)})</td>
<td>118</td>
</tr>
<tr>
<td>6.10</td>
<td>Two-way ANOVA with repeated measures. A: network construction method</td>
<td>119</td>
</tr>
<tr>
<td></td>
<td>((\varepsilon\text{-neighborhood vs KNN}), \text{B: visualization method (cloud \text{vs network})})</td>
<td></td>
</tr>
<tr>
<td>6.11</td>
<td>Result from cloud 1 (participant #10)</td>
<td>120</td>
</tr>
<tr>
<td>6.12</td>
<td>Result from network 1 (participant #10)</td>
<td>120</td>
</tr>
<tr>
<td>6.13</td>
<td>Result from cloud 2 (participant #9)</td>
<td>121</td>
</tr>
<tr>
<td>6.14</td>
<td>Result from network 2 (participant #9)</td>
<td>121</td>
</tr>
<tr>
<td>6.15</td>
<td>Participants’ rating of the groups generated from community detection</td>
<td>122</td>
</tr>
</tbody>
</table>
# List of Figures

1.1 Brentano’s tree of Aristotle’s categories ........................................... 3
1.2 Tag cloud snapshot in librarything .................................................... 9

3.1 A bipartite graph of tags and books in librarything .............................. 48
3.2 A unipartite graph of tags in librarything ........................................... 48
3.3 Interface of Prefuse and Wordle ....................................................... 68
3.4 Visualizations of top authors tagged as ‘classic’ in librarything ............ 69
3.5 Visualization of top authors tagged as ‘classic’ in librarything based on  
   second-order co-occurrence and KNN method ...................................... 70

4.1 Distribution of tags over items in librarything .................................... 73
4.2 Distribution of tags over authors in librarything .................................... 74
4.3 Number of items vs rank in librarything ........................................... 75
4.4 Occurrence vs rank in librarything .................................................... 76
4.5 Number of authors vs rank in librarything ........................................ 77
4.6 Occurrence vs rank in librarything .................................................... 78
4.7 Occurrence vs rank for individual books in librarything ....................... 79
4.8 Occurrence vs rank for individual authors in librarything ................. 80
4.9 Distribution of path lengths in author network in librarything ............ 81
4.10 Distribution of similarity matrix elements in author network in librarything 82
4.11 Dependence of the clustering coefficient on the filtering coefficient ... 83
4.12 Dependence of the average degree on the filtering coefficient .............. 84
4.13 Dependence of the average path length on the filtering coefficient ...... 85
4.14 Dependence of path length distribution on the filtering coefficient (I) 86
4.15 Dependence of path length distribution on the filtering coefficient (II) 87
4.16 Degree distribution of author network in librarything ....................... 88
4.17 Distribution of links across nodes in author network ........................... 88

5.1 Dependence of the modularity value on the filtering coefficient ........... 92
5.2 Distribution of similarity matrix elements in author network in librarything 93
5.3 Standard TF weight ............................................................................ 94
5.4 Fully weighted TF-IDF ......................................................................... 94
5.5 KNN graph ......................................................................................... 97
5.6 ϵ-neighborhood graph ........................................................................ 97
5.7 The “thriller writer” community in the network visualization ............ 102
List of Symbols

$\epsilon$  threshold of edge weights used in $\epsilon$-neighborhood method
$k$  degree of node in a network
$n$  number of nodes in a network
$m$  number of edges in a network
$l$  average path length of a network
$Q$  modularity of a network
$C$  (average) clustering coefficient
Chapter 1

Introduction

Thirty years of development notwithstanding, current online library catalog systems still suffer from the problem that they do not incorporate sufficient understanding of information seeking tasks other than known-item search, for example, exploratory information seeking and knowledge discovery. For an average user with only a vague conception as to what is to be retrieved from the catalog, current systems offer little help.

The social tagging paradigm is widely considered an extension beyond keyword-based indexing and hierarchical classification schemes. The new massive manual indexing method characterized by social tagging differs from automatic indexing that lays the foundation of modern information retrieval in that its manual nature obviates the common pitfalls of computer-based automatic indexing. It also complements traditional manual indexing since tag word distribution reflects the opinions of a large number of people with various background and knowledge instead of domain experts who are dominant in the classification and cataloging undertakings.

The social aspect of tagging offers the possibility of making semantic connections between concepts and entities within a library collection, document collection or photo collection. With the aid of co-occurrence analysis, small-world network model, community detection methods and other statistical analysis, this study is intended to concentrate on semantic relations between entities in the social tagging universe and demonstrate how analysis and visualization of these relations can facilitate exploratory information-seeking tasks and knowledge discovery.
1.1 Background

1.1.1 Hierarchical Classification Schemes

A choice of ontological categories has long been deemed the key to constructing a knowledge base ever since Aristotle’s Organon that enumerates all the possible kinds of thing which can be the subject or the predicate of a proposition.

The philosophical insights that helped the evolution of hierarchical classification structures of things can be found in the history of philosophy in both east and west. Lao-Tzu in ancient China states that

The Tao gave birth to the One;
The One gave birth to the Two;
The Two gave birth to the Three;
And the Three gave birth to the ten thousand things.

while Aristotle, considering the physical world to be the ultimate reality and treating *forms* proposed by Plato as abstractions derived from sensory experience, presented ten basic categories for classifying anything that may be predicated about anything, which were organized by Franz Brentano as the leaves of a single tree, as illustrated in figure 1.1.

Albeit abstruse, these ideas of categorization represent the philosophical thoughts behind the hierarchical taxonomy and classification schemes extensively used in science, technology and social life.

The original classification scheme of the Library of Congress, used between 1800 and 1814, was based on the philosophical works of Sir Francis Bacon. From 1814, the influence of Thomas Jefferson, who reclassified the library, can be seen on the Library of Congress collection [1]. Nowadays, in the field of library and information science, the term “classification” involves the orderly and systematic assignment of each entity to one and only one class within a system of mutually exclusive and nonoverlapping classes [2].
1.1.2 OPAC and Exploratory Information Seeking

In the spirit of the far-reaching top-down hierarchical schema, the hierarchical classification system and categorization-based indexing mechanism have dominated the practice of the online library catalog as the gateway to libraries which represent the most used and studied knowledge base.

But how do they perform in fulfilling the diversified information needs of average users? Imagine a possible scenario that is not infrequent in information-seeking practice but could demonstrate what is missing from current online library catalogs:

Jane is a graduate student in the Department of Civil Engineering. In a Computer-Aided Design course, she is required to develop a program based on AutoCAD. With little previous C++ programming experience, she will have to learn basic programming in C++ in a relatively short period of time. She needs some books on C++ that are for a layman and easy to follow. To browse through subject heading by “C++” can be awful drudgery and, perhaps quite confusing and misleading. Instead, on the homepage of the online catalog she clicks the “browse best-seller books on the Web” button, arriving at another page. In that page, she inputs “C++” in a text box (she can as well browse through a subject hierarchy like “Computers & Internet...
– Programming – Languages & Tools – C & C++ – Programming”, which can be more time-consuming), and then clicks the “browse” button. The result page is just like that of Google, showing a list of tens of online book sellers that sell books on that specific subject. She chooses “Amazon.com”.

There seems to be a long list of C++ books. Amazon provides very detailed editor reviews and reader reviews, from which Jane learns that *The C++ Programming Language* was written by the creator of C++, but it is more like a text book for a computer science major. *Effective C++: 50 Ways to Improve Your Programs and Design* is great, too. But it serves the needs of those who have hands-on C++ programming experience and want to improve even more. By tracking the link “Customers who bought this book also bought”, reading the reviews and iteratively repeating the process, she eventually gets an idea of the books that are more suitable for her: *Accelerated C++: Practical Programming by Example* for step-by-step learning from example programs, and *Thinking in C++, Volume 1: Introduction to Standard C++ (2nd Edition)* for deep and detailed reference whenever necessary.

Coming to the online catalog without a specific title in mind is a more difficult task, especially for those who are not very familiar with the subject matter. Current online catalogs are of little, if any, help to this kind of reader. There are 371 entries under the heading “C++ (Computer program language)” in the University of Illinois online catalog. A list of 371 catalog entries with simple catalog metadata probably will not make much sense to a layman. As reported by Micco, many users are just overwhelmed by the size of the result set of a single catalog search [3], not to mention it is only a list with little cue about semantic relevance.

At this point the requirements of users go beyond the capability of a query-answering machine. To aid users in these scenarios, certain kinds of semantic links should be used to bridge the gap between separated items. The borrowing of successful business practice could remarkably move the online catalog along the way from a query-answering machine to a question-answering machine, which is a great progress since it encourages users to express
and reiterate their unstructured or semi-structured information-seeking instead of merely inputting structured query expressions.

Borgman summarizes the problems with online library catalogs, which are typical IR systems, in two articles, with the second published ten years after the first \[4, 5\]. She particularly points out that the design of online catalogs does not incorporate sufficient understanding of searching behavior, and the basic functionality of online catalogs has changed little since the late 1980s \[5\]. Pejtersen, among others, also suggests that OPAC design should not be solely focused on known item and subject searches as the two types of goals \[6\].

As envisioned by Bush in 1945, when he imagined a future personal device “memex”, the important thing is the process of associative indexing, the basic idea of which is a provision whereby any item may be caused at will to select immediately and automatically another \[7\].

The emergence of social tagging provides a new way of making associations between users, items, creators and tags, and also a new paradigm in which Bush’s expectation of associative indexing might be realized. As argued by Chen and Paul in a knowledge visualization study, the integration of citation and co-citation patterns provides a rich, ecological representation of a knowledge domain \[8\]. Patterns found in tag relations are expected to play a similar role.

\textbf{1.1.3 The Social Tagging Phenomenon}

Tagging using uncontrolled keywords as a different indexing scheme has been gaining increasing popularity on the web for the past few years. A December 2006 survey has found that 28\% of internet users have tagged or categorized content online such as photos, news stories or blog posts \[9\]. Social tagging, also referred to as collaborative tagging or simply tagging, is the practice and method of collaboratively creating and managing tags to annotate and categorize content and objects (anything with a URL) for search, browse and retrieval. The tagging process is performed in a shared and open environment. The act of tagging is
done by the person consuming the information [10]. The collective effect of users tagging items is that tags are not only for the taggers’ own retrieval, but also for search, browse and recommendation for everybody else on the same site, as demonstrated by Manish Gupta et al’s list of tagging motivations [11]:

- future retrieval
- contribution and sharing
- attract attention
- play and competition
- self presentation
- opinion expression
- task organization
- social signalling
- money
- technological ease

Some saw tagging as a popular form of manual indexing on the web that extends beyond the scope of traditional manual classification [12]. For example, on the music tagging site last.fm\footnote{http://www.last.fm} the Chilean classical pianist Claudio Arrau is tagged as “chile”, “classica”, “classic”, “classical”, “classical pianist”, “classique”, “piano”, etc., with “classical” being the most frequently used tag. Another example is the scientific work Small worlds: the dynamics of networks between order and randomness by Duncan J. Watts. On the book
it is tagged “complexity”, “networks”, “science”, “social network analysis”, “mathematics”, “graphs”, with “networks” being used the most. These tags arguably convey more information than the subject terms associated with its Library of Congress Classification number: QA166.W38, which indicates “Science - Mathematics - Graph Theory - General Works”. Noll and Meinel argue that, for HTML documents, tags provide additional information not directly contained within a document and thus help improve or augment traditional document classification and retrieval techniques, especially for broad classification of documents [13].

Another reason social tagging gained popularity within a relatively short period of time lies in the interplay between the personal and social aspects of social tagging. A user of dogear, a social bookmarking service experimented with in the intranet of IBM, reports that there are a number of personal benefits for using bookmark tagging so that even if the social aspect of tagging is ignored, it still yields a lot of benefit to users. This user refers to the added convenience of social bookmarking in managing and recalling sites marked through dogear from anywhere on the internet, compared to the traditional bookmarking restricted to a particular machine [14].

Also important is the economic factor behind social tagging. A large amount of tagging data are literally free to generate and easily accessible to users online given an established community of taggers. In contrast creating formal classification entries and index terms for books are relatively expensive and the results may not be convenient for online access.

Table 1.1 shows how social tagging differs from manual indexing in several aspects.

The connectionist nature of social tagging is reminiscent of what was envisioned by Bush about “memex”:

*The owner of the memex, let us say, is interested in the origin and properties of the bow and arrow. Specifically he is studying why the short Turkish bow was apparently superior to the English long bow in the skirmishes of the Crusades. He has dozens of possibly pertinent

²http://www.librarything.com
books and articles in his memex. First he runs through an encyclopedia, finds an interesting but sketchy article, leaves it projected. Next, in a history, he finds another pertinent item, and ties the two together. Thus he goes, building a trail of many items. Occasionally he inserts a comment of his own, either linking it into the main trail or joining it by a side trail to a particular item. When it becomes evident that the elastic properties of available materials had a great deal to do with the bow, he branches off on a side trail which takes him through textbooks on elasticity and tables of physical constants. He inserts a page of longhand analysis of his own. Thus he builds a trail of his interest through the maze of materials available to him [7].

The process of building semantic trails in Bush’s personal document repository system could be compared to an active user annotating a significant number of relevant resources on a social tagging site, forming a main trail and some side trails connected by tags.

The abundance of public tagging data not only offers the possibility of conducting extensive correlation analysis of tags, items and users, but also suggests an angle to detect the dynamics and trends in different communities and structures of users, items and creators.
The quantitative information implied in tagging data also gave rise to a new paradigm for browsing a collection, through a simple but interesting visual construct called "tag cloud".

### 1.1.4 Tag Clouds as Interface to the Tagging World

**Tag clouds** (or tagclouds) are visual presentations of a set of tag words that are the most popular in the site or associated with a particular portion of the items, in which text attributes like font size, weight or color represent features like frequency of the tags. It has been a visually arresting construct found in almost all the mainstream tagging sites including del.icio.us, librarything and flickr. Practically tag clouds may serve different purposes such as search, browse, impression formation or even as part of the portal page.

The tag cloud translates the diversified vocabulary of folksonomy into a social navigation tool in Dieberger et al.’s definition, i.e., a system whereby user navigation is guided and informed by information about what other people are doing [15].

A portion of the tag cloud on librarything is shown in figure 1.2.

![Figure 1.2: Tag cloud snapshot in librarything](image-url)
The effectiveness of tag clouds for various information-seeking tasks and user satisfaction has been argued and evaluated. What most researchers agree upon is the effect of tag cloud on high-level tasks such as impression formation and user satisfaction. Rivadeneira et al., for example, conducted two studies to evaluate the effectiveness of differently constructed tag clouds for different tasks they can be used to support. The effect of font size was found to be robust; there was no effect of layout on recognition and recollection; there was a moderate but statistically significant effect of layout on high-level processes such as impression formation [16]. Kuo et al.’s studies also indicate that a tag cloud performs better than a list of search results in presenting descriptive information and in reducing user frustration [17]. In a more recent study with 89 participants, Sinclair and Cardew-Hall concluded that where the information-seeking task required specific information, participants preferred the search interface; conversely, where the information-seeking task was more general, participants preferred the tag cloud. Also, a tag cloud is considered to provide a visual summary of the collection that requires less cognitive load than formulating specific query terms [18]. All these studies, and the increased popularity of tag clouds in public tagging sites, suggest that compared to a search interface, a tag cloud, as a visually arresting and cognitively easier web page construct, is particularly useful for browsing, non-specific information summary and impression formation.

The inferiority of tag clouds has also been argued and possible improvements proposed in different studies. Halvey and Keane, for example, suggest that alphabetization and varying font sizes can aid users to find information more easily and quickly [19]. Hassan-Montero and Herrero-Solana, on the other hand, argue that alphabetical arrangements of displayed tags neither facilitate visual scanning nor enable the inference of semantic relations between tags, proposing a tag cloud layout based on clustering of tag words whereby semantic density of tags is reduced while visual consistency is improved [20]. Hearst and Rosner also state that tag clouds are inferior to a more standard alphabetical listing, suggesting that the main value of tag cloud visualization is as a signal or marker of individual or social interaction.
with the contents of an information collection, and functions more as a suggestive device than as a precise depiction of the underlying phenomenon [21].

Particularly relevant to this research is the finding that relations between tags are less pronounced in current tag cloud layout. Kaser and Lemire’s studies indicate that the typical layout of tag clouds does not account for relationships that may be known between tags [22]. Kuo et al. also point out, after a study comparing the PubCloud tag cloud summarization of query results with the standard result list provided by PubMed, that the tag cloud interface is advantageous in presenting descriptive information and in reducing user frustration, but less effective at the task of enabling the user to discover relations between concepts [17]. Hearst and Rosner also note that the problem with the typical tag cloud layout is that items with similar meaning may lie far apart, and so meaningful associations may be missed [21]. None of these studies, however, suggest any new method or visualization scheme where associations between tags can be retained or better presented.

Given empirical studies about the effectiveness of tag clouds and applications of information visualization in information retrieval, the problem of missing relations between concepts in a tag cloud might be alleviated by a graphic paradigm whereby tags are clustered based on their relations while the functionality of giving impression or ‘big picture’ is retained. A recent study in that direction is Lohmann et al’s comparative analysis of three different tag cloud layouts, in which thematically clustered tag cloud layout has proved to perform best in the task that involves finding tags that belong to a given topic [23]. This study takes a step further and explores the possibility of using network paradigm to connect related tags in addition to geographic proximity.

1.2 Purpose of the Study

Hodge groups different knowledge organization systems (KOS) into three general categories that are distinguished by an increasing degree of language control and growing strength of
semantic structure [24]:

- term lists, which emphasize lists of terms often with definitions. Examples are glossaries, dictionaries and gazetteers.

- classification and categories, which emphasize the creation of subject sets. Examples include the Medical Subject Headings (MeSH) and the Dewey Decimal Classification.

- relationship lists, which emphasize the connections between terms and concepts. Examples are thesauri, semantic nets and ontologies.

A tag cloud is essentially a term list, albeit a more attractive one, because it does not offer any strengthened semantic structure. But with relations between tags or authors and visual cues of meaningful network structures, networks of tags generated from tagging analysis definitely fall into the third category. This suggests a direction to address the issue of “presenting relations” that was proposed yet unresolved by previous user studies on tag clouds [17, 22].

The purpose of the study is to conduct co-occurrence analysis and social network analysis on tags and authors in the social cataloging site librarything, identify clustering structures in networks of tags, items and authors, employ visualization techniques to visually present the relations between concepts and entities and the clustering structures formed by nodes in the graph, and provide guidelines for interface designers of tagging sites and online library catalogs as to how tag analysis offers a new angle for generating relationship groups as semantically strengthened structure that facilitates information-seeking tasks such as topic summary, impression formation and knowledge discovery. In Hodge’s term, this helps transform tag or category list into a semantically rich knowledge organization system where connections between concepts and entities are emphasized.

To further exemplify the semantic problem with social tagging and tag clouds, it would be helpful to make comparison to wiki as another important web 2.0 phenomenon.
Wiki and social tagging are both web 2.0 phenomena characterized by content creation through collaborative efforts of a large number of users varying in background, expertise and geographic area. The “wiki” technology was invented by Ward Cunningham in 1995 [25]. In a wiki, each page has an “edit” link through which users could have access to an editing view of the content of the page, thereby allowing anybody to make and submit changes to the content. Therefore, there are often multiple authors creating and editing the same page. The largest wiki on the publicly-accessible web, Wikipedia, has received systematic study, both quantitative and qualitative, for the past few years. The Wikipedia Networks Team from Zagreb, Croatia examines a number of statistical network characteristics of Wikipedia and presents evidence of a possible common growth process [26]. Capocci et al.’s analysis focuses more on the dynamics of Wikipedia, suggesting that its growth can be adequately described by preferential attachment [27]. Korfiatis et al. study the structure of co-authorship networks and visualize a network instance, in addition to analyzing the statistical characteristics of articles and the networks [28]. Holloway et al. take a step further as their team identifies and visualizes the semantic structure and the age of the categories in this free online encyclopedia, and the content coverage of the authors [29]. These studies attempt to analyze Wikipedia as a complex network, which is frequently found in social networks in the physical world and the web alike.

All these studies are made possible by the unified structure and mechanism of Wikipedia. Most articles on Wikipedia are assigned one or more categories, rendering the topic categorization much easier than performed on social tagging sites. The distinction between social tagging sites on different kinds of items like books, photos and music, however, complicates any systematic and unified attempt to study the semantic structure and dynamics of mainstream social tagging sites.

Previous studies on social tagging provide interesting results on the distributions of frequencies of tags, users and items and their evolution over time. Lin et al., for example,
analyze the convergence nature of tagging on flickr[^flickr], a popular social photo website, suggesting that aggregated frequencies of 30% of tags account for 70% of total occurrences of tags, seemingly at odds with the widely accepted ratio of 80/20 in bibliometrics study[^30]. Halpin et al. also present the power law distribution of tag occurrences in del.icio.us[^4], proving that the logarithm of relative position of a tag is inversely proportional to the logarithm of the number of times the tag is used[^31]. The same distribution is also observed by Voss on flickr, del.icio.us as well as Wikipedia[^32].

Much less has been found in terms of extraction of structures in social tagging that are helpful in performing information-seeking tasks, such as the semantic trails Bush had expected of “memex” that can help the recollection and sharing of previously stored documents. A rough hierarchical structure of tags is reportedly found in audioscrobbler[^5], a music tagging site, and citeulike[^6], a collaborative bibliographic site[^33]. Mika also visualizes the network-based semantic structure found in del.icio.us[^34].

These early studies, albeit small in scale, demonstrate the seemingly possible convergence of network-based social tagging and tree-based classification schemes, as Halpin et al. conclude after the identification of semantic structure of social tagging in a preliminary study:

> It seems quite plausible that folksonomies and ontologies, which are merely new incarnations of the age-old distinction between categorization and classification respectively, are not mortal enemies, but fundamentally compatible, as tagging-based categorization in our data exhibits emergent consensus.[^31]

Other researchers attempt to split frequently used tags into several semantically independent groups based on their co-occurrence relationships[^35]. This approach is technically

[^flickr]: http://flickr.com/
[^4]: http://del.icio.us
[^5]: http://www.audioscrobbler.com
[^6]: http://www.citeulike.org
straightforward as it is a special case of the traditional graph clustering problem [36]. Despite its categorizing nature, clustering has been a well-studied leverage for analyzing structural and statistical characteristics of a graph or a network.

The application of various clustering algorithms to the tag network is validated by the fact that networks (co-occurrence graphs) of tags or authors are instances of small-world and scale-free networks, characterized by a large clustering coefficient, small average path length and power-law distribution of node degrees, compared to random graphs with the same number of vertices and edges, as will be proved in this study. Therefore, as hypothesized here, there exist clusters of tags and items, and communities of authors and users, as there are groups and even cliques of people in a social network. These clusters of tags are expected to convey semantic information about tags as people form groups or even cliques by their age, interest, occupation, religion, and so on.

The result of tag analysis should be presented to users in the forms that are more diverse and semantically richer than online catalogs and even tag clouds. Fluit et al. suggest that visualization might be a solution to the not infrequent “zero-match” problem, whereby either a “no match” message or a long list of partial matches is shown. The user gets neither a clear overview of the results nor suggestions for further exploration [37]. He also demonstrates how a visualized system interface can show beautiful and instructive result sets. Also, Chen and Paul successfully applied a link-reduction method to the analysis of co-citation of literature in several fields, proposing a concept named “knowledge landscape”, whereby knowledge in a specific domain could be represented in a 3-dimensional space as objects and links that seem like a landscape [8]. There are some other similar studies that present the landscape of authors, documents or concepts in a knowledge domain through different visualization techniques [38, 39]. These early information and knowledge visualization experiments demonstrate how visualization could possibly help reveal the “big picture” of the underlying collection. But none of them evaluate the power of visualization in facilitating high-level information-seeking tasks by any kind of user studies other than anecdotal evidence. Social
tagging provides a data source beyond author co-citation and word co-occurrence for this kind of visualized knowledge mining.

Given previous research on information retrieval, online catalog design, small-world network model and community detection, this study is conducted in an effort to systematically analyze tag relations in the social cataloging site librarything, the web 2.0 equivalent of a traditional library catalog and a largely under-studied subject in current social tagging literature, thereby providing insight into how tag analysis could contribute to the practice of high-level information seeking. The structure, as opposed to dynamics, of the social tagging universe will be the predominant subject.

The proposed study is intended to fill the niches in social tagging and information visualization research noted above, or specifically,

- systematic study of the co-occurrence and clustering patterns of tags and items on a social tagging site, particularly librarything;
- the proposed but unresolved issue with “presenting relations”, and presenting “the big picture” in general, associated with tag cloud visualizations;
- user study of the effect of network visualization in facilitating high-level information-seeking tasks in the context of social tagging.

The results from this study may benefit the following aspects of research and practice on social tagging:

- Proven methodology and heuristics on similarity analysis and network construction may shed light on the analysis of tag networks or other kinds of networks derived from two-mode networks in general.
- Tag networks, if properly created and evaluated, may represent an alternative way to visualize and present information in the tagging universe that emphasizes relations and helps guide navigation.
- Ideas from tag network visualization can also be borrowed by tag cloud designers so that clues on groups and relationships are reflected in tag cloud layout even though the lines between the words (as in a network) are not present, as in the thematically clustered tag cloud layout \[23\].

1.3 Research Questions

The central question that is planned to be answered in this study is: how effective are visual network constructions based on tagging analysis in answering the basic question in an exploratory IR setting: “What is out there?”, in comparison to tag clouds?

With the focus on the collected librarything data set, the specific questions to be answered in this study are

- What are the connectivity, degree distribution and modularity of networks of tags and authors generated from co-occurrence analysis?

- How will different term weighting schemes, network construction methods and filtering coefficients (defined in detail in section 3.3) affect the parameters of the resulting similarity network and also its community structure?

- Compared to tag clouds, could network visualizations based on tagging analysis facilitate information-seeking tasks, particularly topic summary, impression formation and knowledge discovery, of typical users?

On the other hand, this study is NOT intended to answer questions like

- Are there any community structures among taggers for a particular social tagging site? How do these structures evolve over time?

- What are the dynamics in tag evolution over time? Does usage of tags change dramatically over time or converge toward some relatively stable status?
• How does information retrieval based on tags differ from traditional information retrieval performed on full texts?

To more clearly define the investigations that are to be conducted in this study, the following hypotheses are formalized:

• Tag occurrences in the librarything data follow power law distribution, and the one-mode networks resulting from similarity analysis are small-world networks.

• One-mode author networks contain strong community structures that are algorithmically detectable and semantically relevant.

• Users will be more successful at identifying related groups of authors when viewing network visualizations than when viewing cloud visualizations.

Later sections will be framed around testing these hypotheses using both quantitative and qualitative methods.

The remainder of this dissertation is organized as follows: chapter 2 offers an overview of previous studies on knowledge visualization, social network analysis, network visualization and tagging analysis; chapter 3 delineates research design and methods being used to answer the aforementioned research questions. Chapter 4 covers the structural properties of tag-author networks and the small-world property of projected one-mode networks that lays the foundation for community analysis. In chapter 5 community detection methods are applied to one-mode author networks, and the results are evaluated and interpreted. Chapter 6 presents the results obtained from the user study and chapter 7 concludes the dissertation.
Chapter 2

Literature Review

2.1 Keyword-based Indexing and Hierarchical Classification

Modern information retrieval research characterized by the development of automatic indexing methods has long been in favor of keyword-based searching as opposed to subject-based searching and browsing. As Schneider has pointed out, keyword-based indexing eventually became dominant partly because detailed classifications are more difficult to develop [40]. Both the vector space model and the probabilistic model have been based on queries that are formulated as a set of keywords. Keyword-based search also dominates current Web search engines.

From a user-oriented perspective, however, the result of information-seeking is better presented to users in forms that are more diverse and semantically richer than the result set of most current IR applications such as search engines, online library catalogs and bibliographic databases.

Despite emphasizing the dominance of keyword-based indexing methods, Schneider makes an insightful comparison of keyword-based and classification-based indexing and details some advantages of classification-based indexing: because classifications specify the generic-specific relations between concepts, they facilitate indexing at any desired level of generality or specificity; in a classification system a high degree of precision in indexing or matching ideas is possible because each concept is clearly expressed in natural language; in classifi-
cations any given category or category number represents the same concept, irrespective of synonyms, abbreviations, grammatical variants, etc. These characteristics of classifications leave open the possibility that classification-based browsing, and indexing also, might be preferable to keyword-based indexing in an IR context.

Borgman has pointed out that years after the emergence of online library catalogs, the basic functionality has changed little and design still has not reflected studies of searching behavior [4, 5]. Geller and Lesk propose that perhaps the most efficient way to do a subject search is to start with a keyword search to locate the correct category and then browse through the classification [41], suggesting the incorporation of classification-based browsing in an IR environment. Fluit et al. also suggest that visualization might be a solution to the not infrequent “zero-match” problem, whereby either a “no match” message or a long list of partial matches is shown. The user gets neither a clear overview of the results when the result is overloaded, nor suggestions for further exploration if there is no match at all [37].

2.2 Information Seeking and Visualization

Browsing seems less attractive without a visualized or graphic environment. The recent advancement in graphic display devices and visualization techniques has been very helpful in this regard. The young and still developing information visualization research community has paid particular attention to the problem of document and knowledge visualization. These visualization studies roughly fall into three categories, even though their underlying technologies might be similar or interchangeable: the connectionist approach to visualizing the interrelationships between documents and/or terms in a connected graph; the projection of a whole collection of documents from high-dimensional feature space to two-dimensional space; and 2-D or 3-D visualization of literature by co-citation analysis.

Doyle was among the first to propose a graphical browsing environment for IR where vertices represent terms and edges semantic relationships [42]. Croft et al. extended this
idea by adding vertices representing documents, generating a network organization consisting of document nodes, term nodes and weighted edges connecting them [43]. The topology of these networks may not necessarily follow a classification hierarchy, but the visualization paradigm undoubtedly provides space for user interaction.

Observing the difficulty of practical IR systems featuring keyword-based search and Boolean logic expression, where the interaction between users and the systems is largely limited and users have little control over the amount of output returned by a query, Godin and colleagues are among the early practitioners to adopt a browsing interface for IR [44, 45]. Employing a lattice structure extracted from the term-document relationship in a collection, their proposed interface permits gradual broadening or narrowing of the user’s query by browsing through a graph of term and document subsets. Each vertex in the graph represents a query formed by a combination of terms with retrieved documents. One can reformulate a query by following edges in the graph. Here the reformulation task is performed in an interactive and user-friendly fashion, enabling the user to better target relevant documents based on existing output.

Representative of the second category of text visualization is Lin’s self-organizing map for information retrieval. He reported an application of an unsupervised learning method, Kohonen’s feature map algorithm [46], to the construction of self-organizing semantic maps for information retrieval [47, 48]. The feature map algorithm maps documents and queries to a two-dimensional space, creating a visual effect through geographic features of the areas on the map. The semantic map also preserves certain clustering and hierarchical structures of the document space, presenting to the user a clear picture of how documents in a collection are distributed and correlative. Analogous to the self-organizing map is the ET-map developed by Chen and his colleagues. They employed Kohonen’s self-organizing map for the Yahoo! website to categorize Internet homepages according to their content and create a tree-map-based classification hierarchy [49]. Usability study demonstrates that the visual elements, novelty, multi-layer browsing and navigation mechanism brought by the map are
exciting, although sometimes users find the map confusing, unable to generalize different levels of abstraction and not useful for searching [50].

Inspired by Lin’s semantic mapping of document space was Chalmers’s work using numerical techniques for multi-dimensional scaling, which is a widely used dimension reduction method for scientific data analysis, to represent articles in a bibliography as particles in three-dimensional space. A spring-based layout algorithm tends to make similar articles closer to one another and dissimilar ones more distant from each other [51]. Chalmers’s prototype system enabled graphically-based exploration of a complex information space, providing users with a natural and connectionist model of document relationships. And the interactive visualization and navigation of the information space becomes an attractive device for browsing and exploring the corpus. The performance of the two-dimensional document layout algorithm was later improved by applying stochastic techniques [52].

Wise et al.’s work on document landscape is widely cited as the pioneer of three-dimensional visualization of free text. Moving from a metaphor of points in space to one of a landscape, their visual representation depicts theme density as elevation of mountains, while valleys, peaks, cliffs and ranges reflect intricate interrelationships among documents and their composite themes, offering a clear visual topical summary of the whole collection. Information analysts reported enhanced insight and time savings such as ‘discovering in 35 minutes what would have taken two weeks otherwise.’ The three-dimensional layout algorithm starts by clustering the document collection in the high-dimensional feature space. Multi-dimensional scaling is applied to the centroids of clusters to reduce the computation cost [53].

Lin, Chen, Chalmers and Wise’s approach differs from the previously mentioned term network or term-document network by using a dimension reduction technique to map the high-dimensional document space to a two-dimensional space and thereby present a big picture of the whole collection and the interrelationship between documents, instead of constructing a network of documents and terms, which does not usually involve dimension
Another province where document visualization thrives, which is not necessarily in IR per se but closely related, is author co-citation and document co-citation analysis. Citation analysis examines the frequency and pattern of citations in articles and books. It is considered one of the most commonly used methods in bibliometrics, which studies or measures texts and information of publications. While citation analysis is most often used in the field of library and information science, instead of social network analysis, its results are often presented as a social network of authors, which bears resemblance to other forms of social network.

Author co-citation analysis is a special kind of citation analysis that focuses on intellectual connections between authors as reflected in citations of scientific publications. Any two authors are linked together if other authors often reference their work together. It is a special case of one-mode analysis in studying a bipartite graph as a representation of affiliation networks whereby authors under study are actors while being referenced in the same article is a special kind of event.

A relatively thorough study of a scientific knowledge domain in a co-citation analysis focuses on information science itself, where White and McCain present an extensive domain analysis of information science in terms of its authors. Names of those most frequently cited in 12 key journals from 1972 through 1995 were studied, yielding automatic classification relevant to the histories of the field [38]. Small presents a two-dimensional graph visualization of the scientific literature based on journal co-citation patterns derived from ISI citation databases. His co-citation clustering method using triangulation [54] produces a unified hierarchical arrangement of documents and thus creates a nested mapping [39]. Chen went a step further to extend and transform traditional author co-citation analysis into a knowledge-visualization and domain-analysis tool where the Pathfinder network scaling technique [55] is successfully applied to the analysis of author co-citation of literature in several fields. The intellectual groupings determined by factor analysis are visualized through a so-called

23
“knowledge landscape” [56, 8]. This landscape highlights predominant research areas as well as authors in the field during the period of study. Not dissimilar to Wise’s theme landscape, this knowledge landscape provides information seekers with an additional means of exploring and browsing the scientific literature by author and field.

The list of visualization-enriched IR systems will never end but it appears that traditional visualization-based searching and even browsing have not gone far beyond research environments where work on information visualization was originally started. The usability study of the ET-map on Yahoo! sheds light on the difficulty of average users to adapt to a browsing mechanism that is seemingly novel and academically convincing but not seamlessly integrated to the keyword-based searching interface which users have been increasingly getting used to. Practically, information retrieval approaches that take into account user’s input and relevance feedback, on which Google, Amazon and the social tagging community have been making endeavors, seem more promising.

In an effort to emphasize connections between concepts and explore the proposed issue with “presenting relations” inherent in tag clouds, this research study makes use of network visualizations in which authors/tags are connected by edges as the result of network analysis.

2.3 Social Network Analysis and Visualization

Social network analysis models social relationships in terms of nodes and ties. It has emerged as a key method in modern sociology, communication studies, organizational studies and information science.

With the advent of computer software to systematically analyze and visualize networks, social network analysis has been employed in different academic and industrial applications beyond its origin in studying social relationships of people, ranging from study of the power grid in a country to the link analysis of articles in the PubMed collection [57].

The field of social network analysis is one of the strongest proofs of Alfred Crosby’s
proposal that visualization and measurement are the only two factors that are responsible for the explosive development of all of modern science. And these two factors are indeed integrated in the field of social network analysis as Breiger pointed out that a particularly notable move from metaphor to analytical method in analysis of social networks is the relatively recent development of highly sophisticated computer programs for producing pictorial representations of social networks [58].

2.3.1 Affiliation Networks and One-Mode Networks

Particularly pertinent to this study is the kind of social networks referred to as affiliation network, two-mode network or bipartite network, which differ in several aspects from mostly studied social networks of people or entities. Affiliation networks or two-mode networks consist of a set of actors and a set of events [59]. The primary subject of this study is actually one-mode networks based on linkages established through the other mode in the two-mode network of authors and tags. This is the simplest approach to two-mode network data analysis, that is to convert it into two one-mode networks and examine them separately.

A typical example of two-mode networks is the network of papers and authors in a specific field. Given this paper-author network, a network of co-authorship can be derived using different projection methods. Lambiotte and Ausloos, for example, present a method to project a co-authorship network that accounts in detail for three-body interactions among scientists, thus demonstrating the importance of “high-order correction” in order to characterize the community structure of scientists [60]. The author co-citation networks, as studied by Chen [56], are another kind of two-mode network frequently found in the field of library and information science, where two authors are connected if they are both cited in the same paper.

Although there is a significant number of notions and methods to study one-mode networks both in traditional social network analysis and in the context of recently developed small-world network theory, a generalizable and widely accepted framework is still missing
when it comes to two-mode networks. Most studies involving two-mode networks simply convert or project them into two separate one-mode networks and apply existing methods in one-mode network analysis. Newman was among the first to study the statistics of the projected networks in the context of the underlying bipartite structure [61]. Latapy et al. present a systematic overview of two-mode network studies in the field of social science, computer science, linguistics and physics, as well as an alternative that extends basic notions for one-mode network to the analysis of two-mode networks [62].

The fact that tags are applied to authors or books makes the affiliation network model fit the social tagging data very well. As is the case for many previous works on various affiliation networks, projected one-mode author networks will be the primary subject of this study.

2.3.2 Small-World and Scale-Free Network

Studies of large-scale networks of various kinds have led to academic interest in the small-world phenomenon.

Based on the Erdos-Renyi model of random graph, Watts proposed the small-world network in a 1998 article in *Nature* [63] and later devoted an entire book to this topic [64].

The Erdos-Renyi model of random network gives estimates of two important parameters of a random graph [65]: the average path length,

\[ l_r \approx \frac{\ln N}{\ln k} \]

and the clustering coefficient

\[ C_r \approx \frac{k}{N} \]
where \( N \) is the number of vertices in the network and \( k \) the average degree of all the vertices.

*Average path length* of a network is a concept in network topology that is defined as the average number of steps along the shortest paths for all possible pairs of network nodes. *Clustering coefficient* of a graph is the average of the clustering coefficient for each vertex, which is defined as the ratio of the number of edges between neighboring vertices of the vertex to the number of all possible edges \[63\], or

\[
C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i - 1)} : v_j, v_k \in N_i, e_{ij} \in E
\]

where \( E \) is the edge set of the graph and

\[
N_i = \{v_j\} : e_{ij} \in E
\]

is the neighbor set of vertex \( v_i \), and

\[
k_i = |N_i|
\]

is the degree of the vertex \( v_i \), or the number of vertices in the neighbor set of \( v_i \).

In this study, the concept of average clustering coefficient will also be generalized to a community or group to differentiate groups yielded by community detection methods, as will be discussed in chapter \[5\].

A *small-world network* is characterized by a similar average path length but a considerably higher clustering coefficient compared to a random graph with the same number of edges and vertices, or
\[ l \simeq l_r, C \gg C_r \]

This definition implies that even though the vertices in a small-world network are no closer than they are in a random graph, they are indeed much more connected and form more clusters or even cliques than in a random graph. “Small-world” is literally misleading because the world is not really smaller than a “random world” - it takes almost the same number of steps to reach an unknown person (through a friend, a friend’s friend, etc.), but the world is more closely connected so that a friend’s friend is more likely to be a friend as well.

Despite its relatively recent establishment, small-world network theory has been applied to different fields because many real-world networks, including social networks, turn out to be small-world networks. In his book, Watts presents examples like the power grid in western states and the actor network derived from co-starring of actors in a movie [64].

The scale-free network was proposed on the basis that neither random-graph theory nor small-world network theory entails a power law degree distribution, especially for large \( k \) (degree), which is frequently observed in real world networks [66].

Mathematically, the defining characteristics of scale-free networks is that their degree distribution follows the power law:

\[ P(k) \sim k^{-\lambda} \]

The structure and dynamics of a scale-free network are independent of the number of nodes the network has. These features were found to be a consequence of two generic mechanisms:
1. networks expand continuously by the addition of new nodes;

2. new nodes attach preferentially to existing nodes that are already well connected.

The scale-free model based on these two assumptions is capable of explaining different aspects of real world networks, especially the degree distribution of nodes, which both random network and small-world network theory fail to address [67].

The scale-free feature of a network suggests that there are a marked number of hubs, like those in the airline network, that are highly connected while the majority of the nodes are relatively poorly connected. Although small-world networks are often found to be scale-free, a network can be a “small world” in the absence of hubs.

### 2.3.3 Social Network Visualization

Network visualization has become the cornerstone of the field of information visualization exemplified by an increasing number of visualization applications as a result of complex computations and unbounded imagination. Since the very beginning of social network analysis, it has been extensively used as a tool to develop structural insights and communicate networks to others in social network visualization. Rapid development in display technology and prevalence of web images render it possible to use color animation or even three-dimensional images to explore structural configurations of social networks. For network visualization based on nodes and edges, the force-directed graph drawing algorithm has proven an effective graph layout paradigm that generates topologically correct and aesthetically appealing graphs [68] and has been widely used in point and line displays in social network analysis, especially after the introduction of network drawing programs such as Krackplot, Pajek and NetVis in the 1990s [69].

Combined with analysis of appropriate statistics, network visualization can be an excellent tool for pattern and knowledge discovery. Perer and Shneiderman, for example, by increasing the threshold of strengths of edges being shown in the network of senators based on
co-occurrences of their votes, were able to discover that Democrats stay more tightly unified than the Republicans, which fails to be revealed by other visualization-centered programs like KrackPlot and ManyEyes [57]. This is an interesting example of proper pre-processing of networks based on graph topology or network statistics as the aid to effective knowledge discovery.

Contrary to a link reduction mechanism that retains edges with strengths above a threshold, topology-based algorithms aim to take the network as a whole and preserve the most salient semantic relations. Pathfinder network scaling, for instance, employs triangle inequality to eliminate counter-intuitive edges, but at the cost of significant computational and spatial requirements, making it hardly scalable for large-scale networks [70].

Observing that many small-world networks actually have a multiscale nature and can be viewed as a network of groups that are themselves small-world networks, Auber et al. propose a recursive method that decomposes a network into hierarchically clustered networks based on edge strength as a measure of its contribution to the cohesion of neighborhood [71]. Wu et al. also approach the problem of analyzing scale-free networks from a data reduction perspective, using graph geodesics (i.e., shortest paths) to cluster large-scale graphs for improved knowledge revelation and visualization [72].

In this study of networks generated from social tagging data, not only visualizations of networks provide an intuitive way to examine the semantic clues in the data, but also these visualizations are used as stimuli to prompt users’ feedback as to how the relations between concepts may be better presented.

## 2.4 Classification and Cluster Analysis

The past decades have seen the application of various established machine learning methods to a number of scientific fields. The irregular nature of natural text and its large availability make it very promising to use machine learning techniques in different aspects and forms of
document retrieval. Some of these methods might also be applied to the analysis of social tagging data.

The k-nearest neighbor (KNN) algorithm \cite{73}, a well-known statistical machine learning approach, has long been adopted in different classification tasks, including text categorization \cite{74, 75}. Despite its long history, it is one of the top performers among categorization algorithms, especially when the category distribution is extremely skewed \cite{76}. The disadvantage of KNN lies in its delayed intensive training at the time of classification. Yang’s example-based learning algorithm, however, achieves an improved time complexity of $O(N \log N)$ \cite{75}.

Support vector machine \cite{77} is a new learning method that has not been applied to text categorization until very recently. Its ability to ensure global maximum, generate non-linear classifier and handle a large number of features paves its way to become one of the most popular and best-performing machine learning techniques for text categorization, and pattern recognition in general. SVM is notoriously more expensive than KNN in terms of training time and memory consumption, but there have been studies showing the possibility of improving SVM’s training speed to a level comparable to computationally easy learners such as Rocchio \cite{78}.

Unsupervised learning techniques such as clustering have also been applied to information retrieval since the early days of IR research and remain a topic of active research. Clustering is useful in revealing the topic or style distribution of a collection. There seems to be a clear shift from early hierarchic clustering algorithms \cite{79} to statistical and vector-based methods \cite{80, 81} in line with the progression in the research on retrieval model.

As will be demonstrated in the following chapter, some of the methods overviewed in this section will be adopted or evaluated in different stages of network analysis conducted in this study. Specifically, the NMF and CF clustering algorithms \cite{80, 81} will be evaluated in calculating the similarity matrix, while the KNN concept will be applied in network construction where a similarity matrix is mapped to a one-mode network.
Chapter 3

Research Design and Methods

This chapter addresses the design of the research including different methods and analyses that will be undertaken to answer each of the research questions presented in chapter 1.

Specifically, section 3.1 describes the procedure for retrieving and storing tagging data from the librarything website; section 3.2 presents steps to prepare the data and the models that lay the foundation for later analysis and experiments; co-occurrence analysis and social network analysis methods presented in section 3.3 will be used to answer the first research question:

- What are the connectivity, degree distribution and modularity of networks of tags and authors generated from co-occurrence analysis?

Comparing the different similarity calculation and network construction methods, the community detection analysis will address the research question

- How will different term weighting schemes, network construction methods and filtering coefficients affect the parameters of the resulting similarity network and also its community structure?

from a complex network analysis perspective.

The network visualization and user study delineated in section 3.4 will further address the above question from a user interface standpoint, and also be dedicated to answering the third question:
• Compared to tag clouds, could network visualizations based on tagging analysis facilitate information-seeking tasks, particularly topic summary, impression formation and knowledge discovery, of typical users?

3.1 Data Collection

The major subject in this study is extensive tagging data from librarything. Some social tagging sites provide a certain kind of APIs that facilitate the retrieval of tagging data (last.fm, del.icio.us), while at the time of writing librarything does not. A web crawler is therefore created to retrieve tagging data by downloading the web pages from librarything, parsing the content of the HTML pages and extracting relevant information.

Relevant information about tagging, such as total occurrences of tags in the site and the number of times each tag is associated with an item or a creator, is stored in relational database tables for future manipulation and retrieval.

3.1.1 Web Crawling

There are two different ways to retrieve tagging data from a public social tagging site:

1. Public API

Take, for example, the APIs provided by last.fm. Much of the data available to view on last.fm is available in several formats through the Audioscrobbler Web Services API. This Web service provides all kinds of tagging data from overall top tags to top artists tagged the most times with a particular tag. This URL, for instance,

http://ws.audioscrobbler.com/1.0/artist/Metallica/toptags.xml

provides the most popular tags applied to the music group Metallica. The top tags will then be returned in XML format:

1http://www.audioscrobbler.net/data/webservices/
2. HTML text parsing

For those websites that do not provide APIs for accessing tagging data, an HTTP request is issued to retrieve the web page for a particular item. For example, the page http://www.librarything.com/work/18362 contains the tagging information about the novel *Uncle Tom’s Cabin*. The tags along with the times they are applied to this novel can be retrieved by scanning the page content and performing regular expression match.

For *librarything* where public APIs are not available, tagging data will be retrieved through crawling and parsing HTML pages.

The data collection procedure for *librarything* is as follows:

1. Crawl the author cloud page to get information about authors. Update the *author* table. This is not exhaustive because there are only about 5,000 authors in the cloud page.

2. Crawl the page of each author to get a list of books by the author and a list of tags annotating that author. Update the *book* table with *book ID, title, author.tag*
table with number of times each tag is applied to this author, and for new tags, tag table.

3. Crawl the page of each book to get tags annotating the book. Update the book.tag table with number of times each tag is applied to this book, and for new tags, tag table.

Definitions of the tables mentioned above are the subject of the next section.

### 3.1.2 Database Store and Schema Mapping

The tagging data are stored in a PostgreSQL database for various kinds of manipulation, transformation and analysis. The tables for tagging data about librarything include the author table, book table, tag table, author.tag table, and book.tag table.

Table 3.1 shows the database schema of the table author.

<table>
<thead>
<tr>
<th>attribute</th>
<th>type</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>INTEGER</td>
<td>ID number</td>
</tr>
<tr>
<td>author</td>
<td>VARCHAR(128)</td>
<td>name of the author</td>
</tr>
<tr>
<td>url</td>
<td>VARCHAR(128)</td>
<td>URL of the author description page</td>
</tr>
<tr>
<td>frequency</td>
<td>INTEGER</td>
<td>number of times this author is tagged</td>
</tr>
</tbody>
</table>

Table 3.1: Schema for the table author

<table>
<thead>
<tr>
<th>attribute</th>
<th>type</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>INTEGER</td>
<td>ID number</td>
</tr>
<tr>
<td>author_url</td>
<td>VARCHAR(128)</td>
<td>URL of the author description page (foreign key)</td>
</tr>
<tr>
<td>title</td>
<td>VARCHAR(256)</td>
<td>title of the book</td>
</tr>
<tr>
<td>copy</td>
<td>INTEGER</td>
<td>number of copies owned by librarything users</td>
</tr>
<tr>
<td>review</td>
<td>INTEGER</td>
<td>number of times reviewed by librarything users</td>
</tr>
</tbody>
</table>

Table 3.2: Schema for the table book

The initial run of data collection generated a data set consisting of 5,241 authors, 35,010 books and 248,008 tags.
### Table 3.3: Schema for the table `tag`

<table>
<thead>
<tr>
<th>attribute</th>
<th>type</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>INTEGER</td>
<td>ID number</td>
</tr>
<tr>
<td>tag</td>
<td>VARCHAR(128)</td>
<td>name of the tag</td>
</tr>
<tr>
<td>url</td>
<td>VARCHAR(128)</td>
<td>URL of the tag description page</td>
</tr>
<tr>
<td>frequency</td>
<td>INTEGER</td>
<td>number of times this tag is applied</td>
</tr>
</tbody>
</table>

### Table 3.4: Schema for the table `author_tag`

<table>
<thead>
<tr>
<th>attribute</th>
<th>type</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>author_url</td>
<td>VARCHAR(128)</td>
<td>URL of the author description page</td>
</tr>
<tr>
<td>tag_url</td>
<td>VARCHAR(128)</td>
<td>URL of the tag description page</td>
</tr>
<tr>
<td>occurrence</td>
<td>INTEGER</td>
<td>number of times this tag is applied to this author</td>
</tr>
</tbody>
</table>

Starting from the top authors on the site, the data collection procedure might under-represent the long tail of books, authors and tags, or those that are least frequently accessed. This bias would not distort the results of this study because it aims at the most frequently accessed resources as the major subjects of user interfaces.

### 3.2 Data Modeling

The most frequently used tagging data in this study takes the form of number of times each tag is applied to each item and creator\(^2\). An item or creator, therefore, is easily associated with a bag of tag words as in traditional information retrieval context whereby documents are represented by the bag of words occurring in them.

#### 3.2.1 The Mathematical Definition of Folksonomy

Mathematically, a folksonomy is defined by

\(^2\)In this study creators refer to copyright holders of the item instead of users applying tags. Examples are book authors and artists.
<table>
<thead>
<tr>
<th>attribute</th>
<th>type</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>book_id</td>
<td>VARCHAR(128)</td>
<td>ID of the book</td>
</tr>
<tr>
<td>tag_url</td>
<td>VARCHAR(128)</td>
<td>URL of the tag description page</td>
</tr>
<tr>
<td>occurrence</td>
<td>INTEGER</td>
<td>number of times this tag is applied to this book</td>
</tr>
</tbody>
</table>

Table 3.5: Schema for the table book_tag

\[ F = (C, T, I, A, B) \]

where the five sets represent creators, tags, items and the links between them: \( C \), the set of all creators; \( T \), the set of all tags; \( I \), the set of all items; \( A \subseteq C \times T \), the set of links between creators and tags; and \( B \subseteq I \times T \), the set of links between items and tags.

To indicate the number of times a tag is applied to an item or a creator, matrices \( M \) and \( N \) like term-document matrix in information retrieval will be used where \( M_{ij} \) is the number of times tag \( i \) is applied to author \( j \) while \( N_{ij} \) is the number of times tag \( i \) is applied to item \( j \).

For simplicity, all users are treated equally and therefore not considered in this model.

### 3.2.2 Affiliation Networks and Bipartite Graphs

Affiliation networks are network models created to study the affiliation relationships between social entities. An affiliation network is a network in which actors are joined together by common membership in groups or clubs of some kind. For a long time it has been used to conduct graph-centric analysis of social networks, for example, networks of individuals connected together by joint participation in social events [82], co-authorship in scientific publications [83] and sitting on the board of director of the same company [84].

Structurally, affiliation networks are often represented as bipartite graphs or two-mode networks, even though sometimes they are considered as unipartite graphs or one-mode net-
works for the purpose of simplicity. A bipartite graph is a graph that consists of two kinds of vertices, one representing the actors and the other representing the groups.

Recently tripartite graph has been employed to model relationships between users, items and tags in folksonomy [33]. For the purpose of this study where the relationships between tags and items, and tags and creators are the major subjects, bipartite graph will suffice. A bipartite graph modeling the tagging world contains tags as actors and items or creators as groups.

3.2.3 Representation: The Bag-of-Words Approach

In the information retrieval context, bag-of-words approach assumes that words are conditionally independent, therefore a document can be modeled as a set of words that occur in it.

Closely related to the bag-of-words approach is the TF-IDF (term frequency and inverse document frequency) method, which found its origin in the vector space model of information retrieval. Salton systematically presented the vector space model of information retrieval in a 1975 paper [85], although some of the thoughts had already appeared in his earlier publications [86, 87, 88]. The essential idea of the vector model is that indexing terms are regarded as the coordinates of a multidimensional information space. Both the document and the query are represented by a vector in this space, or ordered list of terms.

Applying the vector space model to tagging, a bag-of-words representation of an item in a social tagging context takes the form

\[ a = [a_1, a_2, \ldots, a_m]^T \]

where \( a_k \) is the number of times tag \( k \) is applied to this item, \( T \) is the matrix transpose operation, and \( m \) the size of the tag vocabulary.
Similarly, a tag word can be represented as

\[ \mathbf{w} = [w_1, w_2, \ldots, w_n]^T \]

where \( w_i \) is the number of times this tag is applied to item \( i \) and \( n \) is the number of items in the collection.

A similar model can be built for the relation between tags and creators.

### 3.2.4 Term Weighting Schemes

Although different term weighting methods have been proposed along with the progression of IR research, the most widely used and proved term weighting schemes in clustering and co-occurrence analysis have not gone far beyond the traditional TF-IDF and its variants, most of which are overviewed by Salton in 1988 [89].

- **binary weight**

  This is a simplistic weighting approach where a tag is only considered to occur or not occur while the number of occurrences is not significant. In vector form, an item can be represented as

  \[ \mathbf{a} = [u_1, u_2, \ldots, u_m]^T \]

  where

  \[
  u_i = \begin{cases} 
  1 & : \text{if } a_i > 0 \\
  0 & : \text{if } a_i = 0
  \end{cases}
  \]
and $a_i$ is the number of times tag $i$ is applied to this item.

This can be improved so that a tag is only considered to occur if the number of occurrences is greater than a threshold $t$, or

$$ w = [u_1, u_2, \ldots, u_m]^T $$

where

$$ u_i = \begin{cases} 
1 & \text{if } a_i > t \\
0 & \text{if } a_i \leq t 
\end{cases} $$

This improvement removes a large number of occasional applications of a tag to an item. In other words, a tag must have been applied multiple times (therefore by multiple taggers) to an item to be considered a legitimate tag for that item.

- **raw term frequency**

  In the social tagging context, this is just the number of times a tag is applied to an item with no further weighting or normalization:

  $$ u_i = a_i $$

- **standard TF weight**

  As larger-term sets are usually assigned to longer documents, the chances of term matches between queries and documents will tend to favor longer documents. To
reduce this bias towards long documents, a normalization factor is often introduced in the context of text retrieval so documents of different lengths can be treated equally for retrieval purposes:

\[
u_i = \frac{a_i}{\sqrt{\sum_{k=1}^{m} a_k^2}}
\]

This standardization may not be desirable in the context of social tag visualization as they actually intend to favor popular tags/items as shown in tag clouds.

- **TF-IDF weight**

The classic TF-IDF weighting is the product of the term frequency and the inverse document frequency:

\[
u_i = a_i \cdot \log \frac{n}{n_i}
\]

where \( n \) is the total number of items, and \( n_i \) is the number of items to which tag \( i \) is applied.

- **Fully Weighted TF-IDF**

Applying both the standard weight and the TF-IDF weight yields:

\[
u_i = \frac{a_i \cdot \log \frac{n}{n_i}}{\sqrt{\sum_{k=1}^{m} (a_k \cdot \log \frac{n}{n_k})^2}}
\]

As term weighting plays an important role in normalization of the tag vector, it will in turn affect the calculated similarity matrix and hence the projected one-mode network. Most
of the weighting schemes listed above will be evaluated in the subsequent network analysis to find the one that is most suitable for developing structural insights into the social tagging system as presented in section 5.1.1.

3.3 Data Analysis

3.3.1 Co-occurrence Analysis

In general, word co-occurrence means concurrence, in a more specific sense, the above-chance frequent occurrence of two terms in a text corpus alongside each other in a particular context. Co-occurrence in this linguistic sense can be interpreted as an indicator of semantic proximity or an idiomatic expression.

Given its emphasis on word relations, co-occurrence analysis has been extensively conducted in the field of natural language processing and text retrieval to tackle problems like word sense disambiguation [90], word classification [91] and lexical distribution analysis [92].

In the context of social tagging, co-occurrences of each pair of tags or authors can be calculated using one of the methods defined in section 3.3.2 and populated into the square similarity matrix $S$ where $s_{ij}$ indicates the normalized number of co-occurrences, or similarity, of tag $i$ and tag $j$ (or author $i$ and author $j$).

In social network analysis, co-occurrence analysis is more frequently referred to as one-mode analysis, co-membership analysis or co-attendance analysis when there is substantive focus on just one of the modes (parties), where a one-mode network is derived from an affiliation network. In matrix form,

$$X^N = AA', X^U = A'A$$

where $X^N$ is the sociomatrix indicating the number of events each pair of actors shares,
$X^U$ is the sociomatrix indicating the number of actors each pair of events shares, and $A$ the affiliation matrix \[59\].

Substituting tag for actor and author for events, this matrix-based co-membership analysis offers a similarity measure for studying the relations between tags and between authors in social tagging, which is equivalent to the co-occurrence based on the binomial model in section \[3.3.2\].

Another factor that makes co-occurrence analysis attractive concerns the phenomenon of synonymy largely found in both natural language texts and social tags. In contrast to automatic indexing for natural language texts, where synonymous terms often do not co-occur in the same document, in a social tagging context, synonymous or semantically close tags are more frequently applied to the same item or author by users with different background and vocabulary.

### 3.3.2 Similarity Measures

Given the vector representation of both items and tag words, there are different ways to determine how close any two items or two words are. The following list is not intended to be exhaustive.

- **Euclidean distance**

  Euclidean distance of item $\mathbf{x}$ and $\mathbf{y}$ is defined by

  $$D = ||\mathbf{x} - \mathbf{y}|| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

  Standardization will be necessary if scales differ.

- **first-order co-occurrence**
First-order co-occurrence is the vector product

\[ S_1(x, y) = x^T \cdot y \]

where \( x \) and \( y \) are vector representations of the two items. Different vector representations yield different variants of the first-order co-occurrence.

When the binary weight vector representation of the two items (the weights are restricted to 0 and 1) is used, the first-order co-occurrence is practically the number of tags that are applied to both items.

When standard weight TF vector representation of the two items is used, first-order co-occurrence becomes the widely used cosine similarity.

Both raw co-occurrence and cosine similarity are generated from conventional vector product formula and are therefore often called first-order co-occurrence.

- second-order co-occurrence

This kind of co-occurrence calculation is based on the co-occurrence vector \( S_1(x) \) for an item \( x \):

\[ S_1(x) = [S_1(x, w_1), S_1(x, w_2), \ldots, S_1(x, w_n)]^T \]

The second-order co-occurrence of two items \( x \) and \( y \) is defined as the cosine similarity of their co-occurrence vectors:

\[ S_2(x, y) = C(S_1(x), S_1(y)) = \frac{S_1(x)^T \cdot S_1(y)}{|S_1(x)||S_1(y)|} \]
• Tanimoto coefficient (extended Jaccard coefficient)

The binary Jaccard coefficient represents the number of 1s in the intersection of two binary vectors divided by the number of 1s in the union of the vectors. In vector form,

\[ T(x,y) = \frac{x^T \cdot y}{|x|^2 + |y|^2 - x^T \cdot y} \]

• similarity based on Latent Semantic Indexing

Latent semantic indexing, a vector-based dimension reduction method, was proposed to address the problem with synonymy and polysemy \cite{93}. As a by-product it also generates the distance matrix of terms and documents. Based on singular value decomposition, a computationally expensive matrix operation, it is only feasible to calculate similarities between a relatively small group of tags given their application to a small group of items.

Given a tag-item or tag-author matrix \( M \),

\[ M = [i_1, i_2, \ldots, i_m] \]

where \( i_k \) is the vector representation of the \( k \)th item, the similarity matrix for all the tags is defined as:

\[ L = T_k S_k^2 T_k' \], where \( T_k, S_k \) are the dimensionally reduced SVD matrices \( T, S \) of \( M \)

\[ M = T S D' \]
where $L_{ij}$ gives the similarity between tag $i$ and tag $j$.

- similarity based on Non-negative Matrix Factorization

Non-negative matrix factorization has proved one of the best-performing document clustering methods based on the classic vector space model of documents [80]. NMF intends to factorize the $m \times n$ term-document matrix $X$ into a non-negative $m \times k$ matrix $U$ and a non-negative $k \times n$ matrix $V^T$ such that the objective function:

$$ J = \|X - UV^T\| $$

is minimum. $\|\cdot\|$ denotes the squared sum of all the elements in the matrix, and $k$ is the number of clusters.

Given the matrix $V^T$, the document similarity matrix can be calculated as

$$ N = VV^T $$

Xu demonstrates that NMF clustering can project the documents into a subspace where the axis corresponds to the cluster and the data points are distributed closely to an axis such that the clustering decision is easily made [80]. This characteristic also makes NMF a feasible technique for network visualization where nodes within the same clique are strongly connected while those of different cliques are weakly connected.

- similarity based on Concept Factorization

Concept factorization is a data clustering method similar to NMF, but it eliminates the non-negative constraint of NMF and therefore can be kernelized [81].
CF intends to factorize the \( m \times n \) term-document matrix \( X \) into a non-negative \( n \times k \) matrix \( W \) and a non-negative \( k \times n \) matrix \( V^T \) such that the objective function:

\[
J = ||X - XWV^T||
\]

is minimum. \( || \cdot || \) denotes the squared sum of all the elements in the matrix, and \( k \) is the number of clusters.

Given the matrix \( V^T \), the document similarity matrix can be calculated as

\[
N = VV^T
\]

The same calculations can be applied to similarities between authors given their tag vectors.

As will be revealed in later evaluation, the effect of specific similarity calculation method is less significant than proper term weighting when it comes to generating one-mode networks with strong community structures.

3.3.3 From Similarity Matrix to Network

A bipartite graph consisting of tags and items as nodes can be transformed into a unipartite graph of tags through the operation called “projection”.

For example, figure 3.1 gives a very quick snapshot of the huge bipartite graph consisting of tags and books in librarything. Here “fiction”, “classic” and “american” have been applied to the book *The catcher in the rye* (they are of course not the only ones), while
“fiction”, “classic”, “British” and “Victorian” are applied to *Jane Eyre*.

![Bipartite Graph of Tags and Books](image1.png)

**Figure 3.1**: A bipartite graph of tags and books in *librarything*

Its unipartite projection will look like figure 3.2 where two tags are connected if they have been applied to the same book. For example, “fiction” is connected with every other tag because they are applied with “fiction” either to *The catcher in the rye* or to *Jane Eyre*.

Not all the associations are created equal, however. “British” and “Victorian” should have a stronger connection than “fiction” with “classic”. To account for the differences between strengths of connections the unipartite graph needs to be weighted, where the weight of an edge indicates the extent to which two tags as nodes are connected, or technically, their similarity in the tagging world.

![Unipartite Graph of Tags](image2.png)

**Figure 3.2**: A unipartite graph of tags in *librarything*

It turns out that the original unipartite graph derived from simple co-occurrence is
strongly connected where some frequently used tags like “fiction”, “non-fiction”, “read” are connected with almost every other tag - even “fiction” and “non-fiction” can be applied, albeit not often, to the same book. This high skewness toward frequent tags can make the graph markedly dense and therefore difficult to analyze and visualize. There are several ways to alleviate the problem:

• focus on the largest connected component.

Tags or authors in the networks often form connected components that are separated from each other. The existence and size of the single largest connected component is a common analysis technique applied in studies of complex networks and may well yield valuable insights into the structure and dynamics of the underlying networks. The largest connected component also represents the best “snapshot” of the entire network and can be used for visualization and subsequent manipulation.

As will be noted later, community structure investigations also rely implicitly on using connected network components. Among all the connected components of a tag/author network, the largest connected component is undoubtedly the most important and interesting subject.

• remove the most frequent tag words.

Removing frequent stop words that do not reveal actual semantics has been a long-term practice in text retrieval. The most frequent word in English “the”, and the second-place word “of” are examples of stop words. Removing them will not impair the quality of index words created while improving the efficiency and performance of the information retrieval model.

The situation in tagging is somewhat different. As can be seen from table 3.6 some of the most frequent tag words in librarything, like “fiction”, “fantasy” and “science-fiction” convey important information about the classification of the item being tagged, while some, like “read” and “tbr”, do not.
Removing most frequent tag words is therefore a double-edged sword. It makes the unipartite graph of tags clearer but loses some potentially important tag words that might serve as the root of the latent taxonomy in the graph. The number of tags that could be removed, or more specifically, which tags could be removed, needs to be fine-tuned to generate an optimized graph structure.

- remove the least frequently used tags.

It is also possible to select the most frequently applied tag words to analyze since their usage patterns are statistically more convincing and therefore immune to various tagging behavior of individuals.

- keep tag words in a controlled vocabulary.

It also makes sense to analyze the usage of tag words that are known to be semantically related, for instance, the words used in a classification system. But without previous knowledge about tags in a social tagging site, the output can be poor because the tag use might not follow the structure of a controlled vocabulary.

- consolidate synonyms and abbreviations.

Not only synonyms but also abbreviations have become popular in social tagging. One of the problems with uncontrolled vocabulary is that different taggers tend to employ
different tags while referring to the same thing. It could be the singular form and plural form, like “classic” and “classics”, semantically similar words like “British” and “English”, or abbreviations like “to-be-read” and “tbr”. Identification of all these forms through automated methods is almost impossible, even though some might be revealed through co-occurrence analysis.

• filter the edges based on weight or topological structure.

The edges could be filtered by either analytical or topological criteria to make the unipartite graph clearer. As a result some nodes that are far from the graphic center and loosely connected with other nodes may also be removed.

Given a number of tag words and similarity between any pair of them, a graph can be formed using the projection method:

\[ G = (V, E) \]

where

\[ V = \text{the set of tag words in question} \]

\[ E = \text{the set of edges connecting two nodes of tag words that co-occur} \]

The graph generated from co-occurrence alone can be strongly connected and too overwhelming to analyze and derive semantic structure. Luxburg proposes several different methods to transform a given set of data points with pairwise similarities (or distances) into a graph [94]:

• The ε-neighborhood graph
In a $\epsilon$-neighborhood graph all nodes whose pairwise distances are smaller than $\epsilon$ are connected. This corresponds to applying a threshold of co-occurrence to the graph generated from tag/author co-occurrences.

- The k-nearest neighbor graph

The goal is to connect vertex $v_i$ with vertex $v_j$ if $v_i$ is among the k-nearest neighbors of $v_j$ and vice versa. A one-way restriction might be applied in the context of a directed graph.

- The fully connected graph

In a fully connected graph all nodes are connected with edges weighted by the similarity of incident nodes.

These different methods for constructing networks, particularly the $\epsilon$-neighborhood method and the KNN method, will be compared and contrasted in both the network analysis and the user study to determine which one is most feasible to map a similarity matrix to one-mode networks that contain strong semantic structures and can be visually approached by users.

### 3.3.4 Detection of Community Structure

The surge of interests in the properties of complex networks, including the Internet, citation networks, social networks, biological networks etc. has led to the study of community structure as an important property of networks, whereby network nodes cluster into strongly connected groups and the inter-group links are relatively weak. The ability to find and analyze such groups can provide invaluable help in understanding and visualizing the structure of networks.

Girvan and Newman initially proposed a community detection algorithm which iteratively removes edges with high betweenness scores and seems to have achieved satisfactory results \cite{95}.
1. Calculate betweenness scores for all edges in the network;

2. Find the edge with the highest score and remove it from the network;

3. Recalculate betweenness for all remaining edges;

4. Repeat from step 2.

The disadvantage of this algorithm is its worst-case time complexity of $O(m^2n)$ on a network with $m$ edges and $n$ nodes. They later devised a better-performing community detection algorithm aimed at reducing modularity of the network on each step, which boasts a worst-case running time of $O((m + n)n)$ [96]:

1. Start with a state in which each node is the sole member of one of $n$ communities;

2. Calculate the increased modularity for each pair of communities if the pair of communities are joined together;

3. Join together the pair of communities that results in the greatest increase in modularity;

4. Repeat from step 2.

A test benchmark of a particular network division, the modularity of a network is

$$Q = \sum_i (e_{ii} - \alpha_i^2)$$

where $e_{ii}$ is the fraction of edges in the network that fall within group $i$, and $e_{ij}(i \neq j)$ is one-half of the fraction of edges that connect nodes in group $i$ to those in group $j$, while $\alpha_i$ is

53
\[ \alpha_i = \sum_j e_{ij} \]

and represents the fraction of all ends of edges that are attached to nodes in group \( i \).

The modularity of the network with respect to a network division offers an objective metric for choosing the number of communities into which a network should be divided. As many factors in the analysis, notably the term weighting scheme, similarity measure and network construction method, will affect the topology of projected one-mode network, hence the modularity value of the network, the methods that lead to networks with stronger community structures should be preferred in practicality when it comes to mining semantic structures in the underlying system. The modularity value of the optimized network division as the output of the community detection algorithm, therefore, will be employed in subsequent analysis as an objective measure to evaluate different term weighting schemes, similarity calculation methods and network construction methods.

### 3.4 Visualization and User Study

#### 3.4.1 Visualization of Network Data

As detailed in section 2.2, visualization of documents often offers a clear visual topical summary of the whole collection and presents a big picture as well as the interrelationship between documents. Visualization of tags or authors in social tagging is conjectured to have a similar effect. It might also spur new design for the browsing interface now dominated by tag clouds in current social tagging sites.

The Java programming language is used throughout this study because it is platform independent, flexible and moreover, there are several mature third-party Java visualization libraries that could be leveraged for this study. Among them are JUNG (Java Universal
Network/Graph Framework) \(^3\) and the Prefuse information visualization toolkit \(^4\) which was used in this study, as will be discussed later.

Another resource that might be potentially useful is the Many Eyes web site which provides collaborative visualization services, allowing users to upload data sets, visualize them, and comment on each other's visualizations \(^98\). Its public accessibility, low barrier to entry and richness in visualization schemes makes it a good choice for presenting mature visualization paradigms for relevant tagging data sets to end users. Its tag cloud visualization is easy-to-use and among the most popular visualization schemes on the site.

A network containing a large number of nodes and edges may well turn to a mess on the screen without appropriate visualization techniques. The aforementioned force-directed graph drawing algorithms can help spread the nodes given the topological structure of the graph. Other methods could also be used to generate a clear and arresting visualization:

- **graph reduction**
  
  Different graph reduction methods as detailed in previous sections will have to be used to emphasize a particular subset of nodes or portion of the graph.

- **font size**
  
  As used in tag clouds, font size variation of nodes may indicate the weights of nodes, thereby distinguishing salient nodes from the rest.

- **spatial proximity**
  
  As an effect of the layout algorithm, items that are closely related, especially those that are members of a clique will appear near each other in a network visualization. This is an important feature of network-based display and will be closely examined in the study.

\(^3\)http://jung.sourceforge.net/
\(^4\)http://prefuse.org/
3.4.2 Experimental Design

The main goal of the experiment is to evaluate the effect of different schemes that can be utilized in presenting tags or authors. Goal-oriented tasks as opposed to free browsing tasks are used because of the difficulties with the simulation of free browsing tasks in controlled experiments [19].

Performing visualization-based data mining tasks in the real world involves a significant amount of knowledge, experience and time. It would be difficult, if not impossible, to reproduce the environment where literary scholars and information scientists iteratively apply different data analysis methods to reveal patterns in the literary world. To make the experiment manageable, participants will not be using the information visualization tools (Prefuse and Wordle) on their own to find out patterns and relationships. Instead, the visualizations will be fine tuned and printed out before the experiment to minimize the efforts of participants.

In the spirit of the tag cloud evaluation carried out by Rivadeneira et al [16], there are several information-seeking tasks that can be supported by graph visualization, including tag cloud:

- search: locating a specific term or one that represents a desired concept (or determining that it is not there).

- browsing: using visualization as a means to browse, often with no specific item, creator, or topic in mind.

- impression formation: using visualization as a means to form a general impression of the underlying collection.

- interrelationship presentation: using visualization as a means to present interrelationships between items or authors in the underlying collection.
• recognition/matching: recognizing whether the item or author in mind is likely to be in the collection represented by the graph visualization.

As suggested in several evaluation studies of tag clouds, current design features of tag clouds, especially alphabetical arrangements of displayed tags, do not seem to have a strong effect on high-level cognitive processes of impression formation and inference of semantic relation between tags, authors or concepts [16, 20, 21]. The controlled experiment, therefore, is intended to evaluate the effect different visualization schemes have on facilitating several high-level cognitive and information-seeking tasks.

The elements of the experiment are as follows:

• Method

The experiment is conducted by presenting printouts of both author cloud and author network visualizations. All presentations have a spatial layout. The experimental design will be a network vs. cloud comparative design.

In this study, subjects were presented with multiple printouts of cloud and network visualizations. They were then asked to identify topics based on impression from the presentation and circle groups of authors with a pen based on their knowledge and preferences.

• Subjects

Ten participants were recruited from a university. All participants have normal or corrected-to-normal vision with no color vision deficiency. The participants are all undergraduate students of literature-related major. One of the prerequisites of the participants is familiarity with American and English literature. The native language of all participants is English.

• Stimuli
Visualizations of Author Names

Two groups of authors, namely A1 and A2, representing different subsets of the librarything collection were selected.

- A1 is a fixed group of authors that are relatively popular and easy to categorize. An example would be the top authors that are tagged as “classic”. Visualizations generated from this group were presented to each of the subjects.

- A2 is a group of authors strongly related to an author or a tag of the subject’s choosing. If one of the favorite authors of the subject, for example, is James Michener, A2 would be the top authors co-occurring with James Michener on librarything.

Creating both a fixed group and a customized group allows easy comparison between responses of subjects and also leaves open the possibility of exploring different aspects of the tagging world.

Two author networks with different parameters were generated for each of these groups, as well as two paired author clouds. As network reduction is involved in both of the two network construction methods (ε-neighborhood or KNN), the set of authors in any of the generated cloud or network will be a subset of the original set of authors (A1 or A2).

As will be shown in chapter 5, first-order co-occurrence calculation combined with fully weighted TF-IDF weighting are sufficient in detecting community structures in the one-mode networks, they will then be used to generate the two networks for the user study. One of them will be constructed using the ε-neighborhood method and the other will be constructed using the KNN method.

Note that the set of authors in the two generated networks will not necessarily be identical, as is the case for the set of authors in the clouds. However, the set of authors
in the paired cloud and network will always be identical.

These clouds and networks will be used for performing the task of identifying groups and relationships in the underlying collection. Varying the network parameters and then observing the responses from subjects may reveal the influence that different network configurations have on user perception and help gain insight into the interplay between network configuration and usability.

Visualizations of Tags

One group of tags, namely T, representing those typically used in the librarything collection were selected. An example would be the 300 most popular tags in the collection. Visualizations generated from this group were presented to each of the subjects.

Size of the Groups

The size of all the groups is around 100, allowing for a sufficient number of tags/authors to be presented in the visualization while avoiding excessive or redundant information. This is also close to what has been used in previous studies [99].

Generating the Visualizations

Specifically, the visualizations include

– normal author cloud in alphabetical order with font size indicating the popularity or number of occurrences.

– network graphs of nodes and edges generated from co-occurrence analysis where nodes represent authors and edges indicate the relations between nodes.

Author networks are Prefuse visualizations of tag analysis results exported in GraphML[5]

format.

Author clouds are generated by the Wordle website \(^6\) (also available on ManyEyes). Tags in the clouds are placed alphabetically starting at the top left as this is the most typical configuration as seen in tag clouds applications.

The visual elements in both visualizations are fine tuned. Prefuse toolkit offers a drag-and-drop interface to zoom the displayed graph, move around the nodes and highlight a node along with its neighbors. To fine tune the layout Prefuse also provides several controls to adjust the drag coefficient, spring coefficient and spring length that are used by its spring force-based layout algorithm. All these controls help fine tune the space between nodes, number of nodes and font size that make the final visualization arresting and the names legible. The controls available at the Wordle site are relatively limited. Most of them change the presentational elements, such as font, color and edge. The tags in its tag cloud can be laid out horizontally vs. vertically, or in alphabetical order vs in random order. There is little space to change the structure of the tag cloud, such as grouping certain tags.

Figure 3.3 shows the interface features of Prefuse and Wordle.

Presentational features that were utilized to strengthen visual effects are:

- font size

  Font size is used in both cloud and network presentations to distinguish nodes and texts based on number of total tags applied to an author.

- edges between items

  An edge between two items in a network indicates a semantic link between these two items. Lengths of the edges or distances between connected items are not significant because the algorithm lays out the items based on criteria such as

\(^6\)http://www.wordle.net
structural coherence and aesthetical appealingness. However, a clique of items or a structure close to a clique in a network where connected items are usually very close to each other suggests strong interrelationships between the items.

Figure 3.4 gives examples of network and word cloud visualizations of 119 authors that are most frequently tagged as “classic”. The author network is based on first-order co-occurrence between authors and filtered using a threshold of co-occurrence and a threshold of relative co-occurrence. The displayed network is the largest connected component in a network of 250 authors that are most frequently tagged as “classic”. The same group of authors is also visualized as a tag cloud. In both visualizations font size is proportional to overall popularity as measured by the number of times tags are applied to the author.

Figure 3.5 presents a different visualization of the same data set using second-order co-occurrence calculation and k-nearest neighbor network construction. This kind of network is characterized by a more spreading layout with a larger number of groups, sometimes cliques, of authors, compared to the previous network.

- Procedure

**Step 1**

Subjects are first instructed about basics of social tagging and configuration of author clouds and networks. For example, words are alphabetically ordered in an author cloud, and font size of a word reflects the popularity of that word. Edges in a tag network indicate semantic relationships between the nodes. The explanation will be accompanied by cloud and network examples that will not be used in later experiments.

**Step 2**

Then subjects are presented with two cloud presentations for author group A1, two cloud presentations for author group A2, and two network presentations generated from
author group A1, A2, respectively, using two different network construction methods ($\epsilon$-neighborhood and KNN).

1. After subjects are presented with an author cloud generated from author group A1, they are then asked to identify groups and relationships within the underlying collection that the author cloud represents. They will use a pen to mark groups on the printout in a limited period of time (3 minutes).

2. After subjects are presented with an author network generated from author group A1, they are then asked to identify groups and relationships within the underlying collection that the author network represents. They will use a pen to mark groups on the printout in a limited period of time (3 minutes).

After each task they will also be asked to give the reason that the authors are grouped together or summarize the groups using a word or phrase, such as “historical drama” or “psychological fiction”.

The same procedure will be performed for author group A2.

Each participant will be presented with 4 clouds and 4 networks generated from 2 groups of authors (A1 and A2), as summarized in table 3.7.

<table>
<thead>
<tr>
<th></th>
<th>$\epsilon$-neighborhood method</th>
<th>KNN method</th>
</tr>
</thead>
<tbody>
<tr>
<td>author group A1</td>
<td>cloud/network 1</td>
<td>cloud/network 2</td>
</tr>
<tr>
<td>author group A2</td>
<td>cloud/network 3</td>
<td>cloud/network 4</td>
</tr>
</tbody>
</table>

Table 3.7: Configuration of stimuli

Participants will be allowed to look at the visualizations while performing the tasks so they will not need to memorize any part of the visualizations. They will be asked to complete an individual task and identify as many groups and authors as possible in 3 minutes.
Both the number of groups and number of authors identified in each task will be recorded for data analysis.

**Step 3**

Then subjects are presented with the cloud visualization for tag group $T$, and the network visualization for tag group $T$.

1. After subjects are presented with a tag cloud generated from tag group $T$, they are then asked to describe the principal topics of the underlying collection.

2. After subjects are presented with a tag network generated from tag group $T$, they are then asked to describe the principal topics of the underlying collection.

**Step 4**

After being presented with the visualizations, subjects will be asked to look at 15 groups of authors as the output of the community detection algorithm.

- After looking at the groups, they will then rate the validity of the grouping on a scale of 1 to 10, with 10 indicating that all the authors in that group are similar by some criteria (country, genre, historic period, gender, etc.) and should fall into the same category. If the groups are meaningful, the subjects will also be asked to give the reason that the authors are grouped together or summarize the groups using a word or phrase, such as “historical drama” or “psychological fiction”.

The scale value will be divided by 10 to obtain a value between 0 and 1 for the purpose of correlation analysis.

**Step 5**

At the end subjects are asked about the general impression on these two presentations and then detailed follow-up questions, for example,
Do you think the author cloud/network is a good way to get the gist of the underlying collection?

Do you think the author cloud/network can help grab users’ attention to a social tagging site?

Do you think the author cloud/network can invite exploration of and participation in the tagging community?

Do you think differences in font sizes and the alphabetical ordering of author clouds are helpful?

Do you think network presentation helps you identify groups of or relationships between authors?

Do you think cloud presentation makes it easy for visually important authors to stand out, in comparison to network presentation?

• Analysis of Results

The scores (number of groups and number of authors identified) based on cloud presentations and network presentations obtained in step 1 can then be compared to see if there is any significant difference in effects that these two presentations have on impression formation and relationship presentation. Particularly, student t-test will be used to test the hypotheses:

True difference between mean of scores attained from cloud presentations

\[ x_1, x_2, ..., x_n \]

and mean of scores attained from network presentations

\[ y_1, y_2, ..., y_n \]

is not equal to 0, where \( n \) is the number of subjects.

The test will be evaluated on both the number of groups and the number of authors identified by each participant in the group identification task.
In addition to independent comparisons, two-factor ANOVA with repeated measures will also be conducted to analyze the interaction between network construction method and visualization method.

Note that the order in which cloud and network visualizations generated using the same network construction method from the same group is randomized for each participant such that in about half of the cases cloud visualization is presented before network visualization is presented, while in the other half of the cases network visualization is presented before cloud visualization. The recency effect will therefore be minimized.

The scores of automatically detected communities obtained in step 2 will also be compared to the average clustering coefficient of the communities to see if the coherence of groups as estimated by human subjects can be approximately measured with their average clustering coefficients.

The results will also be evaluated qualitatively, especially the differences and similarities between topics and groups identified while the participants are looking at network visualizations and those identified while they are looking at cloud visualizations. For example, does a cloud visualization help the participants identify a larger number of topics? Or does a network visualization lead to larger groups of tags/authors being identified?

Also, the consistency or variance across the responses of different subjects, for example, needs to be examined qualitatively, to reveal whether a given visualization leads to consensus over groups or topics identified or largely variant responses. The variance across the rates obtained in step 2 will also be closely examined to see if there is any consensus among users’ perception on groups of authors as a result of community detection. If subjects have different opinions on the quality of the grouping, it may be
worth further investigating subjects’ labeling and summarization of groups in visualizations so the discrepancy can be narrowed down to deficiency of network generation and community detection, or users’ background and knowledge.

Equally important to this study is the outcome from the final interview, especially the qualitative aspects that cannot be easily quantified, for example:

- effects of network configuration
  
  It is speculated that tag/author network can help identify groups and relationships of authors/tags. If this is demonstrated in the study, the interview will further explore the basis on which participants report groups and relationships in tag/author network presentation, for example, genre, gender, country of origin, etc are easily identified by groups of connected nodes. If this cannot be supported by the study, the interview can help find out why the seemingly obvious groups of tags/authors in the network presentation fail to make sense to the participants.

- interplay between visual elements
  
  Font size has been observed as one of the major features that will make tags visually important. But the edges between tags in a presentation can downplay the effects of fonts and shift the focus of readers toward groups, shapes or other visual elements formed in a network presentation. The interview may help find out if this kind of interplay between visual elements exists.

- differences between networks generated using different term weighting, similarity calculation and network construction methods
  
  The selection of weighting schemes, similarity measures and network construction methods has been reported to play a significant role in the semantic analysis of the folksonomy system. The networks generated through different methods can be topologically and visually different. A promising outcome will be the heuristics as to how to select and apply different similarity calculation and network
construction methods to create semantically relevant and aesthetically appealing one-mode networks.

3.5 **Summary**

This chapter delineates the methods that are used to address the research questions and test the hypotheses. Specifically, statistical methods, data plots and small-world models will be used to answer the first research question

- What are the connectivity, degree distribution and modularity of networks of tags and authors generated from co-occurrence analysis?

Chapter 4 will cover this part of the work.

While similarity analysis and community detection methods will be employed to address the second question

- How will different term weighting schemes, network construction methods and filtering coefficients affect the parameters of the resulting similarity network and also its community structure?

The results will be presented in chapter 5.

Chapter 6 will be dedicated to tackling the third research question

- Compared to tag clouds, could network visualizations based on tagging analysis facilitate information-seeking tasks, particularly topic summary, impression formation and knowledge discovery, of typical users?

with findings from the user study.
Figure 3.3: Interface of Prefuse and Wordle
Figure 3.4: Visualizations of top authors tagged as 'classic' in librarything
Figure 3.5: Visualization of top authors tagged as 'classic' in librarything based on second-order co-occurrence and KNN method
Chapter 4

Tag Distribution and Small-World Properties

Despite the intensive study of a wide variety of real world complex networks starting from the late 90s, a well grounded framework is still missing when it comes to two-mode networks as they pose new challenges for methodologies in structural properties, dynamics and community characteristics of one-mode networks.

A natural way to tackle the problem in two-mode network research lies in projecting two-mode networks into links of one kind, then applying common methods in social network analysis to resulting one-mode networks. A classic example is Newman’s study of scientific collaboration networks in which two scientists are considered to be connected if they have authored a paper together [83].

Lambiotte and Ausloos extend this method by applying cosine similarity to reduce online social networks and systematically removing links in the resulting networks, a method that finds its origin in percolation theory [100] but also corresponds to manipulating $\epsilon$ of the $\epsilon$-neighborhood graph method widely used in spectral clustering [94].

In this study other similarity calculation methods than cosine similarity are also evaluated, along with the popular KNN graph in addition to $\epsilon$-neighborhood graph for network construction.

This section is framed around statistical distribution and network characteristics of tag-author networks, with the focus on the small-world properties of projected one-mode networks. Both average path length and clustering coefficient have implications on community formation, as will be shown in this and the next chapter.

It is worth mentioning that the data set used in this study is neither an exhaustive list
of the whole tagging site nor a significant portion of it. By looking at the tip of the iceberg, however, some statistical characteristics of social tagging are studied, as generally performed in other studies, in an effort to extrapolate to the larger social tagging world.

4.1 Tag Distribution in Two-Mode Network

Not dissimilar to what has been reported on natural language corpus [101], Wikipedia [29] and other social tagging sites [31, 102, 103], tag occurrences in the librarything data set, either across all authors/books or over individual authors/books, as will be demonstrated in this section, follow the power law distribution. This implies that difference in tag word usage exists in a way similar to what was found in both natural language texts created by individual authors (books) or through collaborative efforts (Wikipedia), and tagging data from other social tagging sites such as del.icio.us. This difference can entail non-negligible semantic clues that might be revealed by methods that have been successfully applied in traditional information retrieval, text categorization and text clustering tasks.

4.1.1 Degree Distribution

Power law distribution in the use of tags has already been observed for social tagging sites del.icio.us, citeulike [102] and flickr [103], and also confirmed by the librarything data, as plotted in figure 4.1.

This scatter plot gives the number of tags (Y axis) that are applied to a given number of items (X axis). Figure 4.2 gives the number of tags (Y axis) that are applied to a given number of authors (X axis).

In a log-log plot, a linear trend indicates a power law distribution. These plots clearly show power law distribution of tags over items and over authors, suggesting that there is a considerable degree of user agreement between tag usage, as one tag word becomes more popular than another, for example, “fiction” vs. “novel”, its usage will increase at a higher
rate, leading to an even greater advance in usage as the number of users and items grows.

As authors and tags are the two “modes” of the author-tag network while items and tags are the two “modes” of the item-tag network, the two plots actually give the degree distribution of the author-tag network and the item-tag network, respectively. In other words, in the original two-mode network directly generated from the data set, the degree of a node (of either “mode”) is governed by the power law distribution.

This is also very similar to what has been witnessed on the relation between number of articles in Wikipedia and the number of categories assigned [29], which is intuitively unsurprising as categories are a controlled form of tags. It seems that whether the vocabulary is controlled (categories in Wikipedia) or not (tags in social tagging sites) does not constitute a difference in the statistical characteristics of tag/category usage.

Combined with Heymann et al’s finding that librarything tags may be competitive with manually entered metadata created by paid taggers and experts [104], this statistical analysis shows that tags are comparable to manually generated metadata (tags, categories
or subject headings) not only in quality, but also in collective characteristics, which is the key to the network analysis employed in this research.

4.1.2 Occurrence vs Rank

Figure 4.3 is a log-log plot of number of items vs rank of tag. Again, a linear pattern is clearly exhibited, meaning the number of items to which any tag word is applied is roughly inversely proportional to its rank.

Figure 4.4 gives the relation between total occurrence of a tag and its rank in the frequency table. As can be seen from this plot, the occurrence of any tag word is also approximately inversely proportional to its rank.

Figure 4.5 and figure 4.6 exhibit the same pattern for tags applied to authors.

In figure 4.6 it is observed that the lowest ranking tags occur less often than would be expected for a power law distribution given the trend of the higher-ranking tags, implying that the usage of those lowest-ranking tags tends to be individual behavior and is hardly
affected by the social nature of folksonomy. This deviation at the lower end of relative position is also corroborated by the observation on del.icio.us data [31].

The power-law pattern shown in figure 4.3 through figure 4.6 is almost identical to what is found in a corpus of natural language utterances, including Wikipedia.

Figure 4.7 gives the distribution of tag occurrence over rank for three individual books, which exhibits the same pattern.

Figure 4.8 presents the relation for three individual authors. It demonstrates that the number of times a tag is applied to an author is inversely proportional to its position in the rank table of that author.
4.2 Small-World Properties of One-Mode Networks

In order to study its network characteristics, a two-mode network is often reduced to a one-mode network using appropriate similarity measure and network construction method. A tag-author network, for example, can be reduced to a tag-tag network and an author-author network. Given the weights of edges and topological structure of the reduced one-mode network, edges can be removed to better reflect the relations between tags/authors and make the salient groups of nodes stand out. The topology of a reduced one-mode network can change dramatically as edges are removed between nodes, and as a result, nodes are removed if their incident edges have been removed, but the small-world property of the remaining network still holds.

The focus of the analysis on the one-mode tag/author network is its largest connected component as is the case in previous studies [100, 105]. The nodes not included in the largest connected component tend to be either part of very small, isolated clusters or are
not connected to other nodes at all.

Table 4.1 presents the small-world property of a network of 2,492 authors and 33,142 edges in librarything after many edges and authors have already been removed using the $\epsilon$-neighborhood graph method ($\epsilon = 0.587$). The data shows that even though a random network with the same number of nodes and edges boasts a significantly shorter average path length, its clustering coefficient is negligibly small, compared to the author network. In other words, while it takes on average a significantly longer path in the author network to walk from one node to another, the tagged authors actually form substantially more strongly connected components.

Despite its small scale compared to what have been studied in social network analysis, the network quantitatively represents some of the most important characteristics of the social tagging world.

Figure 4.9 gives the distribution of path length in the graph along with the average path length. It seems the path lengths follow a normal distribution.
4.3 Dynamics of Small-World Properties

In order to further examine collective effects in social tagging, the behavior of average degree, average path length and clustering coefficient in relation to change in the filtering coefficient ($\epsilon$) in the $\epsilon$-neighborhood network is studied in detail.

The most important parameter in $\epsilon$-neighborhood graphs is the filtering coefficient ($\epsilon$), or the threshold for edges that are removed from the original network. As the filtering coefficient is raised, edges with similarity values larger than the coefficient are retained, making those stronger semantic structures stand out.
4.3.1 Distribution of Similarity Values

Before applying the $\epsilon$-neighborhood graph method, it may be helpful to take a closer look at the distribution of similarity values in the one-mode author network.

In figure 4.10 the number of pair-wise similarity values in each of the intervals of 0.01 from 0.01 to 1.0 is plotted.

The plot exhibits a clear linear pattern on a log-log scale, indicating a power law distribution of the similarity values. The large proportion of edges represented by low similarity value will soon be removed as the filtering coefficient increases from 0.0, which corresponds to a virtually fully connected network. This plot also suggests that to obtain a reasonably connected network with stronger community structure, the filtering coefficient should be well over 0.1, more towards the end of 1.0.
4.3.2 Clustering Coefficient

As noted before, the clustering coefficient is a measure of the density of triangles, and therefore social effects in complex networks. In other words, it measures to what extent two neighboring nodes of a given node are also neighbors.

In figure 4.11, the clustering coefficient of author networks is plotted against the change in the filtering coefficient, along with the number of nodes in the largest connected component of the filtered network as a percentage of the original network.

The plot shows a sharp decrease of the size of the largest connected component when the filtering coefficient increases from about 0.5 to about 0.7; this corresponds to the interval in which similarity values are relatively dense. A large number of nodes are removed when the filtering coefficient increases in this interval, while fewer nodes are removed when the filtering coefficient is smaller than 0.4.

In stark contrast to the drop in the size of the largest component, the clustering coefficient
remains relatively stable, taking a value larger than 0.6 as the filtering coefficient changes from 0.2 to 1.0, indicating consistently strong social effects in the one-mode author networks.

### 4.3.3 Average Degree

The dependency of the average degree on the filtering coefficient is plotted in figure 4.12. The average degree of the networks decreases, dramatically at the beginning, then slowly, and then stabilizes at about 30, as the filtering coefficient increases and a growing number of edges are removed, meaning there are consistently dense connections among the nodes in the largest connected component, another sign of strong social effects.

### 4.3.4 Average Path Length

As can be seen in figure 4.13, the dependency of the average path length on the filtering coefficient is more complex. The average path length of the networks increases from about
Figure 4.10: Distribution of similarity matrix elements in author network in librarything 2, reaching its maximum at about 12 when the filtering coefficient is around 0.66, then falls to about 2 again. This demonstrates that the network is extremely dense and connected for small filtering coefficient, such that a node is almost always connected to another node, or can be connected to another node in one or two steps, leading to a small average path length. As the filtering coefficient increases and relatively weak edges are removed, the strongly connected components stay, with some “bridge nodes” connecting the components together, thus increasing the average path length. Further increasing the filtering coefficient will remove the “bridge nodes” or the connections between components, leaving individual components as separate “islands”, thereby reducing the average path length. As the filtering coefficient approaches 1.0, few connected components survive, rendering an average path length less than 2.

The dependency of path length distribution on filtering coefficient is more clearly shown in figure 4.14 and 4.15 which depict the distribution curves for multiple $\epsilon$ value. When the filtering coefficient is moving up from 0.55, the path length distribution curve exhibits a
more flattened shape than the nearly normal distribution curve at 0.55. The peaks at 0.6 and 0.65 are substantially lower while the long tails are more pronounced, suggesting that a smaller percentage of paths are close to the average path length, while a larger number of paths are significantly longer than the average path length. In other words, the “diameter” of the network increases when the filtering coefficient is increased from 0.55 to 0.65, which is largely reflected in the stronger community structure of the resulting networks, as will be described later.

The average path length reaches its maximum when the filtering coefficient is around 0.65. When the coefficient further increases, the average path length drops dramatically, as demonstrated by the gap around the center of figure 4.15 - further increasing the coefficient leads to some of the “bridge nodes” being removed from the network and the network has been divided into several strongly connected components. This is also confirmed by the steep slope in the same area in figure 4.11 or the number of nodes in the largest connected component drops substantially when the filtering coefficient is increased from 0.65 to 0.75.
Figure 4.12: Dependence of the average degree on the filtering coefficient

Further increasing the filtering coefficient will remove more “bridge nodes”, separating the network into even smaller connected components with smaller “diameter”, as indicated by both the average path length and path length distribution.

The tagging world is indeed “small” in the sense that the longest path between two nodes in the network is 40, and most of the paths are less than 15. The same might be true for a random graph, but there is a significant difference in clustering coefficient between the librarything network and a random network with the same number of nodes and edges, suggesting a more condensed structure in the author network. Two authors that are similar to another author are very likely to be similar as well. The condensed clusters of authors/tags are believed to convey semantic clues about the tags and resources in the tagging universe, as groups and cliques of people found in social networks.
4.4 Degree Distribution of One-Mode Networks

As depicted in figure 4.16 the power law distribution of the projected one-mode network (author network) is not as strict as those of the original two-mode network. The scale-free property of the one-mode network, however, appears to be evident. This is unsurprising given that tags, authors and items continuously grow over time and that more frequent tags are preferable over less frequent tags when new users tag new authors and items, satisfying the two theoretical assumptions of formation of scale-free networks: continuous growth and preferential attachment.

The plot of link distribution in figure 4.17 shows that 20% of mostly connected nodes are attached by over 60% of links in the network. This is less pronounced than what is reported on other online social networks [105, 106], but still demonstrates that there are a significant number of densely connected nodes or “hubs” in the projected author network.
Figure 4.14: Dependence of path length distribution on the filtering coefficient (I)

4.5 Summary

The power law degree distribution of the tag-author and tag-item networks, and the scale-free property of projected one-mode networks, suggest that the tag frequency, degree property and clustering structure convey semantic clues that might be revealed by automated methods. For example, those frequently occurring tags and tags with high degree in the tag network are probably close to the top of a hierarchical classification system, while clusters of tags in the tag network are likely to contain tag words that are semantically related or adjacent in the classification system. A previous study in this direction is Heymann et al’s algorithm for converting a large corpus of tags annotating objects in a tagging system into a navigable hierarchical taxonomy of tags [102]. Another is a semantic smoothing technique proposed by Eda et al that uses the level of tag generalization to form the objective tags into a hierarchy [107].

The small-world property, especially the unusually high clustering coefficient of projected
Figure 4.15: Dependence of path length distribution on the filtering coefficient (II)

One-mode author/tag networks, indicates the presence of strong clustering structures, the properties of which warrant further analysis and evaluation. The next chapter will then be dedicated to systematically analyzing and interpreting these structures using community detection methods.
Figure 4.16: Degree distribution of author network in librarything

Figure 4.17: Distribution of links across nodes in author network
Chapter 5

Identifying Groups in One-Mode Networks

In complex network analysis, a *community* is defined as a group of nodes that are relatively densely connected to each other but sparsely connected to other dense groups in the network [108]. This notion of community structure has been widely adopted but remains theoretically elusive. In practicality, one of the first-principles definitions is to identify indivisible subgraphs as communities, whereby a network is iteratively divided into subgraphs and subgraphs of subgraphs, and the iteration stops when all the subgraphs obtained are indivisible [109].

As strong community structures pervade social networks, biochemical networks and information networks, the ability to detect the communities or groups in a network is of significant practical importance for a wide variety of complex systems. Sometimes communities correspond to functional units, such as cycles or pathways in metabolic networks, or collections of web pages for a topic. Also networks can have properties at the community level that are different from their properties at the level of the entire network. Identifying communities in networks has substantial implications for the understanding of the underlying systems that the networks represent.

In this chapter community detection algorithms are applied to one-mode author networks as projected from tag-author networks. Complementary to the more rigorous analysis of small-world property conducted in the previous chapter, the groups of authors in the networks as the result of community detection are demonstrated to constitute concrete examples of strong community structures in the network, as they convey semantic information about the relations between authors.
Different term weighting schemes, similarity calculation methods and network construction methods are tested, and the analysis suggests that independent of similarity calculation methods and network construction methods, the community structures in author networks are stronger than many of the online networks studied, as measured by modularity.

5.1 Fine Tuning the Parameters

There have been a few existing studies of the structural properties of one-mode networks derived from online social networks, notably the analysis of web-downloaded data on people sharing their music library [100], and the study of shared vocabularies in the social bookmarking site del.icio.us [110].

In many of existing studies cosine similarity is used to calculate the similarity between entities (tags or music groups for example), thereby projecting the two-mode networks into one-mode networks. As will be demonstrated in this section, however, cosine similarity is actually not the best way to do this kind of calculation in terms of maximizing the modularity value.

In light of the framework proposed in the previous section, three major steps are involved in building one-mode networks from vector representation of entities:

- normalization of vectors
- similarity calculation
- constructing network from similarity matrix

Different patterns of clustering have been observed in networks generated using different term weighting schemes, similarity calculation methods and network construction methods. By varying the parameters in these three steps, this section is intended to explore the effect of the parameters on the resulting one-mode networks, thus evaluating the methodology and
heuristics for building one-mode networks in the social tagging universe, and potentially in other applications where two-mode networks are to be reduced to one-mode networks.

5.1.1 Term Weighting Schemes

Term weighting has proved a basic yet important step since the inception of research on information retrieval. Applying different kinds of term weighting in the IR context may significantly affect the performance of IR tasks. Several classic term weighting schemes are evaluated in this section and it is suggested that other conditions being equal, the fully weighted TF-IDF is the top-performing weighting scheme.

Fixing first-order co-occurrence as the similarity calculation method, \( \varepsilon \)-neighborhood as the network construction method, three author networks are generated using the binary weight, standard TF weight or fully weighted TF-IDF as the term weighting scheme. Note that the largest connected component of the resulting one-mode network is always the subject of analysis, as described in previous sections.

All the three networks (largest connected components) contain about 2,500 authors. As there are about 3,500 authors in the data set, these 2,500 authors in the largest connected component are a reasonable representation of the data set and also the librarything data.

The average degree, clustering coefficient, average path length, and particularly the modularity of three networks are summarized in Table 5.1:

<table>
<thead>
<tr>
<th>scheme</th>
<th>average degree</th>
<th>( C )</th>
<th>( l )</th>
<th>modularity</th>
<th>number of communities</th>
</tr>
</thead>
<tbody>
<tr>
<td>binary weight</td>
<td>328.3996</td>
<td>0.7756</td>
<td>2.2130</td>
<td>0.1538</td>
<td>10</td>
</tr>
<tr>
<td>standard TF weight</td>
<td>131.0624</td>
<td>0.7548</td>
<td>6.6859</td>
<td>0.4295</td>
<td>26</td>
</tr>
<tr>
<td>fully weighted TF-IDF</td>
<td>26.5987</td>
<td>0.7251</td>
<td>9.0484</td>
<td>0.8610</td>
<td>38</td>
</tr>
</tbody>
</table>

Table 5.1: Characteristics of networks generated using different term weighting schemes (\( C \): clustering coefficient, \( l \): average path length)

It is evident that binary weight performs poorly as the detailed frequency information is discarded and an author is considered either associated or not associated with a tag.
The resulting one-mode network is extremely dense, as demonstrated by the large value of average degree and small value of average path length.

Applying standard TF weight leads to a substantial improvement in terms of generating a network with large modularity and therefore strong community structure. Fully-weighted TF-IDF performs even better and is by far the best term weighting scheme, other conditions being equal. The resulting network is represented by a reasonably large average path length, implying the network is larger in diameter and boasts a larger number of communities.

Figure 5.1 plots the change of modularity value in relation to the filtering coefficient when fully weighted TF-IDF is used, which demonstrates that modularity stays consistently above 0.5 for a wide range of filtering coefficient (from about 0.2 to 0.75). And the range where modularity peaks (from about 0.5 to 0.7) is roughly overlapping with the range where the average path length is significantly large, as shown in figure 4.13. It seems that longer average path length tends to lead to larger modularity, although networks with small diameter may well exhibit strong community structure.

Figure 5.1: Dependence of the modularity value on the filtering coefficient
Figure 5.2 plots the number of pair-wise similarity values in each of the intervals of 0.01 from 0.01 to 1.0, for the similarity matrix generated using standard TF weight as the term weighting scheme, and the one generated using fully weighted TF-IDF.

Both plots exhibit a power law distribution, but the similarity matrix from simple normalization have more entries for any $\epsilon$ larger than 0.05, making its corresponding network significantly denser. The steeper line formed from the “fully weighted TF-IDF” also indicates that edges will be filtered out at a higher rate as the filtering coefficient $\epsilon$ increases, making the salient community structures easily stand out.

To get an idea of the communities formed through similarity measure based on standard TF weight vs fully weighted TF-IDF, the networks of top authors tagged as “classic” are obtained. The parameters of the two networks are shown in table 5.2.

Both networks are generated through first-order similarity measure and $\epsilon$-neighborhood graph. The network constructed using fully weighted TF-IDF boasts a substantially greater modularity value. If both networks are visualized with nodes representing authors and color
Table 5.2: Characteristics of networks generated using different term weighting schemes

(C: clustering coefficient, l: average path length)

<table>
<thead>
<tr>
<th>scheme</th>
<th>average degree</th>
<th>C</th>
<th>l</th>
<th>modularity</th>
<th>number of communities</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard TF weight</td>
<td>26.2</td>
<td>0.6920</td>
<td>2.6938</td>
<td>0.1828</td>
<td>5</td>
</tr>
<tr>
<td>fully weighted TF-IDF</td>
<td>7.9835</td>
<td>0.6698</td>
<td>7.3707</td>
<td>0.7675</td>
<td>7</td>
</tr>
</tbody>
</table>

of a node representing its community, the difference is even more obvious, as demonstrated in figure 5.3 and 5.4.

Figure 5.3: Standard TF weight

Figure 5.4: Fully weighted TF-IDF

Even though the number of communities detected in both networks are close, the fully-weighted network is structured such that all the communities are visually discernible and clearly separated. In the standard TF network, on the other hand, a handful of communities are intertwined with each other and are difficult for human readers to interpret.

In the case of first-order co-occurrence, standard TF weight yields cosine similarity. Based on the relatively low modularity of the generated one-mode networks, it is clearly seen that cosine similarity is not the preferred way to calculate a co-occurrence network that is targeted at revealing community structures in the network. In constrast fully weighted TF-IDF leads to networks with significantly boosted modularity value and should be used whenever possible.

The key role of TF-IDF weighting in co-occurrence calculation demonstrates that among
all the tags applied to an item or an author, semantic information is mostly conveyed by
those with heavy weight (with respect to TF-IDF), similar to keywords used in text retrieval.
Properly distinguishing these tags from the rest is crucial to grouping similar items/authors
together and separating disparate items/authors apart.

5.1.2 Similarity Calculation Methods

As the fully weighted TF-IDF weighting has proved best performing among different term
weighting schemes, it will then be used in subsequent evaluation of different methods.

The parameters of networks generated using six different similarity calculation methods
and the ϵ-neighborhood network construction are summarized in table 5.3. As previously
specified all the networks (largest connected components) contain about 2,500 authors.

<table>
<thead>
<tr>
<th>method</th>
<th>average degree</th>
<th>C</th>
<th>l</th>
<th>modularity</th>
<th>number of communities</th>
</tr>
</thead>
<tbody>
<tr>
<td>first-order</td>
<td>26.5987</td>
<td>0.7251</td>
<td>9.0484</td>
<td>0.8610</td>
<td>38</td>
</tr>
<tr>
<td>second-order</td>
<td>54.4634</td>
<td>0.7492</td>
<td>9.6596</td>
<td>0.8109</td>
<td>26</td>
</tr>
<tr>
<td>Tanimoto</td>
<td>26.5409</td>
<td>0.7249</td>
<td>9.0617</td>
<td>0.8611</td>
<td>38</td>
</tr>
<tr>
<td>LSI</td>
<td>29.8258</td>
<td>0.7653</td>
<td>8.2401</td>
<td>0.8607</td>
<td>30</td>
</tr>
<tr>
<td>NMF</td>
<td>20.3734</td>
<td>0.7075</td>
<td>5.4941</td>
<td>0.8004</td>
<td>24</td>
</tr>
<tr>
<td>CF</td>
<td>17.896</td>
<td>0.7248</td>
<td>5.8418</td>
<td>0.8217</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 5.3: Characteristics of networks generated using different similarity calculation
(C: clustering coefficient, l: average path length)

The differences between the modularity of the networks are relatively insignificant. Both
first-order and Tanimoto co-occurrence have shown to yield networks with large modularity,
suggesting that simple first-order co-occurrence is sufficient in revealing community struc-
tures among the nodes under study. While LSI, NMF and CF, despite their overhead in
running time, have not proven superior in terms of generating networks with a higher value
of modularity or a larger number of communities being detected.
5.1.3 Network Construction Methods

Networks constructed using KNN method have demonstrated marginally larger modularity and number of communities detected, which is unsurprising, as keeping the top $k$ edges of a node and removing the others will always lead to networks with fewer edges but more strongly connected groups or even cliques.

Characteristics of $\epsilon$-neighborhood network and KNN network based on fully weighted TF-IDF and first-order co-occurrence are summarized in table 5.4.

<table>
<thead>
<tr>
<th>method</th>
<th>average degree</th>
<th>$C$</th>
<th>$l$</th>
<th>modularity</th>
<th>number of communities</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon$-neighborhood</td>
<td>26.5987</td>
<td>0.7251</td>
<td>9.0484</td>
<td>0.8610</td>
<td>38</td>
</tr>
<tr>
<td>KNN</td>
<td>7.2494</td>
<td>0.5613</td>
<td>10.3042</td>
<td>0.9087</td>
<td>48</td>
</tr>
</tbody>
</table>

Table 5.4: Characteristics of networks generated using different network construction ($C$: clustering coefficient, $l$: average path length)

The same trend has been observed in networks based on other similarity calculation methods. Network modularity has been improved to over 0.90 for KNN network from about 0.85 for $\epsilon$-neighborhood network, accompanied by a significantly larger number of communities. This is also evidenced by a noticeably increased average path length as KNN networks are more dispersed and characterized by smaller groups or cliques and therefore it takes a longer path to move from one node to another.

This also agrees with the practice in spectral clustering, from which the $\epsilon$-neighborhood graph and KNN graph originated, that the KNN graph should be used as the first choice, the reason being it is simple to work with, results in a sparse adjacency matrix, and is less vulnerable to unsuitable choices of parameters [94].

Table 5.5, figure 5.5 and 5.6 compare the KNN network and $\epsilon$-neighborhood network of top authors tagged as “classic”.

In KNN graphs, the number of incident edges of each node is limited to $k$, which explains the significantly smaller average degree, accompanied by a longer average path length. Its larger diameter is also evidenced in the network visualization. The modularity of KNN
network is marginally better and more communities are detected. This makes KNN graph suitable for finding a larger number of small- to medium-sized groups.

Based on a wide range of empirical studies, Newman suggests that values of modularity above 0.3 indicate a strong community structure for the given network \cite{96}. The networks generated in this study, therefore, exhibit very strong community structure, given modularity values as high as 0.8-0.9.

Fine tuning different parameters of the network building process using this data set has achieved better results than previously reported. Robu et al., for example, experimented with network construction and community detection using a data set from the del.icio.us social bookmarking site and have obtained modularity value between 0.2 and 0.6 on networks of significantly smaller scale (50 tags) \cite{110}. Results from the current study not only reveal the strong semantic structures of networks of tags/authors in the social cataloging world as hypothesized in previous sections, but also demonstrate that the selection of proper term

<table>
<thead>
<tr>
<th>scheme</th>
<th>average degree</th>
<th>$C$</th>
<th>$l$</th>
<th>modularity</th>
<th>number of communities</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>3.3273</td>
<td>0.5006</td>
<td>10.3776</td>
<td>0.8384</td>
<td>12</td>
</tr>
<tr>
<td>$\epsilon$-neighborhood</td>
<td>7.9835</td>
<td>0.6698</td>
<td>7.3707</td>
<td>0.7675</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 5.5: Characteristics of KNN network and $\epsilon$-neighborhood network ($C$: clustering coefficient, $l$: average path length)
weighting schemes and network construction methods can make a difference in the generated networks and therefore the semantic structures being studied.

As first-order co-occurrence calculation combined with fully weighted TF-IDF weighting have proved sufficient in detecting community structures in the one-mode networks, they will then be used by default in subsequent evaluation, and particularly in generating visualizations for the user study.

5.2 Closer Examination of Results from Community Detection

After analysis of overall network characteristics that reflect the “big picture” of network construction, it will be interesting to closely examine communities detected in smaller networks to gain more insights into the details of semantic structures. Moreover, as there is no community metadata associated with the tags and authors on the social cataloging site to which the inferred communities can be compared, this first-hand intuitive interpretation of the obtained results becomes necessary.

5.2.1 Cliques and Average Clustering Coefficient

Community detection algorithms are intended to make network divisions such that the modularity value is maximized and they have yielded satisfactory outcome in theory and practicality. It would be premature, however, to assume that the communities generated by those algorithms are homogeneous in topology and structure. Some communities may happen to be cliques where each node is connected to all the rest of the community, while some communities may just be a handful of loosely connected nodes whose merging into a community slightly increases the value of objective function - modularity. A clique can be a group of close friends in a social network, a research group in a co-authorship network, or items of
the same category in an Amazon purchasing network. The members of a clique will likely bear stronger relationship than would the members of an average community. Clique has been an important concept in complex network analysis, and is also significant in network visualization, as cliques usually stand out in a network visualization to the extent that proper layout method is applied. To further evaluate the coherence of nodes in a community or the degree to which nodes in a community tend to cluster together, the concept of average clustering coefficient of a community is introduced. The average clustering coefficient of a community $C_k$, is defined as the average of the clustering coefficient of its member nodes, or

$$ C_{c_k} = \frac{1}{n} \sum_{i=1}^{n} C_i : v_i \in V_{C_k}, n = |V_{C_k}| $$

where $C_i$ is the clustering coefficient of node $v_i$ as defined in section 2.3.2 and $V_{C_k}$ is the set of all nodes in community $C_k$. The average clustering coefficient of a community (or any component of a network) takes a value between 0 and 1. The larger the average clustering coefficient, the more closely the nodes in the community cluster together, and the closer the community is to a clique.

5.2.2 Two Author Network Examples

A list of top 250 authors tagged as “classic” is drawn from the data set. The similarity matrix is obtained by calculating the first-order co-occurrences of the authors. The matrix is then used to construct a KNN network, of which the largest connected component contains 161 nodes. The fast community detection algorithm is then applied to this largest connected component to identify community structures.
Table 5.6 shows the network parameters of the largest connected component:

<table>
<thead>
<tr>
<th>method</th>
<th>average degree</th>
<th>$C$</th>
<th>$l$</th>
<th>modularity</th>
<th>number of communities</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>3.8634</td>
<td>0.4886</td>
<td>8.8592</td>
<td>0.8561</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 5.6: Characteristics of networks generated using the “classic” author data set ($C$: clustering coefficient, $l$: average path length)

Table 5.7 gives the output of the community detection algorithm. The communities are listed in descending order of average clustering coefficient.

Most of the groups contain authors of the same country, genre, or historic period, especially those with a high average clustering coefficient. For example, authors in group 1:

Marcus Aurelius, Friedrich Nietzsche, Aristotle, John Stuart Mill, Thomas More, Niccolo Machiavelli, Michel de Montaigne, Ralph Waldo Emerson

...can be labelled “philosophers”. While group 2 consists of well recognized writers of the German language:

E. T. A. Hoffmann, Theodor Fontane, Johann Wolfgang von Goethe, Friedrich Dörrenmatt, Bertolt Brecht, Thomas Mann, Hermann Hesse.

Group 3 consists of Victorian writers:

Elizabeth Gaskell, Emily Bronte, R. D. Blackmore, Charlotte Bronte, Thomas Hardy, Anne Bronte, George Eliot, George Meredith.

The genre of authors in group 4 are less coherent:

<table>
<thead>
<tr>
<th>#</th>
<th>group</th>
<th>clustering coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Marcus Aurelius, Friedrich Nietzsche, Aristotle, John Stuart Mill, Thomas More, Niccolo Machiavelli, Michel de Montaigne, Ralph Waldo Emerson</td>
<td>0.7952</td>
</tr>
<tr>
<td>2</td>
<td>E. T. A. Hoffmann, Theodor Fontane, Johann Wolfgang von Goethe, Friedrich Döremmatt, Bertolt Brecht, Thomas Mann, Hermann Hesse</td>
<td>0.7810</td>
</tr>
<tr>
<td>3</td>
<td>Elizabeth Gaskell, Emily Bronte, R. D. Blackmore, Charlotte Bronte, Thomas Hardy, Anne Bronte, George Eliot, George Meredith</td>
<td>0.7292</td>
</tr>
<tr>
<td>4</td>
<td>Clement Clarke Moore, Ludwig Bemelmans, H. A. Rey, Virginia Lee Burton, Robert McCloskey, Margaret Wise Brown, Don Freeman, Michael Bond</td>
<td>0.6333</td>
</tr>
<tr>
<td>5</td>
<td>Oliver Goldsmith, Samuel Beckett, Moliere, Eugene O'Neill, Jerome Lawrence, Christopher Marlowe, Lorraine Hansberry, Henrik Ibsen, Arthur Miller</td>
<td>0.6111</td>
</tr>
<tr>
<td></td>
<td>Jacob Grimm, Carlo Collodi, Andrew Lang, Aesop, Peter S. Beagle, Lloyd Alexander, E. Nesbit, Norton Juster, Mary Norton, Roald Dahl, Kenneth Grahame, Frances Hodgson Burnett, Charles Kingsley, Susan Coolidge, Maud Hart Lovelace, Michael Ende, Lord Dunsany, William Goldman, George MacDonald</td>
<td>0.6000</td>
</tr>
<tr>
<td>6</td>
<td>Anais Nin, Henry Miller, John Cleland, Anne Frank, John Knowles, Robert Cormier, S. E. Hinton, Aldous Huxley, William Golding, Anthony Burgess, Lois Lowry</td>
<td>0.6000</td>
</tr>
<tr>
<td></td>
<td>Sinclair Lewis, Saul Bellow, Theodore Dreiser, F. Scott Fitzgerald, Ernest Hemingway, Nathaniel Hawthorne, Henry James, James Fenimore Cooper, Herman Melville, Washington Irving, Frank B. Gilbreth, Zora Neale Hurston, Toni Morrison, Maya Angelou, Ralph Ellison, Frederick Douglass, Benjamin Franklin, Charles W. Eliot, Richard Henry Dana</td>
<td>0.5556</td>
</tr>
<tr>
<td>7</td>
<td>Herodotus, Julius Caesar, Marcus Tullius Cicero, Apuleius, Roger Lancelyn Green, Anonymous, Thomas Bulfinch, Edith Hamilton, Robert Graves, Edward Gibbon, Nikos Kazantzakis, Homer, Will Durant, Alexander Hamilton</td>
<td>0.5250</td>
</tr>
<tr>
<td>9</td>
<td>Henry Fielding, Samuel Richardson, Fanny Burney, Daniel Defoe, Samuel Johnson, Anthony Hope, H. Rider Haggard, Jack London, Fred Gipson, Jean Craighead George, Scott O’Dell, Hugh Lofting, James Herriot</td>
<td>0.4143</td>
</tr>
</tbody>
</table>

Table 5.7: Output of the community detection algorithm on “classic” authors

Carolyn Keene, Daphne Du Maurier

Most of them are thriller writers except Thomas a Kempis, John Bunyan, Dietrich Bonhoeffer and Jonathan Edwards, who are Christian writers or theologians. As can be seen from the visualization in figure 5.7, these four indeed form a clique connected with the rest of the community through G. K. Chesterton, whose diverse output included Christian apologetics, fantasy and detective fiction, thus connecting the two parts. It seems that although the community detection algorithm fails to separate the two smaller groups (suspense and Christian) as the split does not result in an increase in modularity, they are still evident in
the visualization and can be easily spotted by human readers.

![Network Visualization Diagram](image)

Figure 5.7: The “thriller writer” community in the network visualization

Authors and illustrators of children’s books dominate group 5:

Clement Clarke Moore, Ludwig Bemelmans, H. A. Rey, Virginia Lee Burton, Robert McCloskey, Margaret Wise Brown, Don Freeman, Michael Bond

Authors in group 6:


are all playwrights.
Authors in group 7:

Jacob Grimm, Carlo Collodi, Andrew Lang, Aesop, Peter S. Beagle, Lloyd Alexander, E. Nesbit, Norton Juster, Mary Norton, Roald Dahl, Kenneth Grahame, Frances Hodgson Burnett, Charles Kingsley, Susan Coolidge, Maud Hart Lovelace, Michael Ende, Lord Dunsany, William Goldman, George MacDonald

are all fantasists or writers of children’s literature with a flavor of fantasy.

Authors in group 8:

Anais Nin, Henry Miller, John Cleland, Anne Frank, John Knowles, Robert Cormier, S. E. Hinton, Aldous Huxley, William Golding, Anthony Burgess, Lois Lowry

are roughly a combination of writers of dystopia and young adult literature, as demonstrated by separate cliques in figure 5.8

Authors in group 9:

Sinclair Lewis, Saul Bellow, Theodore Dreiser, F. Scott Fitzgerald, Ernest Hemingway, Nathaniel Hawthorne, Henry James, James Fenimore Cooper, Herman Melville, Washington Irving,

are mostly 19th century to 20th century American writers that are either Nobel prize winners or have won worldwide recognition. A handful of them are somewhat related to Europe. For example, F. Scott Fitzgerald and Ernest Hemingway are strongly connected to the “lost generation” in Europe; James Fenimore Cooper and Washington Irving are among the first American writers to earn acclaim in Europe, while Nathaniel Hawthorne and Herman Melville were encouraged by Washington Irving; Henry James has been primarily known for the series of novels in which he portrays the encounter of Americans with Europe.
Authors in group 10:

Frank B. Gilbreth, Zora Neale Hurston, Toni Morrison, Maya Angelou, Ralph Ellison, Frederick Douglass, Benjamin Franklin, Charles W. Eliot, Richard Henry Dana

is dominated by a number of African American novelists, along with several authors that are more closely connected to other groups, as shown in figure 5.9.

Particularly interesting is group 11:

Herodotus, Julius Caesar, Marcus Tullius Cicero, Apuleius, Roger Lancelyn Green, Anonymous, Thomas Bulfinch, Edith Hamilton, Robert Graves, Edward Gibbon, Nikos Kazantzakis,
Homer, Will Durant, Alexander Hamilton

They are either classicists or subjects studied by classicists. This is an example of multiple senses of the same tag word (“classic”) and the ability of network analysis to identify them.

Group 12 is also interesting:

William Faulkner, Flannery O’Connor, Carson McCullers, Harper Lee, Edna Ferber, Margaret Mitchell, Stephen Crane

Among them Harper Lee, Carson McMullers, Flannery O’Connor, and William Faulkner are all Southern Gothic writers of America. This is demonstrated by the clique in figure 5.10. The connection between Harper Lee and Margaret Mitchell can be explained by the fact that both of them are women writers, and both of their best-known and Pulitzer-winning
works (To Kill a Mockingbird and Gone with the Wind) are centered in the south and have been adapted into widely acclaimed films. While the American Civil War as their common subject likely makes the connection between Margaret Mitchell and Stephen Crane.

Authors in group 13:

Jerome K. Jerome, Stella Gibbons, Ambrose Bierce, H. P. Lovecraft, Richard Matheson, Shirley Jackson, Gaston Leroux, Matthew Lewis

is characterized by several authors of the “horror” genre as depicted by the clique at the center of figure 5.11.

Authors in group 14 are somewhat diverse:

Henry Fielding, Samuel Richardson, Fanny Burney, Daniel Defoe, Samuel Johnson, Anthony
Figure 5.11: The “horror fiction” writers community in the network visualization

Hope, H. Rider Haggard, Jack London, Fred Gipson, Jean Craighead George, Scott O’Dell, Hugh Lofting, James Herriot

It is more clearly seen in figure 5.12 that Henry Fielding, Fanny Burney, Samuel Richardson and Daniel Defoe form a clique as they are all 17th century English novelists, while more recent writers such as Jack London and Scott O’Dell are only loosely connected to the clique.

While authors in group 15:

are fairly diverse except they are all 20th century novelists.

This examination strengthens the concept of “clique” as the entity that conveys close semantic relationships in the author network, or other kinds of interrelationships among nodes in social or biologic networks. The average clustering coefficient roughly reflects the extent to which the nodes in a community are related. An exception is group 4, where authors of two different genres (suspense and Christian) are in the same community, due to the community detection algorithm’s objective to maximize modularity value, instead of identifying cliques. The network visualization, however, through appropriate layout methods, can present the structure of the community, as the two cliques are clearly visible in the visualization in figure 5.7. In other words, even though community detection techniques have long been used in identifying communities in various kinds of networks, clique analysis should also be employed as a complementary tool to reveal more detailed clique structures.
which might be hidden in the output of community detection.

In the second example a less ambiguous tag, “fiction”, is chosen. The top authors tagged as “fiction” are drawn from the data set. The similarity matrix is obtained by calculating the first-order co-occurrences of the authors. The matrix is then used to construct a KNN network, of which the largest connected component contains 100 nodes.

Table 5.8 summarizes the network parameters of the largest connected component.

<table>
<thead>
<tr>
<th>method</th>
<th>average degree</th>
<th>$C$</th>
<th>$l$</th>
<th>modularity</th>
<th>number of communities</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>3.16</td>
<td>0.3953</td>
<td>8.9168</td>
<td>0.8122</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 5.8: Characteristics of networks generated using the “fiction” author data set ($C$: clustering coefficient, $l$: average path length)

Table 5.9 gives the output of the community detection algorithm along with the label that can be used to describe the authors in the group.

Again, as the average clustering coefficient drops, it becomes increasingly difficult to describe authors in the group using a single label. The groups whose average clustering coefficient is less than 0.5 (group 4 through group 11) are noticeably more diverse than the first three groups in their genre, nationality, topic of their works, etc.

It seems that in the projected one-mode network, authors do not necessarily form groups by one dimension (for example country of origin); similarity in other aspects can also draw authors together in the network. An interesting example is group 11 in which most of the authors’ work is part of Oprah’s book club selection (figure 5.13). This suggests that social tagging covers a wider scope than traditional classification system and the projected one-mode network is capable of capturing commonalities of tagged items in different aspects.
5.3 Summary

Strong community structures in one-mode author networks have been identified by both the community detection algorithm and intuitive close examination.

To generate networks with strong community structures as measured by large modularity value, fully-weighted TF-IDF should be used in preference to standard TF weighting, which leads to widely used cosine similarity. To the extent that fully-weighted TF-IDF weighting is applied, the effect of similarity calculation method is less significant as simple first-order co-occurrence calculation is adequate in generating networks with strong community structures. KNN-graph method can often generate networks with larger modularity value and number of communities, while $\epsilon$-neighborhood graph is also capable and more suitable for fine tuning the resulting one-mode network as $\epsilon$ can be continuously adjusted.

Table 5.9: Output of the community detection algorithm on “fiction” authors

<table>
<thead>
<tr>
<th>#</th>
<th>group</th>
<th>clustering coefficient</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Frank Miller, Alan Moore, Neil Gaiman, Warren Ellis, Michael Chabon</td>
<td>0.9200</td>
<td>comics</td>
</tr>
<tr>
<td>2</td>
<td>Margaret Mitchell, Geraldine Brooks, E.L. Doctorow, Charles Frazier,</td>
<td>0.8333</td>
<td>American historical fiction</td>
</tr>
<tr>
<td></td>
<td>Stephen Crane, John Jakes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Gordon Korman, L.M. Montgomery, Douglas Coupland, Michael Ondaatje,</td>
<td>0.6429</td>
<td>Canadian fiction</td>
</tr>
<tr>
<td></td>
<td>Alice Munro, Robertson Davies, Margaret Atwood</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>W. E. B. Griffin, Bernard Cornwell, Patrick O'Brian, Alexander Kent,</td>
<td>0.4583</td>
<td>historical fiction</td>
</tr>
<tr>
<td></td>
<td>C.S. Forester, Philippa Gregory, James A. Michener, Dorothy Dunnett,</td>
<td></td>
<td>science fiction</td>
</tr>
<tr>
<td></td>
<td>Audrey Niffenegger, Diana Gabaldon, Iain Banks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>E.L. Konigsburg, Lois Lowry, Scott O'Dell, Avi, Judy Blume,</td>
<td>0.4524</td>
<td>American children</td>
</tr>
<tr>
<td></td>
<td>Kate DiCamillo, Jack London, Jean Craighead George</td>
<td></td>
<td>young adults</td>
</tr>
<tr>
<td>6</td>
<td>John Barth, Don DeLillo, Jonathan Safran Foer, Thomas Keneally,</td>
<td>0.3667</td>
<td>historical fiction</td>
</tr>
<tr>
<td></td>
<td>Jerzy Kosinski, Ursula Hegi, Irene Nemirovsky, Joanne Harris,</td>
<td></td>
<td>holocaust</td>
</tr>
<tr>
<td></td>
<td>Nancy Mitford, Peter Carey, Colleen McCullough, Robert Harris,</td>
<td></td>
<td>postmodern</td>
</tr>
<tr>
<td></td>
<td>Lindsey Davis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Charles Bukowski, Seamus Heaney, Geoffrey Chaucer, Anonymous,</td>
<td>0.3615</td>
<td>ancient Greece</td>
</tr>
<tr>
<td></td>
<td>Robert Graves, Homer, Louis de Bernieres, Mary Renault,</td>
<td></td>
<td>mythology</td>
</tr>
<tr>
<td></td>
<td>Nikos Kazantzakis, Ken Kesey, Jack Kerouac, William S. Burroughs,</td>
<td></td>
<td>poetry</td>
</tr>
<tr>
<td></td>
<td>Richard Brautigan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Sarah Dunant, Donna Leon, Tracy Chevalier, Nick Bantock,</td>
<td>0.3286</td>
<td>Italy</td>
</tr>
<tr>
<td></td>
<td>Edward Gorey, Giovanni Boccaccio, Italo Calvino, Umberto Eco,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dante Alighieri</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Robert Cormier, Francesca Lia Block, Ann Brashares, Meg Cabot,</td>
<td>0.3111</td>
<td>American young adult</td>
</tr>
<tr>
<td></td>
<td>S. E. Hinton, John Knowles, Joseph Heller</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Sinclair Lewis, Saul Bellow, Bernard Malamud, F. Scott Fitzgerald,</td>
<td>0.3083</td>
<td>American</td>
</tr>
<tr>
<td></td>
<td>Nathaniel Hawthorne, Herman Melville, Henry James,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>James Fenimore Cooper, Theodore Dreiser, Raymond Carver,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>O. Henry, Washington Irving</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Chris Bohjalian, Wally Lamb, Jane Hamilton, Janet Fitch,</td>
<td>0.2857</td>
<td>Oprah’s book club</td>
</tr>
<tr>
<td></td>
<td>Joyce Carol Oates, Richard Ford, Andre Dubus, Elizabeth Berg,</td>
<td></td>
<td>American short stories</td>
</tr>
<tr>
<td></td>
<td>Sue Miller</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.9: Output of the community detection algorithm on “fiction” authors
Figure 5.13: The “Oprah’s book club” community in the network visualization

Close examination of the one-mode networks and the detected communities suggests that the communities arguably correspond to groups of authors of similar or related genre, country of origin or historical period. Average clustering coefficient can be used to approximate the coherence of a group. And it can be shown that cliques in networks convey stronger connections of members - this kind of information may be hidden in the output of community detection algorithms but can be easily spotted in a network visualization.

After network analysis and basic examination of the networks and the detected communities, a user study is expected to evaluate the effect of network-based visualizations from users’ perspective. This is the subject of the next chapter.
Chapter 6

User Study

As a popular method for visualizing and linking socially-organized information on websites, tag clouds have been used extensively in various contexts, while tag networks offer an alternative. These visualizations attempt to represent a wide range of variables with various visual properties, making it difficult to predict what will appear visually important to a viewer. A user experiment was therefore carried out to address the question as to whether network-based visualizations can foster a viewer’s visual experience compared to cloud-based visualizations.

Comparative in nature, this study is intended to closely examine users’ responses to the two different methods (network- and cloud-based visualizations) and further our understanding of their strengths and weaknesses. The procedure of the user study is described in detail in chapter 3. As a step further from the analysis performed in the last two chapters, this user study is intended to be concentrating on aspects that are less emphasized in previously published studies on tag clouds, for example, the groups and relations found in visualizations and their effect on impression formation.

The main measure in the study, therefore, was the degree to which the visualizations facilitate a participant’s ability to group similar items together and to detect relations between items. Each participant was asked to group items by any criteria of his/her choosing in a limited period of time (3 minutes). The number of groups identified, the number of items identified as well as the participants’ ratings of computer-generated groups would be evaluated in statistical tests.

This study, albeit small in scale, has reinforced some of the points made in previous
chapters, suggesting that although tag and author clouds are clear and visually arresting, tag and author networks can better help a viewer extract information about groups and relations and thus have positive consequence on impression formation, judging from the number of authors identified by participants in limited-time grouping exercises and their responses to interview questions at the end of each session.

6.1 Introduction

Ten Northwestern undergraduate students from the College of Arts and Sciences were recruited for this study. All the participants have normal or corrected-to-normal vision. The study was conducted in a 3-week period from September to October 2010. A session for each participant took about an hour. Each session started with several grouping tasks when the participant was viewing the presentations, followed by the group rating task, and concluded with an interview which was recorded with the participant’s consent.

When asked the question

- On a scale of 1 to 10 (10 being “very good”), how do you rate your knowledge of English and American literature?

the responses of the participants are shown in table 6.1

| 7 | 10 | 8 | 6 | 9 | 6 | 6 | 7 | 6 |

Table 6.1: Rating of familiarity with literature

In general the participants were modest in rating their familiarity with literature. This rating, however, may not be taken as an objective measure of their knowledge in literature.

When asked the question

- Have you ever seen a tag cloud?
6 responded seeing a tag cloud on some website while 1 responded seeing a tag cloud in class. The other 3 participants have not seen a tag cloud before.

6.2 Apparatus

As described in chapter 3, the subjects were presented with two cloud visualizations for author group A1, two cloud visualizations for author group A2, and two network visualizations generated from author group A1, A2, respectively, using two different network construction methods (\(\epsilon\)-neighborhood and KNN). Analysis from chapter 5 suggests that fully-weighted TF-IDF performs the best among different weighting schemes while first-order co-occurrence will suffice in revealing community structures among the nodes in networks, so all the networks used in the user study are generated from first-order co-occurrence calculation whereby fully-weighted TF-IDF is used for normalization.

In summary, each participant will be presented with 4 clouds and 4 networks generated from 2 groups of authors (A1 and A2), as summarized in table 6.2.

<table>
<thead>
<tr>
<th></th>
<th>(\epsilon)-neighborhood method</th>
<th>KNN method</th>
</tr>
</thead>
<tbody>
<tr>
<td>author group A1</td>
<td>cloud/network 1</td>
<td>cloud/network 2</td>
</tr>
<tr>
<td>author group A2</td>
<td>cloud/network 3</td>
<td>cloud/network 4</td>
</tr>
</tbody>
</table>

Table 6.2: Configuration of apparatus

Note that the set of authors in the two generated networks (using \(\epsilon\)-neighborhood and KNN, respectively) from the same author group (for example network 1 and 2) may not necessarily be identical, as is the case for the set of authors in the clouds. However, the set of authors in the paired cloud and network generated from the same network construction method on the same author group are always identical. For example, in table 6.2 the set of authors in cloud 1 and network 1 are identical, the set of authors in cloud 2 and network 2 are identical, so on and so forth. While the set of authors in cloud/network 1 are not identical with the set of authors in cloud/network 2.
As $A_1$ is a fixed set of authors while $A_2$ varies from participant to participant, the statistical tests are only applied to data gathered when the participants looked at different visualizations of $A_1$. The results of the tests are analyzed in the following sections.

Moreover, as the authors in visualizations generated from author group $A_2$ are more specifically related to a particular author or subject, they tend to be less popular and therefore, most subjects could not recognize an adequate number of authors in those visualizations to make sense of the visualizations. The results from those generated from $A_2$ will not be used in qualitative analysis.

The four different visualizations of author group $A_1$ viewed in common by all the participants are presented in appendix B.

### 6.3 Number of Groups Identified

For the task of identifying groups in the following four different visualizations:

- author cloud based on the $\epsilon$-neighborhood method (cloud 1)
- author network based on the $\epsilon$-neighborhood method (network 1)
- author cloud based on the KNN method (cloud 2)
- author network based on the KNN method (network 2)

the number of groups identified by each participant is summarized in table 6.3.

Applying paired t-test to the number of groups identified when viewing the visualization generated from the $\epsilon$-neighborhood method (table 6.4), the two-tailed P value equals 0.7304. The mean of first group minus the second group equals -0.30. And the 95% confidence interval of this difference is $(-2.21, 1.61)$.

Although the participants identified 0.3 more groups when viewing the author network than when viewing the author cloud, by conventional criteria (at significance level of 0.05), this difference is considered to be not statistically significant.
Applying paired t-test to the number of groups identified when viewing the visualization generated from the KNN method (table 6.5), the two-tailed P value equals 0.8793. The mean of first group minus the second group equals -0.10. And the 95% confidence interval of this difference (-1.55, 1.35).

Although the participants identified 0.1 more groups when viewing the author network than when viewing the author cloud, still this difference is considered to be not statistically significant.

To analyze the interaction between network construction method (ϵ-neighborhood vs KNN) and visualization method (cloud vs network), a two-factor ANOVA with repeated measures was conducted. The result is summarized in table 6.6.
Table 6.6: Two-way ANOVA with repeated measures. A: network construction method (ε-neighborhood vs KNN), B: visualization method (cloud vs network)

The ANOVA summary table shows that by conventional criteria, none of the effect of visualization method (cloud vs network), the effect of network construction method (ε-neighborhood vs KNN), or the interaction effect between the two is statistically significant.

In other words, the subjects could not identify significantly more groups when viewing a display generated from any particular kind of visualization method (cloud vs network), or any particular kind of network construction method (ε-neighborhood vs KNN), or any combination of the two.

6.4 Number of Authors Identified

For the same task, the number of authors identified by each participant is summarized in table 6.7.

Applying paired t-test to the number of authors identified when viewing the visualization generated from the ε-neighborhood method (table 6.8), the two-tailed P value equals 0.0593. The mean of first group minus the second group equals -5.70. And the 95% confidence interval of this difference (−11.68, 0.28).

Although the participants identified 5.7 more groups when viewing the author network than when viewing the author cloud, by conventional criteria (at the 0.05 significance level),
this difference is considered to be not quite statistically significant. While this difference can be considered to be statistically significant at the 0.10 significance level.

<table>
<thead>
<tr>
<th></th>
<th>cloud/ε-neighborhood</th>
<th>network/ε-neighborhood</th>
<th>cloud/knn</th>
<th>network/knn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>19</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>20</td>
<td>28</td>
<td>35</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>32</td>
<td>24</td>
<td>42</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>22</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>26</td>
<td>34</td>
<td>22</td>
<td>34</td>
</tr>
<tr>
<td>6</td>
<td>38</td>
<td>31</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>19</td>
<td>21</td>
<td>26</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>40</td>
<td>34</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>9</td>
<td>15</td>
<td>37</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>10</td>
<td>21</td>
<td>32</td>
<td>7</td>
<td>24</td>
</tr>
</tbody>
</table>

Table 6.7: Number of authors identified

Applying paired t-test to the number of authors identified when viewing the visualization generated from the KNN method (table 6.9), the two-tailed P value equals 0.0801. The mean of first group minus the second group equals -5.30. And the 95% confidence interval of this difference (-11.38, 0.78).

Although the participants identified 5.3 more authors when viewing the author network than when viewing the author cloud, still this difference is considered to be not quite statistically significant by conventional criteria. While this difference can be considered to be statistically significant at the 0.10 significance level.

<table>
<thead>
<tr>
<th></th>
<th>cloud</th>
<th>12</th>
<th>13</th>
<th>23</th>
<th>18</th>
<th>26</th>
<th>38</th>
<th>19</th>
<th>40</th>
<th>15</th>
<th>21</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>network</td>
<td>19</td>
<td>20</td>
<td>32</td>
<td>22</td>
<td>34</td>
<td>31</td>
<td>21</td>
<td>34</td>
<td>37</td>
<td>32</td>
</tr>
</tbody>
</table>

Table 6.8: Number of authors identified (ε-neighborhood)

To analyze the interaction between network construction method (ε-neighborhood vs
KNN) and visualization method (cloud vs network), a two-factor ANOVA with repeated measures was conducted. The result is summarized in Table 6.10.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects</td>
<td>741.9</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>144.4</td>
<td>1</td>
<td>144.4</td>
<td>1.19</td>
<td>0.303672</td>
</tr>
<tr>
<td>Subject × A</td>
<td>1087.6</td>
<td>9</td>
<td>120.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>360</td>
<td>1</td>
<td>360</td>
<td>6.1</td>
<td>0.035581</td>
</tr>
<tr>
<td>Subject × B</td>
<td>531</td>
<td>9</td>
<td>59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A × B</td>
<td>0.9</td>
<td>1</td>
<td>0.9</td>
<td>0.06</td>
<td>0.811990</td>
</tr>
<tr>
<td>Subject × A × B</td>
<td>140.1</td>
<td>9</td>
<td>15.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3005.9</td>
<td>39</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.10: Two-way ANOVA with repeated measures. A: network construction method (\(\epsilon\)-neighborhood vs KNN), B: visualization method (cloud vs network)

The ANOVA summary table shows that by conventional criteria, the effect of visualization method (cloud vs network) is statistically significant (P value is 0.035581), while the effect of network construction method (\(\epsilon\)-neighborhood vs KNN) and the interaction effect between network construction method and visualization method are not statistically significant.

In other words, the subjects could identify significantly more authors when viewing network visualizations than they could when viewing cloud visualizations, irrespective of the network construction method being \(\epsilon\)-neighborhood or KNN.

### 6.5 Closer Examination of Results from the Grouping Task

To gain deeper understanding of the results from the grouping task when the participants were viewing different visualizations, the identified groups and authors are further examined.

Some participants identified groups with exactly the same labels while viewing the cloud visualization and the network visualization (containing the same set of authors), as shown
Although the participant identified the same five groups, none of the groups in table 6.11 overlap with their counterparts in table 6.12. In general, size of groups in table 6.12 is larger. Specifically, this participant identified more authors in 3 (poet, female American and myth) out of the 5 groups when viewing the network visualization. In the case of group 3 female American, the participant identified 13 authors when viewing the network visualization vs 4 authors when viewing the cloud visualization. This indicates that the community structures in network visualizations were appreciated by some participants so they were able to spot a larger number of closely connected authors in the groups that they had in mind.

Participant #9’s responses present a different picture. It seems that when performing the grouping tasks on different visualizations of the same set of authors, not only the identified groups can be different, but also the authors in groups with the same label are not necessarily overlapping.

When participant #9 was viewing cloud 2, four groups were identified: mystery, African American, children’s and adventure (table 6.13). But when the participant was viewing
Table 6.13: Result from cloud 2 (participant #9)

<table>
<thead>
<tr>
<th>#</th>
<th>group</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Agatha Christie, Dashiell Hammett, Arthur Conan Doyle, Carolyn Keene, Raymond Chandler</td>
<td>mystery</td>
</tr>
<tr>
<td>3</td>
<td>Lloyd Alexander, Roald Dahl</td>
<td>children’s</td>
</tr>
<tr>
<td>4</td>
<td>Thomas Mann, John Stuart Mill, Friedrich Nietzsche, Aristotle, Marcus Aurelius</td>
<td>philosopher</td>
</tr>
</tbody>
</table>

Table 6.14: Result from network 2 (participant #9)

network 2 (containing the same set of authors as cloud 2), a mystery group containing the same set of authors was identified, a children’s group was also identified (with only 1 author - Lloyd Alexander - overlapping with the other children’s group). But the group African American and adventure were replaced with American novelist and philosopher (table 6.14). Examining network 2 suggests that authors in either of the American novelist and philosopher group lie close together or even in a clique in the network. In other words, the response of the participant on a network visualization was strongly guided by community structures in the network such that the participant may identify completely different groups than what he/she would otherwise identify when viewing a cloud visualization containing the same set of authors.

To summarize, even though the difference between the numbers of groups identified by participants when they were viewing network visualizations and the numbers of groups identified when they were viewing cloud visualizations is not statistically significant, the difference between the number of authors identified is statistically significant. While there is no significant interaction between the effect of visualization method and network construction method. In other words, the participants could not identify significantly more groups when viewing network visualizations than they could when viewing cloud visualizations, but they
could identify significantly more authors when viewing network visualizations, irrespective of the network construction method being $\epsilon$-neighborhood or KNN. Close examination of the identified groups suggests that participants may identify different groups when viewing network visualizations and when viewing cloud visualizations, as they were strongly guided by community structures in network visualizations and the geographic proximity entailed by those structures helps them identify authors that they might not otherwise have grouped together. In the case that they do identify the same groups, community structures can help them spot more authors that are related and include them in the groups.

6.6 Rating of Groups

When asked to rate the groups generated by the community detection algorithm (table 5.7), the responses of the subjects are summarized in table 6.15.

<table>
<thead>
<tr>
<th>group</th>
<th>cc</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
<th>#6</th>
<th>#7</th>
<th>#8</th>
<th>#9</th>
<th>#10</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7952</td>
<td>3</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td>9</td>
<td>3</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>7.5</td>
</tr>
<tr>
<td>2</td>
<td>0.7810</td>
<td>9</td>
<td>5</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>4</td>
<td>3</td>
<td>9</td>
<td>9</td>
<td>6.9</td>
</tr>
<tr>
<td>3</td>
<td>0.7292</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>10</td>
<td>10</td>
<td>6.9</td>
</tr>
<tr>
<td>4</td>
<td>0.7212</td>
<td>4</td>
<td>9</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>10</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>5</td>
<td>6.3</td>
</tr>
<tr>
<td>5</td>
<td>0.6333</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4.9</td>
</tr>
<tr>
<td>6</td>
<td>0.6111</td>
<td>10</td>
<td>7</td>
<td>8</td>
<td>1</td>
<td>8</td>
<td>10</td>
<td>5</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>6.6</td>
</tr>
<tr>
<td>7</td>
<td>0.6</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>10</td>
<td>4.0</td>
</tr>
<tr>
<td>8</td>
<td>0.6</td>
<td>3</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.4</td>
</tr>
<tr>
<td>9</td>
<td>0.5556</td>
<td>9</td>
<td>9</td>
<td>6</td>
<td>9</td>
<td>6</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>10</td>
<td>10</td>
<td>8.1</td>
</tr>
<tr>
<td>10</td>
<td>0.5250</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>8</td>
<td>9</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>8</td>
<td>6.4</td>
</tr>
<tr>
<td>11</td>
<td>0.5071</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5.3</td>
</tr>
<tr>
<td>12</td>
<td>0.4400</td>
<td>7</td>
<td>7</td>
<td>10</td>
<td>7</td>
<td>8</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>10</td>
<td>7.9</td>
</tr>
<tr>
<td>13</td>
<td>0.4143</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>4.0</td>
</tr>
<tr>
<td>14</td>
<td>0.3444</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>8</td>
<td>5.2</td>
</tr>
<tr>
<td>15</td>
<td>0.1861</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>4.2</td>
</tr>
</tbody>
</table>

Table 6.15: Participants’ rating of the groups generated from community detection

The clustering coefficient of each group is copied in column 2. The last row gives the correlation coefficient between the clustering coefficient and each participant’s rating. To focus on the possible linear relationship between a participant’s response and the average clustering coefficient, the Pearson correlation coefficient as opposed to rank correlation coefficient is adopted here.
As can be seen from the table, all the ratings of the participants exhibit positive correlation with the clustering coefficients of the groups, except for the ratings of participant #8. The highest correlation coefficient 0.61261 is yielded by the ratings of participant #3, who had demonstrated extensive knowledge in various literary works during the session. The correlation coefficient between the clustering coefficients and the average ratings of all the participants is 0.40401, which indicates that the participants’ overall ratings are strongly correlated with the clustering coefficient values of the groups, of which the participants were not aware.

As a comparison, the correlation coefficient between the clustering coefficient values and 15 randomly generated numbers between 1 and 10:

3, 9, 5, 4, 7, 1, 6, 5, 4, 10, 7, 9, 6, 10, 5

is -0.2893.

Although correlation cannot be used to infer a causal relationship between the variables, and a value of 0.40401 is far from demonstrating a strictly linear relationship, the relatively strong correlation between human-generated ratings and computer-generated clustering coefficients is satisfactory. It is believed that a higher value may be achieved if the participants are more familiar with the subject matter, and the clustering coefficient of a group may therefore be employed as an objective measure of coherence of group members that is more or less in line with human perceptions. The analysis of average clustering coefficient and cliques can then be employed as another tool in complex network researchers’ toolbox in addition to regular community detection which treats all the communities equally.

Interestingly, the authors in the two groups with the highest average rating, namely group 9 and 12, are all American novelists:

- **group 9**: Sinclair Lewis, Saul Bellow, Theodore Dreiser, F. Scott Fitzgerald, Ernest Hemingway, Nathaniel Hawthorne, Henry James, James Fenimore Cooper, Herman Melville, Washington
Irving.


While group 2 which consists of well recognized writers of the German language:

E. T. A. Hoffmann, Theodor Fontane, Johann Wolfgang von Goethe, Friedrich Dörrenmatt, Bertolt Brecht, Thomas Mann, Hermann Hesse.

and group 5 which is dominated by authors and illustrators of mid-20th century children’s books:

Clement Clarke Moore, Ludwig Bemelmans, H. A. Rey, Virginia Lee Burton, Robert McCloskey, Margaret Wise Brown, Don Freeman, Michael Bond

received relatively low ratings. This shows that users tend to give higher ratings to the groups that they are more familiar with or confident about. In other words, even though participants may not appreciate some of the groups generated by automated methods, it can be partly accounted for by their lack of knowledge about the authors in those groups.

It will be beneficial, therefore, to control the group of subjects in future studies such that the variance of knowledge in relevant literary fields is minimized. For example, a group of subjects are drawn from students attending a German literature class while another group of subjects are drawn from students attending an American literature class.
6.7 Overall Impression

As they responded to the open-ended questions, the participants were looking at both the cloud and network visualizations generated from author group A1 using the KNN method (cloud 2 and network 2 in table 6.2). The two printouts were placed side by side in front of them in order for them to easily compare the two while answering the questions.

The interview questions asked in the user study are listed in appendix A.

- cloud vs network

7 out of the 10 participants think network visualizations are more helpful than cloud visualizations in terms of forming impression about the underlying collection and inviting exploration. A participant (#1) said:

... I think the network is more helpful because you can see directed connections in addition to popularity levels ... the network ones help me more because you can see the ones that are strongly connected. And I would lean more towards those because I would know if that one is strongly connected I might like that ...

Another participant (# 4) said:

... it’s easier to see groups in this one (network). A lot of the groups are close together so it’s easy to see which writers are similar.

Another (# 10) said:

... I think the network is the most helpful because I can find something that I like and find things that are related to it. It makes it easier because they are connected. ... Seems like it’s grouped more logically. It’s not as cool (as cloud) but I think it’s easy to navigate ...

While some of them did admit that cloud visualizations are more visually arresting and require less efforts from the viewers. For example, one of participants (# 4) said:
... this one (network) is probably more inviting. But if I’m in a hurry and I’m just
glimpsing and skimming the web page I’ll probably look at this one (cloud) just because
it’s simpler, and the words are bigger, and they stand out more - the popular ones at
least. They don’t overlap. And this one (network), you have to put in some effort, to
un-overlap some of the names...

The other three participants actually prefer cloud visualizations. One of them (#3)
said:

I would be somewhat unlikely to pursue it (network) further. This doesn’t compel my
mind. ... But this kind of things with lines for some reason just don’t appeal to me.
... I think I wouldn’t (explore the connected ones). A lot of times when I go online
I have a very specific subject that I want to look at. I prefer just to get to what I’m
looking for ... I think definitely spatial proximity is very important, but this type face
and this boldness (in cloud) jump out to me visually and I can just say “oh let’s look
at the bibliography of Huxley - I know there’s a lot of stuff by him that I haven’t read”
... I think the aggregation per se is less important to me than having it being very easy
for me to visually target and physically scan.

Another (#6) said:

... It (network) overwhelms me... I prefer this one (cloud). I like how some things
are bigger (in cloud) and that (network) has it too but not as pronounced. And it
(cloud) just feels more contained. ... If I’m interested in Arther Miller, I probably
actually wouldn’t be interested in Ibsen at that time. Maybe they’re near each other
since they’re both playwrights but it’s specific sort of intention but not intuitive to me...

In summary, the majority of the participants prefer network visualization for more
detailed browsing and exploration of the underlying collection, given its presentation
of groups and relations through edge connections and spatial proximity. The fact that
some participants prefer cloud visualizations as an “eye catcher” is also instructive - network visualizations should be created to be more visually clear and arresting and incur less cognitive load. This is also exemplified by another user study in which participants’ choices of preferred tag cloud layout for a given task were not only driven by rational factors but largely influenced by aesthetic aspects [23]. A possible combination of the two visualization techniques, and also both the aesthetic and mechanical factors, is a cloud in which items are laid out such that those in the same group or associated are placed near one another or are in the same color, retaining the appearance of a cloud while showing group and relation information emphasized by a network.

• $\epsilon$-neighborhood network vs KNN network

All participants prefer the KNN network to $\epsilon$-neighborhood network except one participant. They characterized visualizations based on KNN method as being “better spread out”, “more insightful” and “more developmental” so they can “see the names more clearly”, while there are “too many overlapping names” in visualizations based on $\epsilon$-neighborhood method. The participant who prefers the $\epsilon$-neighborhood network said “I would prefer the one that’s more contained because it seems like it’s more specific”.

This agrees with the analysis in chapter 5 which demonstrates the better performance of KNN method than the $\epsilon$-neighborhood method in terms of modularity, and the visually clearer layout of networks generated from KNN method.

• alphabetical ordering of tags in clouds

Similar to what has been reported by Hearst et al [21], none of the participants realized that cloud visualizations used in this study are regularly organized into alphabetical order. One participant reported that the alphabetical order might have helped impression formation and the grouping tasks “subconsciously” but not in any obvious way. This implies that the tag clouds are actually “scanned” rather than “read”, es-
especially in a grouping task whereby participants attempted to locate associated names all around the display.

- **font size**

  Although some participants admitted that font size is less helpful than spatial layout in forming impression about the underlying collection, 9 participants said it is important in aesthetically appealing and eye-catching cloud visualization. One participant (#9) would rather not have font size variation, preferring a regular plain list to socially organized treatment:

  ... *(font size indicates popularity) in society or in culture but not in terms of what I would look for. In fact I would want to know what the smaller names are ... That (larger font size) would never influence me ... I don’t go by best-seller lists. I don’t tend to read things that are most popular but I go for what I like.*

  Despite this participant’s opinion against font size variation, font size based on popularity of an item is the defining characteristic of a tag cloud that conveys socially organized information. Again it is acknowledged by participants in this study and has been shown to be one of the reasons that some prefer cloud visualizations to network visualizations for easy and quick impression formation. This confirms previous findings that in tag clouds font size has a consistently strong influence on users’ perception of clouds [99, 23].

- **lines**

  Opinions varied as to the usefulness of line for grouping tasks. 6 participants thought lines in network visualizations are important because they can indicate “which authors are strongly connected” and invite exploration. While 4 other participants stated it is the spatial layout of names that makes the network helpful in identifying groups and relations and they “do not pay much attention to lines”.

128
It is somewhat surprising that not all the participants appreciate lines in a network visualization. This actually makes a grouped or colored tag cloud more promising in which line-based grouping is less or not at all pronounced.

Among the factors evaluated, spatial proximity and font size variation prove significant in affecting users’ perception of the visual displays. Both spatial proximity and font size of items in a visualization strengthen the semantic implication of items. Spatial proximity even conveys information about semantic relation between items, thereby making the tag/author network a “relationship list” in Hodge’s definition of knowledge organization system [24]. The effect of this kind of “relationship list” is largely exemplified in this study as the participants appreciated the added information about groups or categories, which is perhaps intuitive but also reflected in the results of quantitative and qualitative analysis.

When it comes to generating networks that are semantically relevant and aesthetically appealing, the network construction method plays an important role, as confirmed by both the modularity analysis and user study. Specifically, KNN network should be used in preference to \( \epsilon \)-neighborhood network when reducing a similarity-based two-mode network to one-mode network. More novel methods to construct networks should also be explored in future study as it is shown that the selection of network construction method can greatly affect the topology (and thus user perception) of the resulting one-mode network.

The effect of other factors such as alphabetical ordering is relatively minor. As evidenced by the user study, although alphabetical ordering may help users locate a specific item, it does not introduce any additional semantic information and therefore is not appreciated in semantics-oriented tasks such as grouping or navigation.

6.8 Summary

Albeit on a relatively small scale, this user study covers both the quantitative measures in evaluating the visualizations and the factors that cannot be easily quantified.
Statistical tests show that the difference between the numbers of groups identified by participants when they were viewing network visualizations and the numbers of groups identified when they were viewing cloud visualizations is not statistically significant; while the number of items identified by participants when they were viewing network visualizations is on average larger than the number of items identified when they were viewing cloud visualizations. And the difference between the numbers of items identified has been shown to be statistically significant.

The correlation coefficient as high as 0.40 on average between user ratings and the clustering coefficients of groups somewhat reiterates the coherence and semantics of groups generated from community detection. It also partly explains the reason that participants identified more groups and items in a network visualization than they did in a cloud visualization - the groups entailed by geographic proximity and edge connections in a network visualization were appreciated by the participants and facilitated their impression formation, especially on groups and relations.

In the interview part of the study, most participants stated that they prefer network visualizations to cloud visualizations for in-depth browsing and navigation because they can see groups and related items in a network, in addition to popularity levels also emphasized by cloud visualizations.

There are cases where the qualitative aspect of a participant’s response seems to be at odds with the quantitative measure. For example, participant #3 identified substantially more authors (42 vs. 24) when viewing network visualizations while still preferring cloud-based organizations. This exemplifies Lohmann et al.’s finding that one’s choice of preferred tag cloud layout for a given task is not only driven by rational factors but largely influenced by aesthetic aspects [23]. Hearst has concluded that tag clouds are primarily a visualization used to signal the existence of tags and collaborative human activity, as opposed to a visualization used for data analysis [21]. Perhaps in Hearst’s terms, the author network is more of a data analysis tool than a social signaller.
Some of the participants’ preference for cloud visualizations for their lower cognitive load also suggests improvements for network visualizations: larger variance in font sizes, less overlapping of items and clearer or more ordered layout. A possible outcome in that direction is a cloud visualization where items are positioned such that those in the same group appear closer or are in the same color.
Chapter 7

Conclusions and Future Research

As evidenced by studies across scientific fields, network analysis has proven a powerful tool to facilitate the understanding of the underlying complex systems that the networks represent. This study was undertaken in an effort to analyze the characteristics of one-mode similarity networks derived from tagging data on a popular social cataloging site, as well as their implications on improving design of visualization schemes aimed at facilitating navigation and exploration of the underlying collection.

This chapter summarizes the results from both network analysis and user study and gives possible directions for future work.

7.1 Conclusions

Many previous studies have focused on statistical analysis of tagging data from social tagging sites such as del.icio.us and flickr and the implications for information retrieval, tag suggestion and spam filtering. Employing methods developed in the complex network community, this study aims at detecting small-world networks, strong community structures and semantic meanings of the structures in tagging data retrieved from the social cataloging site librarything. Based on findings from network analysis, this study further evaluates the effectiveness of the proposed network visualization as an alternative to the traditional cloud visualization widely used in social tagging applications.

Similar to what has been reported on del.icio.us data sets, one-mode author networks generated from librarything data also exhibit patterns of small-world network. But con-
trary to what has been believed in other studies, simple cosine similarity which does not fully account for weighted term frequency and inverse document frequency may not be the preferred similarity measure when used to project a two-mode network into a one-mode network. While fully weighted TF-IDF turns out to be a superior weighting scheme in terms of maximizing modularity value of the generated one-mode network.

There are arguably strong community structures in the generated one-mode networks from librarything data, as measured by modularity value as high as 0.8 to 0.9. Networks generated using different network construction methods will differ in topology and network characteristics. Particularly, $\epsilon$-neighborhood networks are characterized by a smaller number of larger communities and therefore more contained, while KNN networks usually have a larger number of smaller communities and are therefore more “developmental”. This study suggests that KNN network should be used in preference to $\epsilon$-neighborhood network to generate one-mode networks that are both semantically relevant and aesthetically appealing, while $\epsilon$-neighborhood method is more convenient when the filtering coefficient needs to be fine tuned.

The strong community structures exemplified by high modularity value are further confirmed by close examination of the communities. In the context of librarything data, they usually represent groups of authors that are related or similar in genre, historic period, country of origin or gender. While average clustering coefficient of a group appears to be an approximate indicator of the coherence of the group - the higher the average clustering coefficient, the more similar (or related) the authors in the group are.

The user study further illustrates that tag/author network is a superior visualization scheme for high-level tasks such as grouping, navigation and exploration. The participants could identify significantly more groups when looking at network visualization than what they could identify when looking at cloud visualization. The majority of the participants also prefer network visualization as the user interface to browse and explore the underlying collection because it conveys information about groups and relations of the items.
The participants’ ratings of the groups generated by community detection bear relatively strong correlation to the average clustering coefficients of the groups, indicating that average clustering coefficient of a group may be used as a practical measure of coherence of group members.

To summarize, this study suggests that

- tag occurrences in the librarything data follow power law distribution, and the one-mode networks resulting from similarity analysis are small-world networks.

- one-mode author networks contain strong community structures that are algorithmically detectable and semantically relevant. To generate one-mode networks, fully-weighted TF-IDF, as opposed to cosine similarity, should be used. To generate a network with high modularity value and a large number of communities, KNN method should be used in network construction.

- communities identified by community detection methods are not equal in topology and coherence. Average clustering coefficient of a community (group) can be employed as an approximate indicator of coherence of nodes in the community (items in the group).

- users will be more successful at identifying related groups of authors when viewing network visualizations than when viewing tag clouds. Network visualizations may be preferred to cloud visualizations for in-depth browsing and navigation of the underlying collection.

These findings can further our understanding of the data set gathered from the social cataloging site librarything, an instance of the social tagging paradigm as the extension beyond keyword-based indexing and hierarchical classification schemes. They suggest that not only social tag clouds, as a straightforward application of tagging data, can serve as a social signaller and eye catcher, but also tag and item networks, derived from automated analysis and processing of tagging data, actually present rich information about communities.
and relations, and therefore may facilitate general-purpose information seeking tasks, especially when a user has merely a vague conception as to what is to be retrieved, for example, from online library catalog, Amazon website or a music streaming site.

7.2 Future Research

This study is intended to frame around the application and evaluation of network-based visualization schemes in the particular territory of the social tagging universe - the social cataloging site librarything. But other topics on social tagging are equally interesting and may well serve the purpose of a natural extension of current research. The following are several examples:

- This study has been focusing on the small-world network properties of the one-mode author and tag networks and their implication, while the scale-free properties of the networks are equally interesting. For example, those “hub nodes” with high degree in a tag network may represent popular tags or subject words and probably lie close to the top of a hierarchical classification system, while clusters of tags in the tag network are likely to contain tag words that are semantically related or adjacent in the classification system. There have been previous studies that transform a large corpus of tags annotating objects in a tagging system into a navigable hierarchical taxonomy of tags [102, 107]. Equipped with similarity calculation methods and network construction methods investigated in this study, these undertakings may be further pursued to reconcile folksonomy generated from group knowledge and ontology created by subject experts.

- Taxonomy-like structures might be easy to form on a social cataloging site given average users’ relative familiarity with book classification schemes. Applications of the same analysis on other tagging sites may generate different results. It would be worth-
while to see if the findings from this study are confined to book folksonomy/taxonomy or generalizable across tagging sites (*last.fm*, *del.icio.us*, etc.).

- With the institutional repository as the increasingly popular and academia-oriented platform for managing intellectual contents, tag clouds and tag groups may find their usage at the entry page of these repositories as a visually arresting construct to present primary topics of the underlying collection and guide navigation of users. Further interviews can be conducted with creators and users of typical institutional repositories to further understanding of needs of this community. The interview framework created during this study and the analysis of results will shed light on procedure specifics and candidate selection criteria for future interviews.

- Differences exist between taggers, especially between taggers who make frequent contribution to the site and those at the “long tail”. Interesting patterns might be revealed if tagging behaviors of individual users are closely examined. Also worth looking at is the interaction between taggers. The collaborative nature of social tagging implies that taggers tend to choose frequent tags that have been heavily used by other taggers. But how this would affect tagging dynamics, especially under the framework of scale-free network theory, is still an open question.

- In a real world scenario supervised categorization might be incorporated in addition to unsupervised clustering and visualization to iteratively categorize and classify items. This may be evaluated in a research setting where the participants are asked to group additional items given a number of computer-generate groups, or in an online environment where users can drag and drop items into or out of visualized clusters of items in a web application.

All these potential topics for future research, along with existing research studies, represent a growing community of researchers across various scientific and humanistic fields
studying the usage, characteristics, dynamics and visualization of social tags. Results from these studies may well benefit both researchers and designers of online information seeking systems, from which this dissertation study originated. And by leveraging socially-organized information, information seeking systems such as the online library catalog may serve users in a new way, as demonstrated by both network and cloud visualizations of tags in research studies and also put into practice by an increasing number of websites.
References


[72] Andrew Y. Wu, Michael Garland, and Jiawei Han. Mining scale-free networks using geodesic clustering. In *KDD '04: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 719–724, New York, NY, USA, 2004. ACM.


1990.


[95] M. Girvan and M. E. J. Newman. Community structure in social and biological net-


at internet scale. *IEEE Transactions on Visualization and Computer Graphics*,

[99] Scott Bateman, Carl Gutwin, and Miguel Nacenta. Seeing things in the clouds: the
effect of visual features on tag cloud selections. In *HT ’08: Proceedings of the Nineteen-
hundred ACM Conference on Hypertext and Hypermedia*, pages 193–202, New York, NY,
USA, 2008. ACM.

[100] R. Lambiotte and M. Ausloos. Uncovering collective listening habits and music genres


[102] Paul Heymann and Hector Garcia-Molina. Collaborative creation of communal hi-
erarchical taxonomies in social tagging systems. Technical Report 2006-10, Stanford
University, April 2006.

[103] Börkur Sigurbjörnsson and Roelof van Zwol. Flickr tag recommendation based on
collective knowledge. In *WWW ’08: Proceeding of the 17th International Conference

[104] Paul Heymann, Andreas Paepcke, and Hector Garcia-Molina. Tagging human knowl-
Search and Data Mining*, pages 51–60, New York, NY, USA, 2010. ACM.


Appendix A

Core Questions for the Survey

The core questions that are for all subjects during the experiment are:

- On a scale of 1 to 10 (10 being “very good”), how do you rate your knowledge of English and American literature?

- On a scale of 1 to 10 (10 being “very good”), how do you rate your knowledge of social tagging?

- Do you know what a tag cloud is?

- On a scale of 1 to 10 (10 being “very familiar”), how do you rate your familiarity with tag clouds?

- Given the tag cloud/network, what do you think the collection is about?

- Given the author cloud/network, can you identify two or more groups of authors of different genres, periods or countries? Please list all groups that have at least 5 authors. Why have you made these selections?

- Do you think the tag cloud/network is a good way to get the gist of the underlying collection? Why or why not?

- Do you think differences in font sizes and the alphabetical ordering of tag clouds are helpful?

- Do you think lines in a network visualization will help you identify groups and form impression about the underlying collection?
• Do you think tag network presentation helps you identify groups of or relationships between tags/authors? Why or why not?

• Which one of the two visualizations do you prefer when you browse, navigate or explore a website?

• Do you have any overall comments about the visualizations?

• How could any of the visualizations be improved?
Appendix B

Visualizations Viewed by All the Subjects

To generate the clouds in Wordle application, the parameters are set as follows:

- Font
  - League Gothic

- Layout
  - Prefer Alphabetical Order
  - Rounder Edges
  - Horizontal

- Color
  - BW (Black and White)
Figure B.1: Author cloud #1
Figure B.2: Author network #1
Figure B.3: Author cloud #2
Figure B.4: Author network #2