NETWORK ANALYSIS OF SHARED INTERESTS REPRESENTED BY SOCIAL BOOKMARKING BEHAVIORS

Jung Sun Oh

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Information and Library Science.

Chapel Hill
2010

Approved by:

Dr. Barbara Wildemuth

Dr. Gary Marchionini

Dr. Stephanie Haas

Dr. Paul Solomon

Dr. Katherine McCain

ABSTRACT

Jung Sun Oh
Network analysis of shared interests represented by social bookmarking behaviors
(Under the direction of Barbara Wildemuth)

Social bookmarking is a new phenomenon characterized by a number of features including active user participation, open and collective discovery of resources, and user-generated metadata. Among others, this study pays particular attention to its nature of being at the intersection of personal information space and social information space. While users of a social bookmarking site create and maintain their own bookmark collections, the users' personal information spaces, in aggregate, build up the information space of the site as a whole.

The overall goal of this study is to understand how *social* information space may emerge when personal information spaces of users intersect and overlap with shared interests. The main purpose of the study is two-fold: first, to see whether and how we can identify shared interest space(s) within the general information space of a social bookmarking site; and second, to evaluate the applicability of social network analysis to this end. *Delicious.com*, one of the most successful instances of social bookmarking, was chosen as the case.

The study was carried out in three phases asking separate yet interrelated questions concerning the overall level of interest overlap, the structural patterns in the

network of users connected by shared interests, and the communities of interest within the network. The results indicate that, while individual users of *delicious.com* have a broad range of diverse interests, there is a considerable level of overlap and commonality, providing a ground for creating implicit networks of users with shared interests. The networks constructed based on common bookmarks revealed intriguing structural patterns commonly found in well-established social systems, including a core periphery structure with a high level of connectivity, which form a basis for efficient information sharing and knowledge transfer. Furthermore, an exploratory analysis of the network communities showed that each community has a distinct theme defining the shared interests of its members, at a high level of coherence.

Overall, the results suggest that networks of people with shared interests can be induced from their social bookmarking behaviors and such networks can provide a venue for investigating social mechanisms of information sharing in this new information environment. Future research can be built upon the methods and findings of this study to further explore the implication of the emergent and implicit network of shared interests.

ACKNOWLEDGEMENT

First and foremost, I would like to thank my advisor, Dr. Barbara Wildemuth, from the bottom of my heart. I have been blessed with many inspiring and caring people around me in my life, especially during my years at SILS, but having her as my advisor and mentor has been beyond blessing. She was there to support whenever I stumbled and she has enabled me to move forward in such an understanding and encouraging way. I know I would have not been able to be where I am, both intellectually and career-wise, without her support. She has been a model to follow, as a researcher and as a person, and will continue to be.

I would also like to express my deepest gratitude to my committee members, Dr. Gary Marchionini, Dr. Stephanie Haas, Dr. Paul Solomon, and Dr. Katherine McCain. I had the honor to work with Dr. Marchionini and Dr. Haas for the first three years in my PhD program. The GovStat project had impacted me in many meaningful ways and what I learned from the great researchers in the team formed the basis of my research ability. I deeply thank them for their support and for being inspirations. As many doctoral students at SILS, I am indebted to Dr. Solomon. While he was with us at SILS, I went to his office a number of times, every time with something to ask for. When I asked him to be on my committee, he gave me the same smile as always and I knew I would have his unconditional support and help, as always.

Lastly, I am very grateful for having Dr. McCain in my committee. I especially thank her for her insightful and challenging comments on my dissertation. In years to come, I will reflect upon her advice while pursuing further research in this area.

There are many current and former Ph.D. students at SILS who have been great companions for years. I want to thank Sheila Denn, Cristina Pattuelli, Sanghee Oh, Yan Zhang and many others for their friendship and for all those conversations we had about research, various aspects of student life, and the unknown future.

I want to thank Mr. Scott Adams and Mr. Aaron Brubaker for their help, support, and patience with my computing needs for this dissertation project. Not only did I consume a large portion of the space and capacity of SILS servers, with their kind consideration, I had a designated machine for the analysis of data for several months.

Finally, I would like to tell my family how much I appreciate them. I found my strength from their love. My mom and dad had to suffer from a terrible car accident when they came to visit me last year. Even in the most painful days, I remember how they worried more about me than their own sufferings. My daughter Kyu Won has been *the* source of joy in my life. She is also the one who has been there to hug me when I am tired and weary. Lastly, for the support and sacrifice of my husband, I cannot find words to express my gratitude, not even closely.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1. Introduction

1.1 Background

The new phenomenon commonly called social bookmarking or social tagging produces a new information environment where users are actively involved, as a part of their own information management strategy, in accumulation of collective knowledge.

One of the most important characteristics of social bookmarking is that, by default, all the resources and activities are open for everyone to see. With this remarkable openness, not only can people collect and organize information resources for their own interests on a social bookmarking site, they can also explore the aggregated collections of resources built by the community as a whole. The openness, in and of itself, entails social implications with a bearing on how people perceive and become aware of objects and other people within the information space. Social bookmarking, therefore, has a dual nature as a personal bookmark management tool and as social software and is, in effect, at the intersection of personal and social information space. This unique characteristic of this phenomenon and its potential implications are the main motivations for this study.

In observance of the rapid adoption and growing popularity of social bookmarking, many research communities have shown interest over the last few years. The unprecedented amount of end-user generated metadata given in the form of tags has been a powerful attraction, with a vision of building a bottom-up taxonomy.

Another, perhaps more important, value of this phenomenon resides in the associations among the three axes of social bookmarking – people, information objects, and tags – drawn from bookmarking activities. These three axes have also caught researchers' attention; however, it appears that the significant ramifications of the dual nature of this phenomenon, bridging personal and social information space, remains unnoticed or unaddressed, by and large.

With the growing recognition of the subjectivity and variability of human conceptions of information, as well as the significant effect of contextual factors, many researchers have pointed out the shortcomings of traditional approaches to information representation, organization, and access mechanisms in information retrieval (IR) systems based on the unified objective model (Bates, 1986; Chalmers, 1999). Therefore, developing a flexible and adaptable information environment for users has become an important problem in the field of information science, and has been pursued in a number of ways. For instance, empirical observations on how people organize and/or search for information in their actual context of work have been conducted to bring insights into designing information tools tailored to meet the needs of users. Studies on organizational behaviors in the area of personal information management, carried out as small-scale qualitative research, fall under this category (Kwasnik, 1989; Barreau, 1995). Some efforts are geared more specifically toward understanding people's needs and interests, typically within the context of a specific information retrieval system. Under the umbrella concept of 'user modeling', researchers have experimented with various approaches to obtaining and making use of information about users' interests or needs. Identifying a user's interests is a

challenging problem. One way to go about it is to get users directly involved in the modeling process, such as asking them to provide a list of keywords describing their needs. However, with the recognition of the problems of this approach, including users' reluctance to put in extra effort or people's general inability to articulate their needs, efforts have been made to develop techniques that can unobtrusively obtain information about users, based on their previous behaviors of information seeking or use (Kelly, 2004). For instance, Kim, Oard, and Romanik (2000) found that people's choice of saving or printing a document can be a useful implicit indicator of the relevance of the document to their interests.

Another line of research tries to take advantage of user behaviors from a different angle, in an attempt to provide better access to information. Instead of modeling individual users, the patterns of information access created by a community of users are detected and used to inform subsequent users. Studies and systems adopting the concept of social navigation (Wexelblat & Maes, 1999; Dieberger et al., 2000) fall under this category. For example, Wexelblat's Footprints system (2003) captures the access history of web links and shows signs of popular paths other people have taken. Just as people use footprints left by others to find their way in unfamiliar land in the physical world, users of the system can make navigational decisions informed by other people's prior behaviors. A related technology, collaborative filtering (Resnick & Varian, 1997), shares the basic idea of taking advantage of others' behaviors, but puts weight on behaviors of *similar* people. People who are by some measure determined to have similar interests or preferences form an implicit group, and the behavior of the group is monitored to inform its members. In either approach, users get to have a degree of awareness of what other users in the system find useful.

With its dual nature as a personal information management tool and social software, social bookmarking is related to both lines of research and development efforts discussed above. In other words, social bookmarking holds potential for exploring and/or exploiting the patterns of user behaviors both at the individual level and at the group level. Studies for understanding the behavior of individuals have been conducted either in a laboratory setting or with a small-scale qualitative design. With the availability of chronological records of bookmarking activities of a wide range of users in their natural setting, social bookmarking opens novel possibilities for investigating human information behaviors on a large scale in an unobtrusive manner. Just as previously-studied interactions with information objects, such as reading (measured by time spent), saving, or printing, can be used as indicators of user interests, the act of bookmarking tells us a good deal about the interests of the person who chose to save it. The accumulated record of such actions, as well as the resulting collection of bookmarks, reflects the range of a user's interests, and represents his/her personal information space built by a series of information seeking episodes.

Social bookmarking also provides a new venue for investigating aggregated behaviors of users, and for developing social information tools. The fact that people's activities are visible to one another in a social bookmarking site is in and of itself an important 'social' feature, in that people can discover, often serendipitously, useful information following the traces other people have left in the public space. In addition, in any social bookmarking site, each and every information object has an indication of how many people have saved it, providing a form of an 'accumulated

wear' (Hill et al., 1992) feature for social navigation. However, the potential value of social bookmarking as a social information tool, we believe, far exceeds free exploration or the primitive support of social navigation.

Not only do people leave traces in the information space of a social bookmarking site, but the information space itself is constructed by the accumulated activities of the users and, in effect, is the aggregate of their personal information spaces. The structure of the information space then emerges from the way personal information spaces of users intersect and overlap. Building upon the interpretation of bookmarking activities as an indicator of a person's interests, overlapping bookmarking activities of certain users can be inferred as an indicator of their shared interests. This, in turn, leads to an intriguing conceptual picture where the large (global) information space of a social bookmarking site consists of a number of local information spaces characterized by shared interests of their respective groups of people. From a user's point of view, the shared interest space, if identified, is an extension of his/her personal information space, and provides a dynamic local 'view' of the overall information space, allowing him/her to take advantage of others' efforts in a way tailored to their interests. In addition, unlike typical social navigation or collaborative filtering systems where only information objects are visible, people sharing interests can *see* one another in a social bookmarking site. In other words, social bookmarking can afford awareness of both information objects and people.

All in all, social bookmarking presents a promising set of data pertinent to the core problems of information science, both for research and practical applications. While most discussions about social bookmarking or tagging phenomena, especially in

the area of information and library science, have centered around the pros/cons of tags as an information organization or access mechanism, this study focuses on the conceptual picture of a shared interest space – the implicit formation of a shared space among a group of individuals, based on the overlap among their respective personal information spaces.

An investigation of a new phenomenon, such as social bookmarking, often calls upon the researcher to find an appropriate research tool to represent and analyze the phenomenon. In this study, we take network analysis as such a tool. In many scientific areas, it has been shown that a wide range of empirical problems can be represented as a network and explored in terms of the structural patterns of the network (Breiger, 2003; Newman, 2003). In social bookmarking, there exist implicit associations among users, information objects (URLs), and tags, made by the aggregated activities of bookmarking. These associations make it plausible to represent the phenomenon as a network, or a set of networks. For instance, information objects jointly saved by many people can be considered as related, or people who have many common information objects can be considered as connected. In other words, relations (links) among the entities of social bookmarking can be drawn from co-occurrences or commonality of bookmarks, to form a network. Given the possibility of network representation, it is the basic stance of this study that social network analysis provides both the theoretical underpinnings and the methodological approaches that enable us to explore this new phenomenon in general, and to study the question of a shared interest space among users in particular. Specifically, a network of users can be created by connecting users based on their common possession of bookmarks,

and the network can be explored for meaningful structural patterns. Adopting the network perspective, which posits that a social phenomenon can be studied in terms of patterns of relations among the actors involved therein, this study has investigated the information space of a social bookmarking site by identifying structural patterns of the network of users. In a broader context, understanding the emergent network of users that arises from their behavioral patterns has great potential for building a flexible and adaptable information environment for users.

1.2 Problem statement

The main purpose of the study is two-fold: first, to see whether and how we can identify and characterize shared interest space(s) within the large-scale information space of a social bookmarking site, based on the aggregated patterns of individual activities; and second, to evaluate the applicability of the theories and methods developed in social network analysis to this end.

In order to address the problem of identifying a shared interest space, this study was carried out in three phases, each building upon the previous phase. In the following section, specific problems to be addressed in each phase will be discussed. The potential contributions of each phase to addressing a broader problem will also be pointed out.

1.2.1 The level of accumulation and overlap in a social bookmarking space

As social bookmarking gains popularity, the potential value of the aggregated

collection of information and human judgments involved therein have attracted great attention from the research community. Whereas early studies of this phenomenon focused on finding regularities in user activities or in tag distribution in an attempt to understand underlying dynamics (Golder & Huberman, 2006), more recent studies have investigated various ways of harnessing collective social knowledge from social bookmarking data. Among others, the possibility of making personalized recommendations based on past activities of bookmarking or tagging is a promising topic (Wu et al., 2006; Jiao and Cao, 2007; Ji et al., 2007). Another line of research addresses the issue of constructing a semantic tool based on user-generated metadata, tags (Begelman et al., 2006; Halpin et al, 2007).

All the above and other approaches to exploring/exploiting social bookmarking data presuppose a certain level of accumulation and overlap of activities with regard to the entity of interest (user, information resource, or tag). For instance, recommendation or filtering of information can only be effective when users have a certain level of shared activities with others so that people with similar interests can be identified. Similarly, deriving relationships among tags requires co-occurrences in a large number of cases.

However, there is little empirical research validating this key assumption of accumulation and overlap of activities in social bookmarking. As stated above, social bookmarking has a dual nature as both a personal and a social information tool. Considering the potentially unlimited range of resources that could be bookmarked by a large variety of users, it might be the case that a large portion of the information space of a social bookmarking site is comprised of resources that are bookmarked

only once or a few times. In other words, individual users may not share resources and interests. On the other hand, it might be that users in the site tend to have many resources in common, partly because they can see resources other people have found valuable and incorporate those resources into their own collections. In that case, the overall level of accumulation and overlap would be high. The first phase of this study examined which is the case that we can actually observe in a social bookmarking site. More specifically, we analyzed bookmarking activities in *delicious.com*, to assess the level of accumulation and overlap across resources (a resource-centric view) and users (a user-centric view).

Gauging the extent to which bookmarking activities are accumulated and overlapped across resources and users would be valuable 1) in understanding the overall characteristics of the information space of social bookmarking, and 2) in validating basic assumptions for designing applications or services, such as a recommender system, based on bookmarking data. In the context of the current study, the act of bookmarking is considered as an indicator of a person's interests. We posit that the level of accumulation and overlap observed in bookmarking activities is a reasonable indicator of the level of shared interests within the community.

1.2.2 A network of shared interests

If a certain level of shared interest is observed within the community of a social bookmarking site, in this case *delicious.com*, we can proceed to ask the second question: Whether and how users of a social bookmarking site can be connected based on their shared interests, in other words, whether and how the bookmarking

activities of individual users can be used to create an observable network of users.

In the second phase of this study, a network of users was induced based on the assumption that the possession of common items indicates shared interests. In the induced network, therefore, relations among users are defined by the bookmarks (URLs) they have posted in common. It should be noted that, since there is little chance that users in a social bookmarking site have direct social ties or interactions, and the relations among users are implicitly drawn from their behaviors, the network of users is not a typical 'social' network. However, the approach is grounded in the intuitive notion of shared interests reflected in previous behavior. In addition, by the way the *delicious.com* site is currently designed, each item saved by a pair of users provides a path through which one can move (navigate) to discover the other in the information space.

Social network theories and methods allow abstract representation of a system of interest in terms of relationships among actors, which in turn can be analyzed to reveal the structure that arises from the patterns of connections. The strength of network analysis in studying an empirical phenomenon is in its ability to find regularities or patterns out of seeming untangled or prohibitively complex situation. With network analytic tools, various structural properties of a network can be examined to characterize and understand the structure of the network, which may not be apparent or discernible at first. Understanding of emerging patterns in a network structure is often instrumental in designing a process that operates on the network system. We expect that constructing a network of users and investigating the structure of the network will provide a useful way for understanding the dynamics of social

bookmarking and for gaining insights on tools and services that can be built on social bookmarking.

It is now well documented that many real-world networks, despite obvious differences in the respective systems they represent, exhibit striking regularities with regard to some structural properties, such as a short characteristic pathlength, a high clustering coefficient, and a highly skewed degree distribution. Research shows that such features can have significant implications for critical issues, such as the efficiency of transmission/navigation on a network. It is of interest to see whether and to what extent such features repeatedly found in other real-world networks are observed in this new information environment.

1.2.3 Community structure

The third phase of the study focused on one particular global property of a network, community structure. In a network, community structure arises from disproportionate distribution of ties. A community, in the context of network analysis, is an empirically discovered group of actors who are densely interconnected within the group, but are only sparsely connected to actors outside the group. The presence or absence of a community structure in general informs us whether the network consists of distinct communities, or whether it represents a relatively homogeneous group. With a network of users induced from their bookmarking behavior, it would be interesting to see whether a community structure exists in the network. Given the fact that the network is constructed such that users assumed to have shared interests are connected, communities in the network, if they exist, may represent different interests.

In social network analysis, finding out the group structure within a system has been a central issue, since it provides a way to investigate the fundamental sociological concept of 'group' in network analytic terms. However, it is important to note that groups or communities in a network are defined solely by the structure. A subgroup analysis is in effect to detect densely knit segments of a network, and to explain possible causes or effects of such structure. Studies of cohesive subgroups are typically interested in seeing whether subgroups identified by structural patterns differ with respect to other variables. This analytic frame is relevant for networks other than a typical social network. In the context of social bookmarking, the discovery of a group/community structure can provide a ground for a more detailed study. For instance, similarities and differences of tagging behavior may be compared within and across network communities.

As mentioned above, however, 'communities' in the network analytic framework are identified based solely on structural patterns. It should be noted that the *structurally uncovered* communities do not necessarily correspond with communities in the conventional sense. In a social network representing direct social relations or interpersonal ties, the existence of overlapping and interlocking patterns of connections may indeed indicate group cohesion, which brings about a 'sense of community' (McMillan & Chavis, 1986; Heller, 1989). In networks based on induced relations such as the networks of interest in this study, however, the interpretation of network communities is less straightforward. In this study, since the relations in the network are based on shared interests (shared information objects), it is supposed that the network communities, if detected, represent *implicit* communities of interest. In

order to verify this assumption of shared interests as the determinant of the community structure, analyses were conducted in the third phase of this study, in addition to the analysis of structural patterns.

The identification of community structure has been emphasized in recent studies of large complex networks for another reason. A community structure characterizes a network at a higher or coarser level. Its analysis, therefore, brings about a better understanding of the overall configuration of the network, providing a way to deal with the scale and complexity. In addition, it allows "an intermediate scale of analysis between local (e.g. clustering, network motif) and global (e.g. connectivity, path length) structure"(Watts, 2004, p.254). Needless to say, social bookmarking constitutes a large complex network, making it challenging to characterize and understand the structure of the information space. Assessing the applicability of analytic approaches based on community structure is an important step for further exploration of the phenomenon.

1.3 Conclusion

In summary, this study seeks to understand the information space of a social bookmarking site with respect to its users' shared interests, by investigating three different sets of problems: the overall level of interest overlap, the structural patterns in the network of users connected by shared interests, and the communities of interest within the network. While the first phase examined the cumulative distribution of bookmarking activities in the information space as a whole, the second and the third phases applied network analytic techniques to a specific part of the information

space defined by a set of sample users and their respective personal information spaces.

Chapter 2. Literature Review

In studying the new phenomenon of social bookmarking, this study places its nature of bridging personal and social information spaces in the center. With this focus in mind, a spectrum of related research that contributes to developing the conceptual framework for the study will first be reviewed in this chapter. On one end of the spectrum stand lines of research related to the 'personal' aspect of this information space. In social bookmarking, users build and organize their own personal information space by means of bookmarks and tags. Choosing information objects of interest and labeling them is in essence an act of categorization. Therefore, a basic understanding of how people categorize objects and artifacts in general and how people organize information objects, in particular, is relevant. In the 'personal information space' section below, categorization research in cognitive psychology will be reviewed and categorization research in information science, including studies in the area of personal information management, will follow.

On the other end of the spectrum, there are two areas of research that inform how to address the 'social' aspect of this information space, on an aggregated or collective basis: citation analysis and collaborative filtering. Note that the shared information space of social bookmarking is neither designed a priori nor deliberately constructed by participating users, but emerges, in aggregation, from individual bookmarking activities. In this context, therefore, it is the patterns of associations

made by bookmark postings that create the shared information space. Both citation analysis and collaborative filtering study patterns of associations among entities, with different sets of problems and assumptions. In this study, it was assumed that a bookmark (an information object) posted by a user represents the user's judgment of its relevance to his/her interests and, therefore, can serve as an implicit indicator of his/her interest. Furthermore, an information object jointly bookmarked by two users was assumed to indicate their shared interests. These assumptions are very similar to those made in citation analysis and collaborative filtering, respectively.

The next part of the review will discuss social network analysis, which enables abstract representation and exploration of this information space in terms of a network. The section starts with a discussion on the theoretical perspectives and methodological approaches of social network analysis in general. Then a special kind of network analysis called an affiliation network will be reviewed in detail, because this specific analysis method bears particular relevance to this study.

Finally, recent studies of social bookmarking will be reviewed. Research addressing social bookmarking or social tagging has increased considerably over the past few years. However, the vast majority of the research has concentrated on understanding tag patterns or exploiting tags. There is little research closely related to this study.

2.1 Personal information space

2.1.1 Introduction

As Bruner (1956) stated early on, there is little disagreement in cognitive science that

"virtually all cognitive activity involves and is dependent on the process of categorizing" (p. 246). People make sense of their surroundings by categorizing them. Categorization is to "render discriminably different things equivalent, to group the objects and events and people around us into classes, and to respond to them in terms of their class membership rather than their uniqueness" (Bruner, Goodnow, & Austin, 1956, p. 1). By grouping stimuli into meaningful groups, categorization enables people to cope with the complex stimuli encountered in the world, given their cognitive limitations. The importance of categorization in human perception, thought, language and so on has been constantly emphasized (Lakoff, 1987).

The prevalence of categorical thinking in human judgments and actions makes the issue of categorization relevant to many areas of research across a number of disciplines. In sociology, for example, researchers are interested in explaining how social categories are formed and how they influence people's perceptions of themselves and others, which is closely related to other issues such as prejudice or stereotypes.

In information science, theories of categorization provide useful frameworks for understanding how people make sense of information objects, systems, and processes in their information environment. The work of de Mey (1982), who argued that any processing of information is mediated by categories, increased interest in the potential links between the area of cognitive categorization and many areas in information science including indexing and searching.

2.1.2 Categorization research in cognitive science

In this section, theories of categorization will be reviewed, focusing on the increasing

recognition of the inherently dynamic and subjective nature of human categorization. The following begins with the seminal works of Eleanor Rosch, and moves on to theories of category formation which emphasize the role of theories and prior knowledge, and to more recent theories taking context dependent factors into account.

2.1.2.1 Category representation and structure

According to Lakoff (1987), in the classical view, "[categories] were assumed to be abstract containers, with things either inside or outside the category. Things were assumed to be in the same category if and only if they had certain properties in common. And the properties they had in common were taken as defining the category" (p.6). In other words, categories have clear all-or-none boundaries defined by necessary and sufficient conditions. This classical view of categories dominated psychology until the 1970s.

The first challenge to the classical view was presented by Wittgenstein (1974). Wittgenstein pointed out that certain categories, such as "game", do not have defining features or common properties shared by all members, and do not have clear fixed boundaries. Instead, he argued that category members have "family resemblance," meaning that, although they do not share the same set of properties, they are similar in various ways (perhaps sharing various combinations of features). A handful of other researchers also discussed specific problems with the classical view and provided evidence that the classical theory could not adequately explain the complex nature of categorization (for a comprehensive summary, see Lakoff, 1987).

It was Eleanor Rosch who integrated and generalized the insights and findings of

previous studies, and developed empirical research methods to demonstrate that the classical view can not provide a full-scale theoretical account of categorization. Rosch (1978) suggested that category systems are structured along a vertical and a horizontal dimension, which are related to the basic-level categories and to the prototype effects, respectively. She also proposed that two basic principles of categorization underlie the formation of basic level categories and prototypes. The first principle states that the function of category systems is to provide maximum information with least cognitive effort. The second principle is based on the notion of a perceived structure of the world, which assumes that objects in the world are perceived to have correlated clusters of attributes that are highly probable to co-occur (e.g., wings and feathers). The perceived world's structure, Rosch asserts, yields the structure of categories of the real world.

2.1.2.1.1 The vertical dimension - Basic level categories

The vertical dimension is concerned with the level of abstraction or the level of inclusiveness in a taxonomic arrangement of categories, where the most basic categories are placed in the middle of the hierarchical structure. In general, an object can be labeled in many different ways with varying degrees of abstraction (from general to specific), but among many alternatives a category at a particular level of abstraction has a primary status. This phenomenon was originally observed by Brown (1965), who also suggested that the basic level is constructed "at the level of distinctive action" (p.321). On the other hand, in the area of anthropology, ethnographic studies of folk taxonomies (Berlin, 1978; for a extensive review see Berlin, 1992) found certain regularities in the way people from different cultures categorize their environment, which also demonstrates the taxonomic category structure and

the primacy of categories in the middle of the taxonomy. For example, in his study of Tzeltal speakers, Berlin et al. (1974) found that Tzeltal natives tend to identify plants and animals at the level of 'genus' (oak, maple) which is in the middle of the folk classification, rather than one of the upper levels (plant, tree, leaf-bearing tree, etc.) or lower levels (sugar maple, live oak) of the classification structure. It is construed that "the genus was established as the level of biological discontinuity at which human beings could most easily perceive, agree on, learn, remember, and name the discontinuities" (Lakoff, 1987, p.34). Interestingly enough, Berlin also noted that folk taxonomies overlap with scientific taxonomies quite accurately at the basic level, but not at the other levels.

Drawing upon the findings from studies on folk classifications of natural (biological) objects, Rosch et al. (1978) proposed that "categories within taxonomies of concrete objects are structured such that there is generally one level of abstraction at which the most basic category cuts can be made… A taxonomy is a system by which categories are related to one another by means of class inclusion." (p.30). Categories at a higher level of inclusiveness are called superordinate and those at a lower level are subordinate categories. The basic level is generally placed in the middle of the hierarchy. Rosch and her colleagues conducted a series of experiments using a variety of measures and found that the basic level is 1) the most inclusive level at which category members have similar appearances with a representative image, 2) the highest level at which category members are used in similar ways (in other words, people interact with them using similar motor movements), and 3) the level of abstraction at which category members have many common attributes distinguishing them from members of other categories, with few distinctive features among themselves[1]. For

[1] In Rosch's (1978) terms, it is the level of abstraction at which average 'cue validity' and 'category resemblance' are maximized. 'Cue validity' is a probabilistic

example, members of the basic level category 'chair' have overall shapes with discernable common features such as legs and a seat, whereas members of a superordinate category 'furniture' do not have a similar appearance and share only a few attributes. In the case of subordinate categories such as kitchen chairs, comfort chairs, etc., although members of each category have common attributes, most of those attributes overlap with other categories at the same level (e.g. different kinds of chairs) because they all inherit attributes from the higher (basic) level. In terms of the second principle of categorization, basic categories best reflect the correlational structure of the environment. The results of a series of experiments show that objects are identified most rapidly at the basic level, and basic level categories (labels) are most frequently used when people name an object. In addition, as discovered in folk classification studies, children learn the basic level categories first and then learn other objects by generalizing (upward) or specializing (downward).

In summary, it has been found that categories of concrete objects (either natural or man-made) are organized in a hierarchy from general to specific and, more importantly, that there is a certain level in the hierarchy, usually in the middle of the hierarchy, at which cognitively basic and primary categories are situated. There is evidence that this basic level is by and large determined by the way people perceive and interact with objects, suggesting that categorization is not solely dependent on the attributes of the objects themselves, but also on the physical and mental capacities of humans.

---

concept defined as: "the validity of a given cue *x* as a predictor of a given category *y* (the conditional probability of *y*/*x*) increases as the frequency with which cue *x* is associated with the category *y* increases and decreases as the frequency with which cue *x* is associated with categories other than *y* increases." (p.30). Category resemblance is defined as "the weighted sum of the measures of all of the common features within a category minus the sum of the measures of all of the distinctive features." (p.31).

2.1.2.1.2 The horizontal dimension – Prototype effects

The horizontal dimension has to do with the existence of prototypes within categories. In order to move beyond the classical view of categories dictating clear boundaries with necessary and sufficient conditions for category membership, Rosch (1978) argued that in many cases categories are conceived in terms of their prototypes rather than formal boundaries. Prototypes are regarded as 'the clearest cases' of a category, and operationally defined by 'goodness-of-example' ratings in her experiments. With the measure of goodness-of-example, it is shown that there exist asymmetries among category members. Certain category members are judged to be better examples or be more representative of a category, and there is a graded structure among category members as a function of how typical they are perceived as a member of the category. For example, 'robin' is judged to be more typical of 'bird' than falcon, which is more typical than 'penguin.' In a number of experiments conducted by Rosch and her colleagues as well as other researchers, it is repeatedly verified that subjects' judgments of typicality are highly reliable and exhibit striking agreements (Rosch, 1975; Rosch & Mervis, 1975).

Given the empirical findings of degree of prototypicality, Rosch went on to test the effects of prototype structure on various psychological variables, including speed of processing, efficiency of learning, etc (Rosch & Mervis, 1975; Rosch et al., 1976). Rosch (1978) concluded that "the prototypicality of items within a category can be shown to affect virtually all of the major dependent variables used as measures in psychological research" (p.38).

Empirical verifications of prototype effects, indicating that categories have internal structures instead of being simple containers as the classical view assumes, entail the need

for a new model of category representation and structure. Since the 1970s a number of models of categorization have been proposed. According to Smith and Medin (1981), these models fall into one of two general views of categorization: the probabilistic view and the exemplar view.

The probabilistic view is directly influenced by Rosch's research on prototypes. This view is based on two basic assumptions: (1) categories are represented in terms of some abstract summary (prototype), and (2) categories do not necessarily have a set of defining attributes (Smith & Medin, 1981). Instead, it is presumed that prototypes consist of clusters of attributes, which reflect the correlational structure of the world. Category membership of an object is then based on how similar it is to the prototype. In addition, the membership is graded depending on the level of similarity. In other words, category structure is established by the probability of objects matching the abstract summary.

The exemplar view posits that categories are represented by a set of exemplars or known instances, rather than by a single abstract summary. The underlying premise is that when people encounter a new object, exemplars stored in their memory from previous experiences play a critical role in processing the new object. Category membership of an object under this view is determined on the basis of the extent to which it is similar to the stored exemplars of the category.

For models based on the probabilistic view, the main problem is to explain the relative importance of features or combinations of features in membership judgments. For example, Rosch's cue validity model formalizes individual features' weight in terms of the frequency with which they appear in category members and nonmembers (Rosch & Mervis, 1975; Tversky, 1977). One the other hand, many models under the exemplar view are

concerned with what triggers particular exemplars being retrieved and used for categorization.

2.1.2.2 Category formation and the effects of prior knowledge

According to Wrobel (1994), recent developments in categorization theories can be divided into two overlapping phases. The first phase is heavily influenced by the experimental paradigm Rosch and her colleagues developed. Various tasks and measures, including instance categorizations, feature listings, and typicality judgments, have been used to investigate the nature of categorization in general. Many of the empirical findings converge on the understanding that categorization is far more complex than the classical view could explain. More specifically, categories have been shown to have internal structures with graded centrality and graded membership.

In the second phase, much research concerns category formation and learning, that is, the process and governing principles by which categories arise. Whereas earlier studies mostly addressed categories of concrete objects, abstract entities began to be taken into account in this phase. In terms of methodology, more qualitative observation-based analysis was introduced. Important developments included theoretical and empirical studies on the role of people's cognitive models of the world in categorization and the impact of background knowledge on category formation and learning.

Generally, there are two different methods of category formation: similarity-based and theory-based. The first assumes that similarity is the basis of categorization. Similarity is one of the most central theoretical constructs used to explain categorization (Medin et al., 1993). Different views of category representation and structure reviewed in the previous section all rely to some extent on the notion of similarity. The more an item is similar to what

is known about a category (either in the form of the abstract summary or the array of stored instances), the more likely it belongs to the category. In addition, perceived equivalence or similarity is assumed to be largely grounded on the structure of the world. Rosch's second principle of categorization states that people perceive the world with correlated attributes, and the most important category cuts are made based on this perceived structure of correlation. In other words, categories are thought to more or less preserve existing similarities among objects. This view was prevalent in the 1970s.

Murphy and Medin's (1985) paper, "The role of theories in conceptual coherence," opened up the argument that similarity alone can not account for human categorization, especially for why particular categories, out of many alternatives, arise as useful. They argued that, "Current ideas, maxims, and theories concerning the structure of concepts are insufficient to provide an account of conceptual coherence. All such accounts rely directly or indirectly on the notion of similarity, and we argue that the notion of similarity relationships is not sufficiently constraining to determine which concepts will be coherent and meaningful. These approaches are inadequate, in part, because they fail to represent intra- and inter-concept relations and more general world knowledge. We propose a different approach in which attention is focused on people's theories about the world" (p.289). Objects can be similar in numerous ways. Therefore, the notion of similarity is meaningless without further specification of the aspects on which objects under consideration are similar. In the theory-based view, it is assumed that theories (that people bring to bear on categorization situations) govern the process of category formation by determining which features or feature correlations are relevant and important for evaluating similarities between objects. In general terms, theories provide a basis for what are 'essential' features representing a concept and

provide causal or explanatory links between features such that certain correlations of features are highlighted over others. In that way, categories are tied to a larger knowledge structure (including people's theories about or models of the world), and coherence of categories depends on the extent to which they fit into the theories or models. Murphy and Medin's approach emphasizes that people's background knowledge (embodied in theories) plays a critical role as an underlying principle in the process of category formation and use in general terms. However, as they acknowledged, they did not address exactly what constitutes people's theories and how they are involved in category formation.

Lakoff (1987) provided a more detailed theory-based account of categorization. While drawing upon the notion of people's cognitive models of the world, Lakoff also tried to explain how people's imaginative capacities as well as basic perceptions work to form conceptual categories of various levels of abstraction. His approach is based on the idea that "human categorization is essentially a matter of both human experience and imagination – of perception, motor activity, and culture on the one hand, and of metaphor, metonymy, and mental imagery on the other" (p.8). Lakoff argued against what he refers to as the objective view of thought and reason, which assumes that abstract symbols (concepts) derive meanings through their correspondence to objects and structures in the world, and thus are independent of the human mind and body. It is this objective view that gave rise to the classical view of categorization. As an alternative to the objective view, Lakoff suggested what is called the experientialist account. Under this view, symbolic structures are divided into directly meaningful constructs that are based on basic perception and concepts that are built up using imaginative capabilities. In earlier studies, it was empirically shown that concepts at the basic level often involve similar appearance and characterizing motor movements (e.g., sitting on

chairs). Lakoff assumed that basic level categories are largely governed by direct perception and mental image, while categories at other levels depend more on complex cognitive models.

Like Murphy and Medin (1985), Lakoff (1987) assumed that categories are related to or constrained by people's knowledge of the world. More specifically, he suggested that category formation is greatly influenced by 'idealized cognitive models,' "which can be viewed as 'theories' of some subject matter" (p. 45). He further argued that prototype effects arise from the interaction of different cognitive models. As an example, he took Fillmore's (1982) famous example of the concept 'bachelor.' While the concept can be defined as 'an unmarried male,' we would not generally categorize a Catholic priest or a homosexual man as a 'bachelor' even though they clearly meet the definition. According to Lakoff's explanation, the concept is defined with respect to an idealized cognitive model of human society in which there are certain expectations of marriage and marriageable age. However, this idealized model does not perfectly represent the actual world. In some cases, such as Catholic priests or homosexual men, the model does not fit. A simple kind of centrality or membership gradience results from the degree to which the idealized model fits the actual cases. Another example of prototype effects due to idealized cognitive models is shown with the concept of 'mother.' He noted that, in a human society, there are many different kinds of 'mothers' besides biological mothers, including surrogate mothers, foster mothers, stepmothers, etc., each with different associated cognitive models. Therefore, a number of individual cognitive models need to combine and form a cluster model to represent the broad range of cases of 'mother.' More importantly, he argued that all the different models converge on the ideal case or the idealized model and different cases (members) are graded by virtue of their relation to the ideal case. It is important to note that the theories, models, or

ideals people have with relation to conceptual categories are basically constructed within a society.

Theory-based approaches commonly posit that perceived similarity changes depending on people's knowledge and experiences, and that background or implicit knowledge largely affects how categories are formed and used. Therefore, many empirical studies have been conducted to demonstrate the effect of background knowledge. Empirical evidence has shown that prior knowledge has effects on categorization, by weighting particular features or feature combinations, by guiding selection and interpretation of features, or by providing initial category representations which would then be gradually updated by integrating observed data (see Heit, 1997, for a review of empirical findings). For example, Wisniewski and Medin (1994) have demonstrated that, depending on intuitive theories and expectations about a given dataset, subjects search for significantly different features. Their test stimuli consisted of 16 drawings divided into two sets. For one group of subjects (called the 'theory group'), meaningful labels were given for the two sets of drawings, i.e., creative drawings vs. non-creative drawings. For the other group of subjects (called the 'standard group'), neutral labels were given, i.e., group A vs. group B. The task given the subjects was to figure out, for each drawing, why it belonged to the set under which it was categorized. The results show that, "when categories are meaningfully labeled, people bring intuitive theories to the learning context. Learning then involves a process in which people search for evidence in the data that supports abstract features or hypotheses that have been activated by the intuitive theories. In contrast, when categories are labeled in a neutral manner, people search for simple features that distinguish one category from another" (p.221).

Beyond the general knowledge people have in the form of intuitive theories,

individual experiences and expertise also might affect the construction and use of categories. Chi, Feltovich, and Glaser (1981) investigated how novices and experts classified physics problems. Novices tended to classify problems based on “surface features,” while experts categorized problems on the basis of underlying principles. The authors pointed out that similarity judgments vary due to expertise, stating that "experts are able to 'see' the underlying similarities in a great number of different problems, whereas novices 'see' a variety of problems that they consider different" based on surface features (Chi et al., 1981, p. 130).

While Chi et al. (1981) demonstrated the influences of the level of expertise on categorization, Medin et al. (1997) showed the effect of different types of expertise. Medin et al. (1997) have investigated the commonalities and differences among categorizations of trees made by three different types of experts: taxonomists, landscapers, and park maintenance workers. The main motivation of the study was to find out to what extent categorical systems are formed by universal principles and constraints (e.g., preserving the correlated structure of the environment) and to what degree they diverge due to different expertise and goals. Medin et al. supposed that, “With the domain held constant, any differences in categorization and reasoning related to type of expertise reflect nonuniversal contributions of the mind to the understanding of the biological world. Similarities, in contrast, suggest universal tendencies in the structure of mind and/or world” (p.90). In the experiment, 24 subjects were asked to categorize the names of 48 tree species typed on index cards and develop a taxonomy of their own. In addition, they were asked to provide justifications for their categorization decisions. Overall, the result showed both similarities and differences. The similarities found across groups were closely related to the basic level

primacy. Approximately 50% of trees were placed into categories that were common to all three groups, and many of those common categories are at the genus level. In addition, "inductive privilege of genus-level categories" was found in the justification of categories in all three groups. These similarities suggest that universal principles or patterns hold to a certain extent. On the other hand, each group of experts produced a category structure that was highly consistent within the group and more or less distinct from the other two. More specifically, while taxonomist's categorical system largely reproduced the scientific taxonomy that is both broad and deep, maintenance workers' categories revealed a broad but shallow structure. The main differences from scientific taxonomies included categories based on utilities such as 'weed trees' and categories with higher weights on certain morphological features. Despite some differences, the maintenance workers' category structure was fairly correlated with the scientific one. On the contrary, landscapers' categories displayed a noticeably weak correlation with scientific taxonomy, and included a large number of categories based on utilities, such as 'weed trees', 'ornamentals', and 'street trees.' These findings support the claim that categorization varies with people's knowledge, theories, beliefs, etc. The subjects in this study all had substantial knowledge about trees due to their job experiences, but the types of expertise and their interests related to trees were different. The fact that each group developed a distinct category system given the same set of trees verifies that people with different knowledge and experience do actually conceive things differently. In addition, the clear indication of a utility-oriented criterion shown in the landscapers' categories further demonstrates that categorization varies in context dependent ways, reflecting people's interests, perspectives, goals, etc.

2.1.2.3 Dynamic and context dependent categories

The core of Murphy and Medin's (1985) argument is that similarity is too flexible to define categories and, thus, a further mechanism of constraint is needed for an account of categorization. People's prior knowledge or intuitive theories about a stimulus situation constitute one of those mechanisms. Much recent research has suggested that similarity perception not only varies with people's knowledge and experience but also varies in context dependent ways (Medin et al., 1993). For example, shifts in similarity perception or typicality judgment have been observed in experiments, depending on the point-of-view adopted by subjects (Barsalou & Sewell, 1984), or the existence of or types of contrasting categories (Franks, 1995), lexical contexts (Roth & Shoben, 1983), etc. It has led some researchers to argue that categories are dynamically constructed in working memory, rather than being static units of knowledge in long-term memory as many models of category representation imply.

Barsalou is well known for his research on context dependency of categorization and theories of ad hoc goal-derived categories (Barsalou, 1982, 1983, 1985, 1991). In his earlier work, Barsalou (1982) demonstrated that categories have context dependent properties which are activated only in the relevant context, as well as context independent core properties. For example, a property 'can be eaten (by humans)' is usually not associated with 'frog', but when the context of 'a French restaurant' is specified, the property becomes active. Some context dependent properties are accessed by inference, as shown in the example of 'pencil – can pierce something', perhaps in a situation where such properties become useful. Barsalou noted that, among many contextual cues, goals or utilities often make certain context-dependent properties salient. This observation led to his original research on ad hoc

categories.

Ad hoc categories refer to categories that are "created spontaneously for use in specialized contexts" (Barsalou, 1983, p.211). Examples include "foods to eat on a diet" or "things to take from a house in case of fire." Since these categories are generally constructed with respect to certain goals and consist of things that can be instrumental to achieve a given goal, they are also called goal-derived categories. In fact, Barsalou uses goal-derived categories as a more inclusive term because, in his view, goal-derived categories can be either ad hoc or well-established, depending on the frequency of use.

Barsalou (1983) suggested that these ad hoc or goal-derived categories are distinct from common taxonomic categories (e.g., dog, chair) in two aspects. First, ad hoc categories often violate correlated structures of environment. That is, the category members do not seem to share correlated properties, as common categories do (Rosch et al., 1976). Second, ad hoc categories do not have well-established representations. Barsalou (1983) found that free recall of lists containing ad hoc category members (without labels) was hardly better than recall of random word lists, whereas recall of lists of common categories was significantly higher than random word lists or ad hoc category lists. In addition, when subjects were presented with members of categories and asked to identify respective categories (e.g., "What category do moth, bee, gnat, and ant all belong to?"), subjects found it difficult to discover ad hoc categories and came up with various different answers when they did not know relevant contexts, while they could easily name and highly agree on common categories without any context. These findings indicate that there are no persistent concepts representing and binding entities in goal-derived categories. It is specific contexts and goals that bind otherwise disparate category members.

Given that goal-derived categories are highly dynamic and context dependent, Barsalou originally conjectured that these categories might not have graded prototype structure as common taxonomic categories do. However, in a series of experiments, it was found that ad hoc categories possess graded structures that are as salient and stable as those in common categories (Barsalou, 1983, 1985). Since members of goal-derived categories do not necessarily show family resemblance or similarity to the prototype, the question is, what determines typicality of these members and how do people construct similar prototype structures for ad hoc categories? Barsalou (1985) assessed the determinants of prototype structure. More specifically, the effect of family resemblance (central tendency), ideals, and frequency of instantiation (familiarity) on similarity judgments and resulting graded structure were assessed in common taxonomic categories and goal-derived categories. Ideals are regarded as "characteristics that exemplars should have if they are to best serve a goal associated with their category" (p.630). For example, for the category "foods to eat on a diet," an ideal would be "zero calories." Note that, for the category of "food" without the context of "on a diet," the property "zero calories" is neither typical nor desirable. Barsalou argued that, given the goal orientation of ad hoc categories, ideals should be central in judging category membership and typicality of each member. The results of experiments indeed showed that, while family resemblance dominantly determines prototype structures in common taxonomic categories as Rosch and Mervis (1975) found earlier, family resemblance does not predict typicality of goal-derived categories. Instead, ideals and frequency of instantiation turned out to be major determinants for prototype structures in goal-derived categories. In other words, members of goal-derived categories are graded depending on how well each member satisfies the goal and how frequently it is perceived as

a member of the category.

With these findings, Barsalou (1990) distinguished two category formation mechanisms: exemplar learning and conceptual combination. Exemplar learning is considered a bottom-up process involving perception and memory. Whenever people experience exemplars of a category, perceived properties of those exemplars are extracted and integrated into the category representation. In this way, category knowledge is primarily induced from people's accumulated experiences with exemplars. In contrast, conceptual learning does not necessarily rely upon exemplars. People can derive new categories based on reasoning and by manipulating existing knowledge. Moreover, by combining concepts, categories are contextualized. This is a more abstract top-down process that requires deliberate cognitive effort. Barsalou posits that exemplar learning is central to the formation of common taxonomic categories, whereas conceptual combination is more important for goal-derived categories. The relative importance of family resemblance as a determinant of graded structures in common taxonomic categories and that of ideals in goal-derived categories demonstrate different underlying mechanisms[2].

Barsalou (1991) further proposed a general framework explaining different yet complementary roles of common taxonomic and goal-derived categories in the cognitive system. According to this framework, "In perceiving the world and storing information about it, people use common taxonomic categories for primary categorizations, as they build and update their world models. These categories form the building blocks of world models

[2] Similarly, Wisneisky and Bassok (1999) propose two different types of processing mechanisms that underlie categorization: comparison and integration. Taxonomic categories can be compared based on shared dimensions of commonalities and differences. On the other hand, thematic categories do not have common dimensions upon which comparisons can be made, but instead have thematic relations. A thematic relation can be defined by many things, including scenarios or goals.

because they specify central tendency information about entities that is useful across many contexts. Following primary categorizations, people use goal-derived categories for secondary categorizations that specify the relevance of entities to particular goals. By linking entities in a world model to attributes in event frames, people store information that will later facilitate their ability to construct plans" (p. 57). As seen in the above statement, Barsalou introduces the concept of primary and secondary categorization, and also the distinction between world models and (event) frames. A primary categorization is a person's initial categorization of entities, and any subsequent categorization is referred to as a secondary categorization. Barsalou's insight was that it is unlikely that people use goal-derived categories without having made a categorization based on common taxonomic categories. Common taxonomic categories, especially basic level categories, carry rich information about entities independent of specific contexts, and thus are useful for constructing a person's general knowledge about his or her environment, a world model. Attending to a particular goal, people retrieve knowledge relevant to achieving the goal in the form of a *frame* which consists of clusters of attributes (Barsalou, 1992; Cohen and Murphy, 1984). In the early stages of planning, attributes of the appropriate frame need to be instantiated by adopting particular values to be used in the current plan. For example, for planning a vacation, the attributes of *vacation* frame, such as *location, departure, activities*, etc., are to be instantiated. It is the role of goal-derived categories to provide sets of potential values for instantiating the attributes. A secondary categorization of entities in world models derives ad hoc categories to instantiate attributes. In instantiating attributes, various optimizations and constraints are involved to define the specialized context for the current goal. In summary, Barsalou's framework is composed of world models representing entities in people's environments and

frames containing knowledge about attributes for achieving goals. Whereas common taxonomic categories provide building blocks for constructing world models, goal-derived categories provide an interface between world models and frames.

Finally, it is worth noting that Barsalou (1991) presented an interesting theory about domain expertise, although it is only briefly mentioned. As described above, derivation of categories for instantiating attributes entails activation of specific optimizations or constraints. In Barsalou's view, experts would have well-established sets of goal-derived categories with various configurations of optimizations and constraints, as a result of their frequent construction and use of those categories in the course of solving problems. Having well-established goal-derived categories in place, experts can directly access relevant categories in the presence of a specific goal. In their study of the impact of different types of expertise on categorization (reviewed in the previous section), Medin et al. (1997) noted that, "Barsalou's analysis of goal-derived categories and his conjecture that frequently used goal-derived categories may become well established in long-term memory (Barsalou, 1982, 1983, 1991) predict the pattern of landscape sorting in striking detail" (p. 93).

2.1.3 Categorization research in information science

In the following, research in information science related to the way people access, categorize, or organize information objects will be reviewed. Discussions on subjectivity and variability of human conceptions of information in the context of information access and organization will be presented first, and research focusing specifically on people's categorization behavior in the context of personal information management will be reviewed next.

2.1.3.1 Cognitive approaches to information systems design

In the area of library and information science, it has been a central problem to provide subject access for effective retrieval. Organizing information in library catalogs, according to Cutter's *Rules of a Dictionary Catalog,* which formed the basis of contemporary subject headings, has two basic functions: finding known items and gathering similar items (Cutter, 1876). It is the gathering function that is traditionally provided through subject access. While the notion of similarity is the basis of this gathering function, it is assumed that determining similarities among documents relies on a universally applicable vocabulary, which controls subjectivities and diversities of language (Olson, 2002). Library classification schemes and subject headings provide such language. Jacob (2004) compares and contrasts two mechanisms for information organization: classification and categorization. According to her, while categorization is characterized by its flexibility and context dependency, classification is designed to have a rigid structure with mutually exclusive and non-overlapping classes in order to provide the stability of reference. She also notes that, while the classical theory of categorization which assumes clear boundaries of category membership defined by essential common features, does not account for the variability of cognitive categorization, it does provide a theoretical basis upon which principles of classification have been established.

Studies of cognitive categories have provided compelling evidence of the dynamic nature of the cognitive processes by which individuals conceive their environment. Given this evidence, the question arises whether cognitive structures constructed by individuals in organizing/accessing information accord with the formal organization imposed by a controlled vocabulary (classification or subject headings) within an information system. Indeed, catalog use studies have repeatedly shown that people experience difficulties

interacting with cataloging systems and often fail to access items due to the mismatch of terms (for example, Krikelas, 1972; Cochrane & Markey, 1983; Borgman, 1986). Indexer consistency studies have further demonstrated that individual differences and variability can not be controlled by the use of a controlled vocabulary. Not only can ordinary system users not find the 'right' terms, but highly trained indexers using the same vocabulary do not agree on indexing terms for a given document in a great number of cases. Even the same indexer indexes the same document differently over time (Jacoby & Slamecka, 1962; Cooper, 1969; Zunde & Dexter, 1969; Markey, 1984). Another line of empirical evidence of individual variability is found in studies on search term overlap. The agreement between searchers on search terms for the same question is shown to be relatively low, regardless of the complexity of the question, of the level of search experience of the searchers, or of specific measures of overlap (Bates, 1977; Fidel, 1985; Saracevic & Kantor, 1988). Based on these studies, researchers have questioned existing approaches to subject representation and information organization. Bates (1986) stressed that the current design of subject access does not accommodate the complexity and diversity inevitably involved in the processes of indexing and searching, and proposed that a new model of subject access needs to be based on philosophical stances fundamentally different from the objectivist view which forms the basis of the current organizational principles.

Recently, more attention has been drawn to cognitive approaches in information science. Understanding the way that individuals categorize information objects has been emphasized, with the basic premise that it can be instrumental to improving the design of information architectures and systems. For example, Carlyle (1999, 2001) asked subjects to sort 47 documents related to a particular work, Charles Dickens' *A Christmas Carol*, into

groups based on their own judgments of similarities between the documents. Groups common across study participants and attributes identified as important in the process of clustering were analyzed and the implications for designing retrieval systems were discussed. In a school library setting, Cooper (2004) investigated how children in various developmental phases (grades) conceptually categorized information in a library. Subjects were asked to name books (topics) that they think should be included in a library, and to categorize the resulting pool of terms (representing topics). Beyond empirical observations of user categorizations, an attempt to develop a theoretical framework is found in Cole and Leide (2006). Their research on cognitive information organization behavior introduces recent theories of categorization in cognitive psychology into the problem of how users, especially domain novices, identify concepts (terms) within the conceptual structure of the domain to formulate search queries representing their needs. Their theory of information organization behavior states that, in response to unfamiliar stimuli in documents or systems, people organize information through formation/reformation of ad hoc categories. They posit that domain novices who often do not have established frames for the topic area rely on metaphorical thinking through which they construct categories of an unknown domain drawing upon their knowledge of a known domain. Based on this assumption, they propose to incorporate the concept and device of metaphor instantiation into IR system design. According to their proposal, users with a vague idea of their problem situation and limited understanding of the relevant domain benefit from an interactive IR system that assists them in finding a metaphorical frame, the structure of a known domain that is applicable to the unfamiliar domain of the current problem. The frame can then serve as cognitive scaffolding for facilitating the process of identification of information needs and exploration of the target

domain.

While there is an extensive body of empirical research on categorization in the area of cognitive psychology, it is important to note that stimuli used in those studies are mostly concrete (natural or man-made objects such as trees, birds, chairs, etc.) or fairly simple (e.g., geometric shapes). It is not clear whether the same set of principles governing the process of categorization of simple concrete objects would hold for categorization of information objects containing a potentially large number of abstract concepts. There is little empirical evidence for that matter. Bates (1998) suggested that, even though information seeking behavior would not be amenable to controlled experiments as carried out by Rosch, much more research is needed to see whether similar patterns of basic level primacy would be observed in people's use of search terms. In addition, she noted that studies of folk classification have found a consistent pattern across many cultures that not only categories have a hierarchical structure consisting of few levels with the generic level (in the middle) being primary, but also the generic level in any folk classification has a strictly limited number of terms, ranging from 250 to 800. If research finds similar patterns in search terms, she argues, there are significant implications for design of access in information systems.

2.1.3.2 Categorization behavior in the context of personal information management

Bowker and Star (1999) stated that, "the categories represented on our desktops and in our medicine cabinets are fairly ad hoc and individual, not even legitimate anthropological folk or ethno classification … everyone uses and creates them in some form, and they are (increasingly) important in organizing computer-based work" (p. 6). In the context of

personal information management, research has been conducted to gain a better understanding of the kinds of categories that are constructed for and interwoven into people's everyday-life interactions with their surroundings. Among other things, how people organize their work space containing a myriad of information artifacts (e.g., books, manuals, memos, forms, etc.) is of particular interest.

Malone (1983) presented one of the earliest studies of individuals' information organization behavior in their own environment. Malone's case study addressed how people organize information objects in their offices, in the interest of designing better office information systems. He interviewed various types of office workers and research scientists and explored the patterns of their behavior. He found that two different organizational units commonly used for grouping things are files and piles. While files are well-organized and labeled, piles are often loosely defined stacks containing mixed content. Malone observed that offices typically have a large number of piles, in part due to the fact that people tend to defer decisions about where to file things or what to do with them. He claimed that the cognitive difficulty of categorizing (including labeling) information is a major factor in explaining this behavior. For example, many subjects in his study mentioned that it is often the case that a document belongs to more than one category or is potentially related to various tasks, which makes it difficult for them to put it into a filing system. It was discovered, however, that there is a certain pattern of organization regardless of whether a stack is well-defined and coherent or not. People tend to arrange files and piles on their desk space according to relative importance and the actions that need to be taken. From this observation, Malone concluded that people organize information in their workspace not only to make it easier to find it later, but also to remind themselves of things to be done with it.

Kwasnik (1989) also explored people's organization behavior in their offices, but specifically addressed "the influence of context on the process by which people organize and classify their own documents in their own information space" (p.145). She interviewed eight university professors and asked them to describe materials in their offices and explain the organization, recollecting "classification decisions" made about those materials. The basic premise of her study was that investigation of people's classification behavior should be conducted in a natural setting because "classificatory decisions are always made in relation to something else" (p.147). Participants in the study were allowed to use their own words rather than to choose terms from a given set, and the investigator did not impose any constraints on how they described things. In fact, terms used by participants to name or label basically similar items showed great variety, often accompanied by further qualifying expressions. Both the noun and qualifying terms were analyzed to identify dimensions along which classificatory decisions had been made. In the analysis, a set of coding categories representing these dimensions were incrementally derived from and applied to the data in order to discover which dimensions were commonly used across participants, and which were most frequently invoked. The results showed that the most frequently occurring dimensions were form, use, time, topic, and circumstance. When groups of dimensions were analyzed, the group related to situational attributes (e.g., use) was the most frequently used and the group related to document attributes (e.g., topic) was the second. These findings indicate that traditional classification systems, which rely almost entirely on document attributes, do not provide adequate support for individuals' needs for organizing their information space. In addition, Kwasnik suggested that the great variety of terms used by participants to label the same kinds of objects demonstrated that a document could be

classified in a variety of ways, and thus lead to the need for supporting multiple classification.

Case (1991) investigated the way in which historians categorize and store information they have collected. General findings from the interview with twenty participants were comparable to those from Kwasnik (1989), in that similar dimensions (such as form, topic, and purpose or use) were found to be important. An interesting observation made in this study was that historians consider four conceptual levels of storage in order: physical space, form, topic, and treatment/purpose/quality. As Case summarized, "it is the physical space that is first given priority in document location, and then the physical form of the document. It is the third level that constrains the factor typically of most concern to the information scientists: that of the specific topics of the document" (p.664). The most specific level, which concerns treatment (e.g., intellectual genre), purpose or use, or quality, is invoked last. Given that people deal with physical package of information within the physical environment, it is not surprising that physical constraints are considered first in filing documents. Perhaps the more interesting part is that participants in this study made a topical categorization and then proceeded to make more specific distinctions based on contextual attributes such as purpose and quality (e.g., 'good' example for a specific argument). In fact, a similar pattern was implied in Kwasnik's (1989) findings. In a large number of cases, 'form' was used as the head noun, followed by qualifiers representing other dimensions (e.g., books to be used for a class). Kwasnik also stated that "neither the document attributes nor the situational attributes can be considered independently" (p.156). There is an interaction of dimensions, and while a document's intended use or value is often the most important factor for a classificatory decision (the end result), its content (topic) is a 'given' factor constituting the basis upon which the further evaluation of its relevance to a task can be made. At this

point, it is worthwhile to recall Barsalou's (1991) theory of the complementary roles of ad hoc goal-derived categories and taxonomic categories. According to his theory, taxonomic categorization is made first to build a world model which is people's knowledge of their physical environment. Then, in the presence of a specific task, goal-derived categories are formed on top of the taxonomic categories. This account appears to adequately explain the above findings.

Whereas the studies reviewed above are concerned with people's organization behavior in physical environments, there are studies addressing the same question in electronic environments. Barreau (1995) and Barreau and Nardi (1995) are well-known early studies. Barreau (1995) explored the factors that influence people's classification decisions in their electronic environment. More specifically, Barreau investigated "whether the factors which influence classification decisions in an electronic environment were consistent with the factors that Kwasnik observed for physical documents in an office" (p. 327). In her study, Barreau interviewed seven managers about their use of personal information management (PIM) systems, using a methodology similar to Kwasnik's. Overall, the findings were analogous to Kwasnik's findings. It was reaffirmed that document attributes were not the sole consideration in making category decisions. The context in which documents are created and used has significant impact on the way people identify and manage documents within their PIM systems. It is noted, however, that document creation and usage in the PIM systems were more dynamic in nature, partly due to the temporal characteristics and the variety of the tasks performed. Documents are organized to support the current project, and "rules that are applied for a period of time to reflect the priorities of the moment may soon be abandoned or forgotten" (p. 337). In addition, a pattern of satisficing strategies was discovered in that the

managers usually did not file documents using subdirectory features and chose to leave them all in one directory.

People's reluctance to use a hierarchical organization of directories or folders was also reported in studies addressing how people manage and organize their bookmarks or emails. Keller et al. (1997) described the various obstacles people encounter using the bookmarking feature in a web browser, and claimed that organizing bookmarks within a hierarchical folder system is particularly challenging because "a single piece of information is often relevant in multiple ways, and thus not easily categorized within a single folder" (p. 1104). Not only does a folder system make it hard to decide where to put a bookmark, it also requires users to remember their decision when they need to access one. Abrahams et al. (1998) stated that, "Users must continually tradeoff the cost of organizing their bookmarks and remembering which bookmarks are in which folders versus the cost of having to deal with a disorganized set of bookmarks" (p. 44). Their survey of 322 Web users and analysis of bookmark archives of 50 users indicated that the majority of users choose not to organize bookmarks into folders. As long as the list is easily scanned (with a threshold of 35 bookmarks), users prefer to have an unstructured list, not only because it is easier but also because they want to retain their chronological order. Beyond the threshold of 35 bookmarks, users create folders incrementally as the number of bookmarks increases. A similar pattern was found in the way people manage their emails. Whittaker and Sinder (1996) found that people usually leave messages in their inbox and do not try to maintain a folder structure. There could be several reasons behind this behavior. First, the cognitive difficulty of filing, as discussed in Malone (1983), was noticeable. Moreover, the resulting folder structure may not be useful in retrieving messages later. As one participant put it, "any piece of information

longer than five lines has at least several axes along which you might want to look it up and it really depends how you're coming at it and what you're thinking about at the time" (p. 279). That is, since the way people conceive information within a message and thus categorize it can be changed over time, filing requires anticipation of future use beyond the current context.

More recently, Gottlieb (2001, 2003) investigated classificatory behavior of users in their creation of folder structures and assignment of bookmarks within the structure. One of the purposes of the study was to see whether the factors identified by Kwasnik (1989) and Case (1991) as affecting people's categorization of information in physical environment are relevant to explain bookmarking behaviors. However, rather than examining people's own collection of materials, participants with similar backgrounds in finance were recruited and asked to categorize the same set of internet documents (web sites). Given the structures created by individual participants, customized questionnaires were developed for each participant to solicit the reasons or motivations of particular classificatory decisions made in the creation of their own folder structure and the placement of specific items. Contrary to the highly contextual basis for organizing materials found in other studies, including Kwasnik (1995) in the physical environment and Barreau (1995) in the electronic environment, content attributes (as opposed to context attributes) were found to be the most frequently cited factors affecting their categorization decisions. As Gottlieb acknowledged, this result might be attributed to the laboratory characteristics of the study. Similar observations were made in the area of cognitive psychology. When a specific context is not given in an experiment, people tend to bring in their knowledge about categories that is most likely relevant across situations. That is, category decisions are made based upon context

independent features. Another interpretation given by Gottlieb is that, even though an attribute such as author or topic is normally considered as intrinsic to a document, its meaning may vary depending on the individual's take on the information. By giving subjects the same set of materials, the investigator could observe not only what factors or attributes were used to make classification decisions, but also whether the resulting classification decisions are similar or different. It was found that, even when people used the identical set of attributes (e.g., topic and publisher) to make their decisions, the resulting classifications could be quite dissimilar; on the other hand, when people relied on different attributes, the end result could be quite similar.

2.1.4 Conclusion

Our understanding of cognitive categorization has been greatly extended over the last several decades. At first, categories were assumed to have clear boundaries with definitions, and to simply mirror the discontinuities in the environment independent of human beings. Both assumptions were proved to be incorrect by Rosch and other researchers in the 1970s. Categorization was shown to depend largely upon human perceptual functions, and empirical evidence indicated that categories have internal structures characterized by central tendency (prototype effects). Many models of category structure were proposed, mainly based on the idea of similarity to central prototypes or of family resemblance among exemplars. From the 1980s, theory-based approaches moved the research a step further, to show the impact of people's knowledge and models of the world on categorization and, thus, the selective nature of categorization. More recently, in addition to the role of background knowledge, other contextual factors were also studied. Nowadays it is generally accepted that categorization is

dynamic and a context dependent process.

Barsalou's research on ad hoc or goal-derived categories has been reviewed in detail in the previous section. Since it provides an account of cognitive behavior in the context of the activities of everyday life, Barsalou's theory of goal-derived categories has been adopted by many researchers in their field of research, including managerial decision making (Kahneman & Miller, 1986), consumer behavior (Ratneshwar et al., 1996, 2001; Felcher et al., 2001), problem solving (Chrysikou, 2005), medical diagnosis (Custers et al., 1996), etc. This theory appears to be of particular value for explaining the cognitive processes behind information seeking, use, and management.

Categorization of information objects involves inextricably related dimensions including the object's physical package, its content, individuals' situations which make them highlight or overlook certain aspects of it, its current usage and potential relevance, etc. Because of this complexity, the cognitive effort required to categorize information can be overwhelming.

Information organization within information systems often imposes a hierarchical structure and relies on controlled vocabulary. However, many researchers have noted variability and subjectivity of individuals' perception of and interaction with information objects. Research in the area of personal information management further demonstrates that the way people organize and access information is highly context dependent. It follows that any information object can be categorized in a variety of ways. Balancing between the flexibility and stability of information access systems remains a challenge.

2.2 Social information space

2.2.1 Introduction

While a personal information space is constructed and maintained by an individual and, thus, research addressing human categorization and information behavior at the individual level bears relevance, research related to social information spaces is concerned with patterns of choices at the group level. In this section, two well-established research areas – citation analysis (bibliometrics) and collaborative filtering (recommender systems) – that study relationships or associations derived from aggregated data will be reviewed. An emphasis will be placed on the fundamental assumptions on which the methodological approach of each area is based.

2.2.2 Bibliometric Methods – Citation analysis

Bibliometrics offers statistical methods for studying the process of scholarly communication embodied in published works and the resulting intellectual structure of scientific disciplines (Pritchard, 1969; Borgman & Funer, 2002). Among those methods, citation analysis is the best known approach. Bibliometric methods including citation analysis traditionally have been used in the context of scholarly works. However, their applicability to a broader range of contexts has been increasingly recognized, especially in relation to hyperlinked data on the Internet.

Essentially, citation analysis is based on the relationships between cited documents and citing documents. A citation relation, in network analytic terms, is a directed relation from the citing document to the cited document. Further relationships among documents can be defined based on citation patterns, especially joint citation. In this sense, citation analysis

shares conceptual ground and formal structure (consisting of nodes and links) with network analysis.

At a high level, the field of bibliometrics is characterized by the magnitude of the data and the interest in aggregated patterns. Bibliometrics and its associated methods may be appropriate for many types of research inquiries concerning interrelationships among objects in a large aggregation.

2.2.2.1 Studies using citation analysis

2.2.2.1.1 Assumptions

A basic assumption underlying most types of citation analysis is that a citation is an indicator of influence. As Wilson (1999) says, "[a] document is cited in another document because it provides information relevant to the performance and presentation of the research, such as positioning the research problem in a broader context, describing the methods used, or providing supporting data and arguments" (p. 126). This interpretation of citation, called the normative theory of citing, draws upon Merton's norms of science, which states that scholars internalize and commit to the norms that obligate them to give credit for ideas and to specify the sources of the knowledge upon which their research is based (Wouters, 1999). From this perspective, citations are regarded as a device for acknowledging intellectual debts. It then can be argued that "the research that scientists cite in their own papers represents a roughly valid indicator of influence on their work" (Cole & Cole, 1973, p. 220). Having assumed that the cited work had an impact on or was instrumental in pursuing the research reported in the citing paper, it follows that the number of citations a paper receives in a subsequent body of literature reflects the degree of its influence or importance in the research area. This stance is

supported by the findings of Cole and Cole (1971) that citation frequency was highly correlated with several other measures of prominence or quality such as peer ratings or grant awards: "The data available indicate that straight citation counts are highly correlated with virtually every refined measure of quality. . . There can be little doubt that large differences in the number of citations received by scientists do adequately reflect differences in the quality of the work"(Cole & Cole, 1971, p.28).

This interpretation of citation count, however, has been continually challenged and questioned. The doubt about this assumption in part results in a number of studies investigating the actual function or role of citation or the motives of citing authors. The development of a distinct line of citation research that is more interpretative and constructive stems from the criticism of the normative theory of citing and the interpretation of citation data based on the presumptive arguments (see the section on Studies on citation behavior).

Perhaps a more important assumption, partly intuitive and partly informed by the normative theory, is the relatedness of content between cited and citing papers. Wilson's (1999) description of citing behavior (quoted earlier) implies that there is a relationship between the substantive content of the two documents. Small (1978) points out that, because of the implicit assumption on a conceptual relationship, we usually expect to be able to connect a certain portion of the citing work to a cited work, and attempt to find the author's rationale for citing particular works. The relationship between cited and citing papers further entails possible relationships between papers cited together. "If two documents are jointly cited by another document, they jointly contribute to the content and impact of that research document, and are associated by their role in that research document. Accordingly, the more two documents are co-cited from a body of literature, the greater is the association of their

content, in the opinion of the authors of that body of literature. This leads to the cocitation analysis and its application in literature mapping and visualization studies" (Schneider & Borlund, 2004, p.528).

2.2.2.1.2 Evaluative studies and relational studies

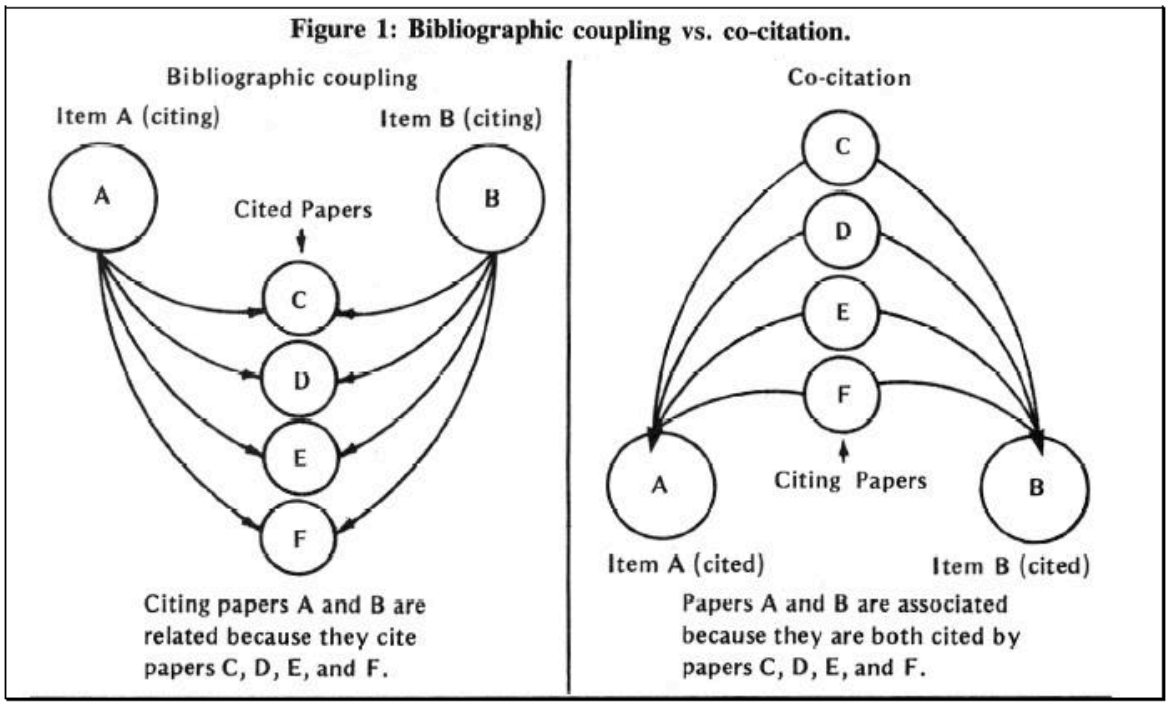
Broadly speaking, studies employing citation analysis can be divided into two groups depending on their general purpose or orientation: 1) evaluative studies, and 2) relational studies (Borgman & Turner, 2002). It can be seen that these two groups of studies are based on the assumption of influence and the assumption of conceptual relationship, respectively.

In evaluative studies, citation counts are used as indicators or measures of impact, quality, or performance. Commonly used units of analyses are individual publications, journals, authors, and research groups (White & McCain, 1997). The number of times a unit (a document, an author, etc.) is cited is the fundamental measure used in most studies. What this count actually measures is usually defined with reference to the unit of analysis and qualified in the specific context of a study. For instance, citation counts can indicate relative performance of a researcher, the level of influence of a paper within a subject domain, or quality of a research institution. The result of this kind of analysis often produces a ranked list, which may be used as a basis for policy decisions (Borgman & Furner, 2002).

While evaluative studies are based on direct counts (or sometimes normalized frequencies) belonging to individual units, in relational studies co-occurrence of certain features (e.g., cocitation) is often used for measuring associations between units (e.g., pairs of highly cited documents). As in evaluative studies, various analytic units such as documents, authors, and journals have been used. The analytic techniques generally involve

measurement of similarities, formation of clusters based on similarity measures, and spatial arrangement of clusters in a way to depict their relatedness (Hummon & Doreian, 1989). The clusters represent topics, specialty areas, or research fields, while links between them show possible relationships (McCain, 1990). As a result of the analysis, a citation network or a visualized map can be produced. These maps or networks are then used to understand the overall patterns of communication and the intellectual structure of a domain (White & McCain, 1997).

Two basic similarity measures used in relational studies are *bibliographic coupling* (Kessler, 1963) and *cocitation* (Small, 1973; Small & Griffith, 1974). Papers are bibliographically coupled when they cite one or more papers in common. On the other hand, papers or authors are said to be co-cited when they are cited together by one or more papers published later. The difference of these two measures is clearly explained in the following figure taken from Garfield (1988). As shown in the figure, for a pair of papers A and B, bibliographic coupling measures the number of papers cited by both A and B, while cocitation measures the number of papers citing both A and B.



The basic premise for measuring similarities based on citation patterns is that authors cite earlier works that are conceptually related and relevant to the current work. A pair of papers is more likely to be related in content when they cite many of the same papers (bibliographic coupling) or when they are cited together in a great number of subsequent papers (cocitation). The level of similarity or the strength of relationships is assumed to be proportional to the count of bibliographic coupling or the frequency of cocitation (Calado et al., 2006).

As a tool for mapping science, cocitation analysis has been more widely used since it allows evolutionary perspectives. Introducing cocitation analysis, Small (1973) posited that, "If it can be assumed that frequently cited papers represent the key concepts, methods, or experiments in a field, then cocitation patterns can be used to map out in great detail the relationships between these key concepts" (p. 265). In his view, the intellectual structure of science is composed of interconnected specialties, each of which is represented by a cluster

of highly cited papers. A point of importance is that cocitation analysis enables dynamic linkages. Cocitation links between pairs of documents are not static properties of those documents. The links are constructed based on citations in later literature, outside the documents being linked. It is therefore possible to capture changes in the intellectual structure as they emerge over time, as well as the connectedness of specialties. Small and Griffith (1974) showed how cocitation analysis can be used to map the intellectual structure of specialties, starting from identifying pairs of highly cited documents linked by cocitation. Subsequently, cocitation analysis has been used to map research specialties of many disciplines (see White & McCain, 1989, pp. 140-146).

While the discussion of cocitation analysis so far has been based on the linkages between documents made by their joint citation in later documents, author cocitation analysis, developed by White and Griffith (1981), traces the linkages between authors and produces maps of prominent authors in selected domains. The unit of analysis is not a single document, but a set of documents by an author, i.e., the author's oeuvre. Just as with document cocitation, it is assumed that "two authors are somehow related to each other if they are often jointly cited and that, the more frequently they are co-cited, the more closely they are related" (White, 1990, p.84). The map of authors is drawn such that authors with perceived similarity (based on the frequency with which their works are jointly cited by other authors) are placed closer to one another. Maps produced from author cocitation analysis provide another representation of the intellectual structure of the chosen domains. Clusters of authors in the map may represent subject areas, specialties, schools of thoughts, etc. (McCain, 1990). Bayer et al. (1990) argue that, "While single works of an individual may precipitate scientific revolutions and new scientific paradigms (Kuhn, 1962), it is more generally the case that a

body of writings by a scientist places that person in the intellectual and influence structure of a field." (p. 444).

2.2.2.1.3 Information retrieval/information filtering applications

Although bibliometric methods, especially citation analysis, are used mainly in the context of scholarly communication, in order to understand its processes and structures, the applicability of bibliometric methods has been discussed and tested in other areas. Most notably, from the outset of the development of citation indexes, Garfield has repeatedly emphasized their value as an information retrieval (finding) tool (Garfield, 1955; 1974; 1990; 1994). In his view, a citation index is similar to a traditional subject heading system in that a citation index can be used to bring related works together as well as finding a specific work. The difference is that a citation index is more flexible because it allows associative links and thus facilitates access from different perspectives: "By virtue of its different construction, it tends to bring together material that would never be collated by the usual subject indexing. It is best described as an association-of-ideas index, and it gives the reader as much leeway as he requires. Suggestiveness through association-of-ideas is offered by conventional subject indexes but only within the limits of a particular subject heading" (Garfield, 1955, p.122). In the context of automatic indexing and information retrieval techniques, Salton noted the value of citation data for representing the subject content of documents and suggested that "documents processed in a retrieval system should normally carry bibliographic citation codes in addition to standard content indicators" (1971, p. 109). In a series of experimental studies, Shaw (1990a, 1990b, 1991a, 1991b) showed empirically that citation information can be employed in document representation in retrieval systems, with the following

conclusion: “In the context of the CF [cystic fibrosis] Database and the single link clustering criterion, the capacity of citation descriptions to associate documents relevant to the same query and discriminate between those that are not is comparable or superior to subject descriptions” (Shaw, 1991b, p.683).

More recently, bibliometric methods have been found useful in the context of web link analysis. Papers connected with citation relationships are analogous to web documents connected by means of hyperlinks. We could assume some kind of implicit value judgment or endorsement behind the citation decision and, even though the contexts are different, the decision to link to a specific web document implies a decision of a similar kind, e.g. quality, usefulness, value, etc. With the equivalent formal structure as well as similar assumptions, bibliometric methods have been successfully adopted for developing searching algorithms. Indeed, Brin and Page’s (1998) PageRank algorithm, which is used in the Google search engine, is based on the same assumption. Specifically, the PageRank algorithm takes account of the number and quality of incoming links to a webpage in ranking the page. Another prominent example is Kleinberg’s (1999) HITS algorithm based on the notion of ‘authority’ and ‘hub.’ The HITS algorithm is designed to search the web for authoritative sources on a topic. Based on the analysis of the link topology of the web, Kleinberg proposes that the web consists of ‘authority’ pages, which are authoritative sources with many incoming links, and ‘hub’ pages, which provide collections of links to authoritative sources. In the original algorithm, authorities and hubs are structurally defined, without relying on semantic information such as titles or link texts. Kleinberg's algorithm was adopted in IBM's Clever search engine (Chakrabarti, et al., 1999). The algorithm also has been applied to perform various tasks, including identification of communities (Kumar, et al. 1999) or clustering of

web documents (Chakrabarti, et al. 1998).

2.2.2.2 Studies of citation behavior

Either in evaluative studies or in relational studies, citation analysis is used as a tool for describing or explaining some phenomenon, based on citation counts or patterns. There is another line of studies in which citation behavior per se is the subject to be investigated (Leydesdorff, 1998; Snyder, Cronin, & Davenport, 1995). In a recent ARIST review, Borgman and Furner (2002) noted a trend of interpretative and constructive approaches to studying citation behaviors and suggest that those studies can be categorized as theoretical.

This trend can be traced back to early critics of citation analysis questioning the assumption that citations can be used as valid indicators of impact, quality, importance, or utility. Edge (1977; 1979) argued that, since citation analysis is concerned only with formal communication manifested in publications and, thus, does not measure intellectual influence made by informal communication or through social relations, the result can not adequately represent the influence structure. Gilbert (1977) interpreted citations as primarily rhetorical devices for authors to appeal to their readers. MacRoberts and MacRoberts (1986, 1987) cast doubt on the normative theory of citing by arguing that citing practices are incomplete and biased. Cole (2000) provided a historical review of the critics of citation analysis, especially criticisms of its use for evaluative purposes.

Controversies surrounding the reliability and validity of citation analysis have given rise to a series of studies adopting a qualitative method called citation context analysis or citation content analysis (Small, 1982). In these studies, the contexts of citations (for example, texts near footnote numbers or reference codes), were examined in order to identify what

functions or roles citations have in cited works. Early findings showing that not all citations are the same type (Chubin & Moitra, 1975; Moravcsik & Murugesan 1975) led to the development of classification schemes or citation typologies based on the different cognitive functions citations may have (Cozzens, 1981). The underlying motivation for developing a classification or typology of citation is the hope that, by distinguishing different types of citations, it would be possible to more precisely define what is being measured in the quantitative analysis of citations and improve interpretation of the results.

While many classification approaches examine the nature of relationships between the citing work and the cited work, others focus more on factors affecting citers and study their reasons and motivations for citing particular works or authors. Surveys and interviews are the most common approaches. Shadish et al. (1995) surveyed authors of psychology journal papers about their reasons for citing. Case and Higgings (2000) provided a review of citer behavior studies and replicated the study of Shadish et al. in the field of communication. Although some differences between citing behaviors in the two disciplines were noted, the common high-level finding is that there is a general tendency for authors to cite what they consider as exemplary work or "concept markers" in the research area, and that there is a spectrum of reasons for citing particular works, varying among authors.

In summary, studies attempting to answer questions about functions and roles of citations or about reasons and motivations of citers have contributed to the understanding that citation practices are more complex and multidimensional than the normative assumption suggests. Moreover, citer behavior is increasingly understood to be subjective, dynamically constructed and affected by the situation. With this broadened understanding of citation behaviors, it is possible to draw an analogy between citation decisions and relevance

judgments (Borgman & Furner, 2002). As Harter (1992) puts it, "An author who includes particular citations in his list of references is announcing to readers the historical relevance of these citations to the research; at some point in the research or writing process the author found each reference relevant. Relevance is the idea that connects IR to bibliometrics, and understanding it in one context should aid our understanding of it in the other" (pp. 612-613).

2.2.2.3 Theories of citation

Debates over fundamental assumptions and disagreements on what is being analyzed consequently led to calls for a theory of citation (Cronin, 1981; Cronin, 1984; Leydesdorff, 1998). On the one hand, studies of individual citation behaviors contribute to a move towards such a theory, because an improved understanding of the nature of citations and the behavior of citers can bring insight as to how to interpret citation counts. In addition, the idea of viewing citation as a kind of relevance judgment makes it possible to place citation behavior studies within a more general theoretical framework. However, as Leydesdorff and Amsterdamska (1990) pointed out, it is important to recognize the differences between studies of individual behaviors and studies of aggregate citation patterns, which is a distinction between micro-level and macro-level research.

In citation analysis, what determines the prominence of authors or makes connections between cited works is the aggregation of citation records. From the outset, this point has been emphasized by the developers of citation indexes in various terms. For instance, Small and Griffith (1974) stated, "Many of the relationships we have uncovered are, of course, known to the specialists themselves, since they were established by their own citing patterns, but the perspective this method offers is far broader than can be achieved by any individual scientist. This is the crux of the method: the observed relationships are in

substance those which have been established by the collective efforts and perceptions of the community of publishing scientists" (quoted in Wouters, 1998, p. 226). While the individual citation decisions are subjective, it is argued that biases tend to cancel out by aggregating those subjective decisions on a large scale (Aaronson, 1975), and the aggregate data represent the degree of consensus of scholars (Davenport & Cronin, 2000). In response to criticisms over citation analysis, White (1990) made an analogy with political studies regarding election data, saying, "Citing an author is in some ways like voting for a candidate" (p. 90). While an investigation of individual voters' behavior, including their reasons to (or not to) cast a vote, constitutes a proper study, there are other studies which completely ignore underlying behavioral or psychological factors of the votes, and are concerned only with the magnitude of the counts and the overall distributions. Likewise, he argues, although studies on citer motivations or functions of citations enhance our understanding of the citation process at an individual level, those studies can hardly dispute the value of aggregated citation counts. In other words, when the data are the overall magnitudes of citation and cocitation, individual differences do not matter. Commenting on Edge's (1979) criticism of author cocitation, he again stresses that a relationship between a pair of authors established by cocitation analysis is not the result of single citation, but by the fact that a large number of subsequent authors over time "*jointly* perceive the relationships (or lack of them) among key writers in their field" (p.92), exceeding a certain threshold.

It is not rare to observe certain regularities or patterns emerge from a large scale dataset in various areas of human activities, despite the diversity of individual behaviors and the differences in circumstances. In bibliometrics, distribution laws such as Bradford's law (the distribution of papers on scientific journals) and Zipf's law (the distribution of word

frequency) demonstrate this tendency.

Recently, theoretical discussions based on this insight have appeared. In an effort to seek a theory of citation, Leydesdorff (1998) suggests that we consider citations as dynamic relational operations, which relate cited and citing pairs in a dual-layered citation network consisting of a layer of social relations among scholars and a layer of cognitive relations among communications (reflected in text). By the 'recursive' nature of the relational operation (meaning that citations link to texts which have citations referring to other texts), a network structure emerges and a citation has a position within the structure. He further argues that, while citations can be observed as individual events, citation analysis is intended to trace the relational operations of the network and the meaning of citations should be studied in terms of distributions at the network level.

2.2.3 Recommender Systems Research

Recommender systems have been studied as a promising approach to cope with information overload. In many e-commerce sites, such as Amazon.com, recommender systems are being used with a certain degree of success. Recommender systems are often classified by how the system is designed, and there are three broad approaches to designing recommender systems – content-based filtering, collaborative filtering, and a hybrid approach (Mirza, 2001; Peruguini et al, 2004). In the next section these approaches will be discussed as traditional approaches. Recently, with increasing interest in social network and link analysis in many related areas such as web search, studies introducing network methods into recommender systems research emerged. Since these studies can serve as bridges between recommender systems research and other research areas amenable to network methods (including social

bookmarking studies), these studies will be reviewed in detail.

2.2.3.1 Traditional approaches

2.2.3.1.1 Content-based filtering

Content-based filtering, also called information filtering, is based on analysis and comparison of the content of items. Commonly, a user profile is constructed using the content of the items the user has rated and their respective ratings. The profile is then used to predict ratings of other items. Items that match well with their profiles are recommended to the users. In the sense that this approach mainly concerns the analysis and matching of content representations, it is similar to information retrieval. Belkin and Croft compared and contrasted information retrieval and information filtering (Belkin and Croft, 1992). According to them, while information retrieval is to meet short-tem information-seeking goals by retrieving items that are relevant to queries, information filtering focuses on removing items that are irrelevant to long-term user interests represented in their profiles.

Content-based filtering has its weaknesses. Since it requires items to be parsed, it only works well with text-based items or items with textual metadata assigned. Content-based techniques provide recommendations based on the degree of matching, which does not have much to do with qualitative factors. Therefore, as the number of items in a given topic or category grows, the effectiveness of the filtering could be diminished. In addition, there is little room for serendipitous finding of relevant items, because the system recommends only items that are similar to those already rated by the user (Shardanand & Maes, 1995; Balabanovic & Shoham, 1997; Claypool et al., 1999).

2.2.3.1.2 Collaborative filtering

The term collaborative filtering was coined by Goldberg et al. (1992), emphasizing the social aspects – sharing collective group knowledge - of this approach. The basic idea is that the system can leverage other people's opinions to provide recommendations to users who have similar preferences. Typically, according to an often cited definition, "people provide recommendations as inputs, which the system then aggregates and directs to appropriate recipients" (Resnick & Varian, 1997, p.56). By relying on people's judgments, either in the form of explicit ratings or implicitly drawn from user behaviors, this approach tries to overcome some of the above mentioned limitations of content-based filtering methods. From the algorithmic point of view, the emphasis shifts from computing item similarities to matching users with similar preferences. User preferences are usually expressed as item ratings (or equivalent measures) and users who have common items with similar ratings form 'neighbors.' Various matching algorithms have been proposed to identify a set of similar users based on correlation coefficients or other similarity measures (Terveen & Hill, 2001). A prediction of what items a user might like or dislike is made based on the ratings or the behaviors of their neighbors. The fundamental assumption is that people's preferences on items are not random and there are persistent patterns in their choices. In other words, people like items similar to those they liked before, and thus people who made similar choices in the past would probably agree on new items (Shardanand & Maes, 1995). In fact, an empirical study using the GroupLens system, which first introduced the concept of 'k-nearest' neighbor group, supports the assumption (Konstan et al., 1997). The result showed that "correlation between ratings and predictions is dramatically higher for personalized predictions [based on nearest neighbors] than for all-user average ratings" (p.81), and that there were systematic

differences in user preferences even within a certain newsgroup where people with relatively close interests gathered (Konstan et al., 1997).

Since collaborative filtering is based on similarity of users, user modeling is the main issue. There are basically two ways to model users – explicitly soliciting ratings/opinions from users or implicitly deriving user preferences from behavioral/activity data. One of the earliest implementations of collaborative filtering, a system named Tapestry, is based on explicit opinions of people as they filter email messages (Goldberg et al, 1992). The GroupLens system, developed for filtering Usenet news articles, asks users to rate articles and uses their ratings to form 'k-nearest' neighbor groups based on Pearson's *r* correlation coefficients. A predicted value for a new article for a user is then calculated with the weighted average of all the ratings of their k-nearest neighbors (Konstan et al., 1997). Explicit modeling methods, however, do not scale well because users are generally reluctant to provide ratings and it is hard to get a sufficient number of ratings for building accurate user profiles, especially in large and heterogeneous systems (Aggarwal et al., 1999). In response to the fact that users are not willing to put time and effort into rating items, researchers turned to the possibility of using other data sources as surrogates for ratings. For example, the time a user spent on reading a document could be an implicit indicator of their interest in that document. Past purchase history is a broadly used rating-surrogate in e-commerce implementations. For discovering patterns or trends from large data sources, various data mining and machine learning techniques have been introduced (Perugini et al., 2004). PHOAKS (People Helping One Another Know Stuff) is a well-known example of using data mining methods for implicit user modeling. PHOAKS examines Usenet postings to find uniform resource locators (URLs) within the messages. The inclusion of URLs is

interpreted as implicit affirmation of interest in or 'endorsement' of the Web sites (Terveen et al., 1997). Siteseer uses personal bookmark folders to model users. A user's bookmark folders are compared with folders belonging to other users to compute the overlaps (set intersection) among the bookmarks and to find qualified recommenders in the context of each folder (Rucker and Polanco, 1997).

Even though a collaborative filtering approach achieves a certain success and constitutes the core technology for many recommender systems, there are well-recognized limitations (Sarwar et al. 2000). The sparsity of ratings (or rating surrogates) always poses a challenge in making an accurate recommendation. In a typical ecommerce system, for example, both the number of items and the number of users are very large and the number of transactions is relatively small. It makes the user-item matrix sparse and, as a result, in a great number of cases the similarity/correlation between two users is zero or too small to be reliable. Many attempts have been made to alleviate the sparsity problem in research prototype systems, but developing a scalable technique to be able to deal with the inherently sparse data is continuing to be an issue (Huang et al., 2004). The so called 'cold start' problem is another common problem. It refers to the situation where the system has no data to make recommendations. When a new user first enters the system, it is not possible to make any reliable recommendation since there is no data on their preferences. Similarly when a new item is added, the system would not be able to recommend this item until a sufficient number of users rate it (Adomavicius et al., 2005; Schein et al., 2002).

2.2.3.1.3 Hybrid approach

In general, two types of information sources are available for a recommender system –

features/attributes of items and transactions/interactions between items and users. Typically a content-based filtering method relies solely on feature/attribute data, while a pure collaborative filtering method makes recommendations based on transaction data without considering item features (Huang et al., 2002). Hybrid approaches combine content-based filtering and collaborative filtering in an effort to utilize both types of information and thus bring the advantages of the two approaches together. The Fab system (Balabanovic & Shoham, 1997) is a representative example. In this system, each user profile is built based on the content of the items they have rated. User similarities are then calculated based on the affinity of their profiles, which in effect are the similarities of the 'content' of the items associated with each user profile. An item is recommended to a user either when the content of the item is similar enough to the user's profile or when the item is highly rated by similar users. By using content information, the system can produce better results especially in those situations where a collaborative filtering method is known to be ineffective, while being able to take advantage of collective group knowledge whenever possible. Specifically, when an item is not rated by many users or when a user does not have enough items in common with other users the system can still make recommendations for the item or the user in question based on content analysis. Many other attempts have been made to combine the two different filtering methods at different stages of the recommendation process with varying degrees of computational sophistication (Basu et al., 1998; Sarwar et al., 1998; Claypool et al. 1999; Condliff et al., 1999; Popescul et al., 2001).

2.2.3.2 Graph-theoretic /network approaches

While traditional approaches emphasize how the recommendations are made, and thus concentrate on mapping users to items and enhancing the accuracy of predictions given

sparse data, other researchers approach recommender systems from a different perspective. This relatively new line of study attempts to conceptualize and model the recommendation process as a mechanism for building a network in which people with some commonalities are connected.

Schwartz and Wood (1993) present one of the earliest attempts to induce social networks based on shared interest evidenced in existing data sources. The authors analyzed email logs obtained from 15 academic/research sites and tried to uncover a social network by tracing email exchanges between people. With a heuristic iterative method, an algorithm is developed to derive a subgraph consisting of people with shared interest in a particular topic, from a large communication graph. Specifically, the approach is to locate a dense subgraph interconnected around a particular 'distinguished' node (person) whose interest/expertise is well-known. 'Interest distance' from the distinguished node is measured by "the proportion of neighbors that two nodes n1 and n2 do not have in common (the symmetric difference set), out of the set of all neighbors of both nodes" (p. 84). After applying the algorithm to about 40 people whose interest is known to the authors, the result was examined to see if it identified people who share the particular interest. Even though some erroneous entries were found, the authors contend that the proximity in the resulting network indeed reflected shared interests between nodes (people) and furthermore the network can be useful for locating experts and potential recommenders on a particular subject.

The Referral Web project at AT&T labs also introduced the network approach for the task of finding experts or recommenders. In this project, a social network was built by analyzing web documents. Their assumption was that, if the names of two individuals appear closely in a document, it implies some kind of connection between them, even though what

exactly is the nature of the connection is not clear. The primary purpose of the analysis is to uncover and represent existing social networks so that users can explore them to find possible connections with other people that they may not be aware of or may have overlooked (Kautz et al., 1997).

More recently, the research front moved further in the direction of investigating and modeling structural properties such as degree distribution or connectivity of the network. As Perugini et al. (2004) point out, "identifying the effects of certain parameters of a developed model for social network graphs is invaluable for setting such parameters when designing a system" (p. 124).

Aggarwal et al. (1999) proposed a graph-theoretic collaborative filtering algorithm, which is in many ways an inspiring attempt to extend existing concepts and employ graph models. In their attempts to address the sparsity of ratings, Aggarwal et al. noted that most collaborative filtering algorithms make use of the ratings of only immediate neighbors. In other words, when predicting the value of a specific item for a particular user, there should be a direct link (based on correlations of their ratings on common items) between the user and those who have rated the item in order to get their ratings. Rather than requiring direct links, the authors developed a mechanism to traverse a graph to find people who have rated the item in question within a short path (with intermediary nodes who have not rated the item) and propagate the ratings along the path. They developed two new concepts, 'horting' and 'predictability', to define the relations between people. User *a* horts user *b* if the number of the items that user *a* and *b* have rated in common is large enough to constitute a major portion of the total items rated by user *a.* If user *a* horts user *b* and a linear transformation can be defined to compute their ratings from one another, it is said that user *b* predicts user *a*. By

definition, predictability is based on the reverse relation of horting. The existence of the horting relationship between user *a* and user *b* (*a* horts *b*) means that, from user *a*'s perspective, there is enough commonality with user *b*, and the linear transformation between their ratings makes it possible to use user *b*'s rating on a particular item to produce a predicted rating of the item for user *a*. Based on these concepts, it is possible to build and maintain a directed graph where nodes represent users and edges correspond to predictability. "The ultimate idea is that predicted rating of item *j* for user *i* can be computed as weighted averages computed via a few reasonably short directed paths joining multiple users. Each of these directed paths will connect user *i* at one end with another user *k* who has rated item *j* at the other end. No other users along the directed path will have rated item *j*" (p. 203). Their graph theoretic approach incorporating indirect links to the recommendation process is influential in other studies.

Mirza et al. (2003) proposed a framework, called 'jumping connections', for studying recommender systems from a graph-theoretic approach. They posited that the basic function of a recommender system is to build a social network of people with shared preferences. They placed their emphasis on the connections a recommendation algorithm makes in a network, rather than on how a recommendation is to be made or how accurate the prediction is. In the 'jumping connections' approach, a dataset consisting of users and items is represented as a bipartite graph, which in social network studies is often referred to as an 'affiliation network.' A bipartite graph is a graph whose nodes can be partitioned into two disjoint sets such that nodes in one set are linked only to nodes in the other set. That is, no two nodes in the same set are linked directly to (adjacent to) each other. Connections among nodes in the same set are induced from their shared relationships with nodes in the other set

(Wasserman & Faust, 1994). Mirza et al. (2003) used a movie rating dataset consisting of a matrix (people x movie) of ratings. A 'jump' function is used to connect people based on movies they have commonly rated. In social network theoretical terms, a 'jump' is a function to transform two-mode affiliation data into a one-mode social network. Different recommendation algorithms can be used to specify various jump functions. For example, a simple algorithm called 'hammock jump' is based on the number of commonly-rated movies (*hammock width, w*). If two people have the specified number (*w*) or more of movies in common, they are linked in the induced social network. After building a social network graph, another graph called a 'recommender graph' is produced such that every movie is attached as a sink node to the existing social network graph. Once the recommender network is built, structural properties of the network can be examined. Since different algorithms or parameters used in the jump function result in different networks, it is possible to examine the effect of different jump conditions on the properties of the resulting networks. For example, in the case of the movie database, the induced network with a hammock jump was well-connected despite the sparseness of the data, until the hammock width (the number of common movies required for a link) was increased to seven. With the hammock width seven, the network was disconnected, but an interesting structure was uncovered. The graph had one large strongly connected component (SCC) and many isolated nodes, rather than a number of small SCCs. The authors attributed this structure to a power law distribution of ratings, and conjectured that other domains where there are distinctive sub-domains (e.g., different genres of books) or a greater variety of tastes, ratings might not follow a power law distribution and thus a recommender graph would have several SCCs representing small communities. As the above example illustrates, by exploring structural properties of each network induced by

different jump functions, it is possible to gain insights into underlying dynamics, and a better understanding of implications of certain parameters in a recommendation algorithm. As the researchers claimed, the jumping connections framework proposed by Mirza et al. (2003) provides a systematic way to design and evaluate recommender systems using graph-theoretic analysis.

Perugini et al. (2004) provides an extensive review of recommender system research from the 'connection centric' perspective, which regards the role of a recommender system as 'bringing people together' in a social network. The perspective is basically the same as suggested in Mirza et al. (2003). Perugini and colleagues claim that mining and using graph structures is "a viable and increasingly popular way to satisfy information seeking goals" (p.125), and the power of this approach is demonstrated by the fact that many successful recommender system designs (e.g. Kautz et al.'s (1997) Referral Web, Mirza et al.'s (2003) Jumping Connections) as well as web search algorithms (e.g. Kleinberg's (1999) HITS) draw heavily upon structural information.

2.2.4 Conclusion

Both citation analysis and collaborative filtering study patterns of associations among entities with different sets of problems and assumptions. In citation analysis, a citation represents a connection between publications judged by an author. Specifically, for example, two papers jointly cited by another paper, or two papers citing the same third paper, are assumed to be connected. Based on the assumption, relationships among published works (or authors) are derived from the aggregation of the individual judgments of connections (i.e., citations), in order to study a higher level

structure, i.e., the intellectual structure of a research domain. The main focus of the review of citation analysis was on how the validity of such assumption has been debated and verified in this well-established scientific field. Citation analysis research, indeed, offers empirical evidence as well as theoretical support that the collective behavior of a group produces meaningful data, regardless of potential differences in motivations, functions, roles, etc.

Another area of research reviewed above, collaborative filtering, is perhaps more directly related to this study. Collaborative filtering is basically concerned with finding and connecting people with similar preferences, interests, or needs. While content-based filtering predicts user ratings based on similarity of content, collaborative filtering recommends resources based on other people's choices and thus provides a better chance of serendipity. The fundamental assumption of collaborative filtering is that people with similar preferences or interests can be identified from the choices they made in the past and these people would continue to agree, to some extent, on new items. In other words, collaborative filtering relies on a persistent pattern of choices at the group level. The considerable success this approach has achieved in many applications, including online shopping sites such as Amazon.com, provides evidence favoring the assumption.

Connection-centric models focus more specifically on the formation of a network of people. The structure of the network is then explored and exploited to better understand relationships between people and resources.

2.3 Social network analysis

2.3.1 Introduction

Social network analysis is the study of relations among a set of actors (Berkowitz, 1982; Wellman, 1988; Wasserman & Faust, 1994; Scott, 2000). Scholars from psychology, anthropology, and sociology have been the primary contributors to theoretical and methodological development in social network analysis (for a detailed account for the development of the field, see Scott, 2000). Its applicability, however, goes well beyond those areas. In fact, network concepts and methods can be applied to any kind of network consisting of nodes and links.

In recent decades, social network analysis has been gaining recognition from various research communities as a promising research approach, and has been successfully applied to a wide range of substantive problems (Breiger, 2003; Wasserman et al., 2005). Its fast growth can be attributed, as Wasserman and Faust (1994) point out, "to the appealing focus of social network analysis on relationships among social entities, and on the patterns and implications of these relationships" (p.21). In addition, its potential capabilities to handle a large amount of empirical data also attract interest.

A multidisciplinary scholarly organization for this field, the International Network for Social Network Analysis (INSNA)[3] demonstrates the field's interdisciplinary character. It holds an annual conference and publishes a refereed journal, "*Connections,*" contributed to by researchers from across many fields including sociology, psychology, political science, economics, organizational research and physics, to name but a few.

As a field of research, social network analysis encompasses a distinctive theoretical

[3] http://www.insna.org/

perspective and a new set of methodological tools to collect and analyze "network" data. The network approach has been used within a variety of substantive areas. Underlying all those studies across different disciplines is a shared theoretical perspective that focuses on relationships among a set of entities (actors) in studying a given substantive problem. The distinctive theoretical perspective emphasizing relational properties, in turn, entails methodological approaches that are distinguished from mainstream social science approaches. Rather than basing the analysis on properties of independent individuals (and aggregations of them) as traditional approaches do, social network analysis primarily concerns *relational* attributes. 'Network' data represent relations.

Social network analysis seems promising for many substantive problems in information science, especially for internet research. Garton et al. (1999) argue that any computer/communication network that connects humans is amenable to social network analysis. There is a large body of research (dominated by sociologists) on 'computer supported social networks' or virtual communities (Wellman et al., 1996), dealing with issues that are relevant for digital libraries. The network approach can also be used to understand user information behaviors where some kind of interaction, either mediated by a communication network or face-to-face, takes place. For example, using a social network perspective and methods, Borgatti and Cross (2003) modeled the way that characteristics of relationships among actors affect information seeking and sharing within an organization.

More obviously, the methods can be used for analyzing properties of the network itself. So called 'small world' network research (Watts, 1999; Bjorneborn, 2004) falls in this category. In addition, some data can naturally be represented as a network and thus analyzed with social network methods. Hummon and Doreian's (1989) study of connectivity in a

citation network is a good example. Studies on scientific collaboration networks (Newman, 2000; Barbási et al., 2002) present another example.

This section will start with reviewing the theoretical perspectives and methodological approaches of social network analysis in general and will move on to a special kind of analysis called affiliation network analysis. Affiliation network data bear particular relevance to this study. Prior studies on different substantive issues using affiliation networks, including the above-mentioned scientific collaboration studies, will be reviewed. Finally, I will discuss the possible application of network theory and methods to the study of social bookmarking systems and their users.

2.3.2 Theoretical perspective

2.3.2.1 Network perspective

The basic theoretical standpoint of social network analysis is that actors are interdependent and that patterns of relations among interacting actors should be taken into account in order to understand human behavior and social processes (Wasserman & Faust, 1994; Wellman, 1988). Social structure, in social network analysis terms, refers to the regularities in or patterns of relations, and it is assumed that individuals' involvement in the social structure has significant impact on their behavior. There are many different models and methods in social network analysis, but all share of the view that a social structure (i.e., network structure) can be conceptualized as persistent patterns of relations.

This conceptualization of social structure departs from the traditional social science approach, which tries to explain social behavior "as the result of individuals' common possession of attributes and norms rather than the result of their involvement in structured

social relations" (Wellman 1983, p. 165) and to describe social structure in terms of the aggregation of characteristics of individuals in a social group. In contrast, social network analysis rejects explanations of human behavior that are solely based on the categorical attributes of individuals and the aggregation of those attributes. It assumes that behavior is not independent among individuals (actors) in a social group, and that the relations among and between the actors have significant effects in constraining or enabling their behaviors, regardless of categorical characteristics. This is what Emirbayer and Goodwin (1994) called the "anticategorical imperative" of social network analysis (p. 1414).

Emirbayer (1997) argued that there are basically two opposing views as to how we conceive of the social world, and consequently how we methodologically approach it. One view, which Emirbayer called 'Substantialism,' regards the social world as 'static substances and entities.' Fundamental units of any inquiry from this view are certain substances which are assumed to be pre-formed and remain unchanged even when involved in dynamic flows such as interaction. In contrast, the other view, called 'Transactionism,' sees the social world as 'dynamic, unfolding relations and processes,' and emphasizes the changing roles played by the units of analysis in a given transaction. In this view, individuals can not be separated from the dynamic relational contexts within which they are embedded. According to Emirbayer, this relational view opens up new directions for investigating social phenomena. In that vein, he argues that key sociological concepts, such as power, inequality, and freedom, can be reconceptualized in terms of relations and emerging patterns of relations among actors.

Social network analysis is consistent with the relational view or 'transactionisim' in Emirbayer's terms. In addition to this general relational view, Wasserman and Faust (1994) suggest the following as central principles underlying the network perspective:

- "Actors and their actions are viewed as interdependent rather than independent, autonomous units
- Relational ties (linkages) between actors are channels for transfer or 'flow' of resources (either material or nonmaterial)
- Network models focusing on individuals view the network structural environment as providing opportunities for or constraints on individual action
- Network models conceptualize structure (social, economic, political, and so forth) as lasting patterns of relations among actors" (p.4).

Similarly, Wellman (1988) posits that the following five paradigmatic characteristics provide "underlying intellectual unity" for structural analysis:

- "Behavior viewed in terms of opportunity and constraints (not inner focus of psychological propensities)
- Analyses focus on patterns of relations between units, not on nominal classifications of units by inner attributes
- Central consideration is how patterned relations affect member behaviors
- Structure is treated as network of networks (overlaps and webs, not discrete groups)
- Methods deal directly with relational nature of social structure to supplement and supplant mainstream statistical methods that make assumptions of independence" (p. 20).

2.3.2.2 Network theory

Emirbayer and Goodwin (1994) posit that "network analysis is not a formal or unitary 'theory' that specifies distinctive laws, propositions, or correlations, but rather a broad

strategy for investigating social structure" (p. 1414). Given a nearly unlimited variety of networks with different types of nodes (actors) and links (relations), it is not feasible to develop a single universal model to explain network phenomena. Instead, network theory building aims at developing and defining network concepts, such as centrality, cohesion, and structural equivalence. With the formal definition of these concepts, relational or structural properties of actors, subgroups, or groups can be studied and structural effects can be tested (Wasserman & Faust, 1994). These formal concepts can be used to investigate a wide variety of phenomena, and provide a way to question cause or effect of social structure in different contexts. In other words, many problems across different contexts can be understood in network terms (Emirbayer & Goodwin 1994). The formal network concepts can then be integrated into substantive theories.

2.3.2.2.1 Granovetter's theorizing of the strength of weak ties

Granovetter's seminal work on the strength of weak ties demonstrates how formal network concepts can be used in empirical studies to address substantive issues and developed into theories. In his work on *Getting a Job* (1974), Granovetter was interested in how people get information about job openings through their social network. Specifically, he was focused on what types of links are involved in transmitting significant information (i.e., information that resulted in job movement), and used the concept of tie strength to feature the different types of links. He selected a sample of technical or managerial workers who had changed their jobs recently. Approximately 56% of his respondents replied that they found the job information through personal contacts, mostly work or work-related contacts. Granovetter noted that family or close friends were rarely mentioned as important sources while their acquaintances

were often credited with providing job information. In other words, in terms of tie strength, weak ties turned out to be more instrumental than strong ties in getting this kind of information. To explain this, he hypothesized "the strength of weak ties," which has been highly influential in network research. While the importance of direct strong ties (such as family or close friends) had been well acknowledged, as Granovetter argued, the implication of weak ties had not been stressed. His theory is that weak ties are more useful than strong ties in many circumstances, especially in transmission or diffusion of information, because weak ties are more likely than strong ties to act as "bridges" between segments in a network (Granovetter, 1973, 1974, 1982). People with strong ties (e.g., close friends, family) often have overlapping contacts and thus form a densely knit network. Information, resources, or influences can easily travel along within their local network, and often redundancy occurs. Weak ties, such as with a mere acquaintance, link individuals in different clusters, and thus enable flow of information between otherwise disconnected parts. Information passed on from a different segment of a network is likely to be new and, thus, useful.

Granovetter's theory on the strength of weak ties has been applied to investigate other substantive themes such as social capital (Lin, 1982), diffusion of ideas (Granovetter, 1982), etc.

2.3.2.3 Network data and datasets

The basic premise of social network analysis is that social structure can be represented as a network. The use of social network analysis methods depends on the availability of relational attributes depicted in a network consisting of a set of nodes and a set of ties (Scott, 2000). Nodes represent actors and ties, sometimes referred to as links, edges, or arcs, represent

relations.

Because the social network theoretical perspective is different from traditional social science in its emphasis on 'relations', the standard dataset for social network analysis is relational rather than attribute data. Wellman (1988) contrasts network analysis with other social science methods, pointing out that network analysis studies social relations while traditional social science research studies personal attributes. Borgatti and Everette (1997) rephrase Wellman's statement in more general methodological terms: "[T]raditional social science studies attributes of INDIVIDUALS (call these monadic attributes) where as network analysis studies attributes of PAIRS OF INDIVIDUALS (call these dyadic attributes)" (p. 243). Borgatti and Everette recognize 'social relation' as just one type of dyadic attribute, and argue that many different types of dyadic attributes can be studied using network analysis. For example, a distance between two points (e.g., the distance between two colleagues' offices) is a dyadic attribute.

Essentially, both actors and relations (or nodes and ties) can be defined in many ways, depending on the substantive research problem. Relations vary in content, direction, and strength (Wasserman & Faust, 1994; Garton et al., 1997). Relational content refers to a specific substantive connection. Typically studied are kinship (e.g., marriage), individual evaluations (e.g., like, respect), or flow of material/non-material resources (e.g., money transfer, information transmission), etc. However, it is important to note that there is no limitation on the possible content of the relation. A relation can be directed or undirected. For example, in the case of a 'friendship' relation there is no specific direction, but the 'supervisor' relation requires indication of the direction. Relations can also be characterized by strength. The strength of relations can be operationalized in a number of ways, such as

intensity of emotion, frequency of interaction, length of duration, etc.

2.3.3 Methodological approaches

According to Wellman (1988), even though the concept of social structure or social network has long been discussed as an important constraining factor of social behavior, this concept has only been used metaphorically. Social science studies have relied heavily on aggregated statistical analysis, for which the assumption of independence of individual units should be made. In other words, traditional social science research methods have no good way to represent and analyze actual 'relationships' directly. Social network analysis, with its formal conceptualization of structural properties and the use of network data representing relations, enables researchers to empirically study patterns of relations.

2.3.3.1 Unit of analysis

Since actors are considered as interdependent, and observations are made on pairs of individuals, the smallest unit of analysis in network analysis is dyads, two actors and the ties connecting them. Larger units constructed from dyads include triads, subgroups, groups, and the whole network. Basically, the unit of analysis consists of a set of actors and the ties among them (Wasserman & Faust, 1994).

Network data can be represented as graphs or matrices. A graph is a very intuitive way to represent connections, and many network concepts such as distance and path are built from graph theoretic concepts. However, since data represented in graphs are not particularly adequate for analytical procedures, matrices are used as input for most software tools[4].

4 There are many computer programs for social network analysis. These programs support a variety of analytical procedures and some incorporate features for

Mathematically, any graph can be transformed into a matrix.

An important concept in the social network context is 'mode.' The number of modes in a matrix is the number of distinct sets of entities represented in it. For instance, in a two dimensional matrix (i.e., a two-way matrix), if the rows and the columns denote different kinds of entities the matrix is called a two-mode matrix, whereas if the rows and columns point to the same kind of entity it is a one-mode matrix.

A traditional social science dataset is usually represented as a case (person) by variable (attribute) matrix, which is a two-way, two-mode matrix. In contrast, a typical data matrix in social network analysis is a square case by case (actor by actor) matrix, called an 'adjacency' matrix, which records a social relation (or other dyadic attribute) among a set of actors. Since the rows and the columns are composed of the same set of actors, an adjacent matrix is a two-way one-mode matrix.

Each cell of an adjacency matrix contains a value denoting the presence/absence of the relation between the corresponding row actor and column actor. Characteristics of the relation are also represented. By convention, if the relation is directed, the senders are recorded in the rows, and the receivers are in the columns. In other words, the dataset is recorded such that a row actor does something to a column actor. The strength of the relation is recorded by the value of the cell, where 1 or 0 represent mere presence or absence of the relation. A number greater than 1 could appear if the strength of the relation is available.

Even though a one-mode matrix is considered the canonical dataset in social network analysis, certainly not all network data are represented as one-mode matrices. A two-mode

---

visualizing networks. An up-to-date list of software tools is available at the INSNA webpage. See http://www.insna.org/INSNA/soft_inf.html. Scott (2000) provides a brief review of network analysis packages in the appendix. Recently, an extensive comparative review has been given by Huisman and Duijn (2005).

matrix called an 'incidence matrix' is commonly used to represent a special kind of network, an affiliation network. Affiliation networks will be discussed in detail later in this paper.

2.3.3.2 Measures

Brass (1995) classifies network measures into three categories according to the corresponding unit of analysis and provides a brief definition of each measure.

1) Typical social network measures applied to ties: indirect links, frequency, stability, multiplexity, strength, direction, symmetry (reciprocity)
2) Typical social network measures applied to individual actors: degree, in-degree, out-degree, range (diversity), closeness, betweenness, centrality, prestige. In addition to these measures, concepts for describing roles of actors are also included in this category: star, liaison, bridge, gatekeeper, isolate.
3) Typical social network measures applied to describe entire networks: inclusiveness, component, connectivity (reachability), connectedness, density, centralization, symmetry, transitivity.

Among these measures, we will take a look at the most broadly used measures for each category, and how these measures are used in some network studies.

2.3.3.2.1.Strength – a measure applied to ties

The strength of a tie is a general notion that describes the nature of the relationship. It can be operationalized in a number of ways depending on the particular context (Marsden & Campbell, 1984). In general, the strength is thought of as a "combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services

which characterize the tie" (Granovetter, 1983, p. 1361).

2.3.3.2.2 Centrality – a measure applied to actors

Centrality is obviously one of the most important concepts in network analysis. Most empirical studies use some kind of centrality analysis to identify the most important or visible actors within the network (Everett & Borgatii, 2005). Conceptually, centrality measures are used to find out who is central or important in a given network or a subgroup network. A wide variety of specific measures have been proposed so far. Centrality measures can be categorized broadly into four groups: degree, closeness, betweenness, and power (Wasserman & Faust, 1994; Faust, 1997). Freeman (1979) suggested categorization of centrality measures consisting of the first three categories and provides exemplary measures for each. The eigenvector-based measure proposed by Bonacich (1972) stands out from the other three categories, and constitutes the fourth category.

Degree centrality is perhaps the most intuitive notion of centrality. The actor with the most ties is considered most important. However, the simple number of ties, the degree of an actor, can be misleading. Depending on the maximum possible degree (determined by the number of actors in the network) or the overall degree distribution in the network, a certain degree measured on an actor could tell quite a different story about the importance of the actor in a network. In order to address this problem, some kind of normalization of the degree measure is often suggested. A more important criticism is that degree centrality only counts direct ties and does not take indirect ties or paths among actors created by indirect ties into consideration. The other category of centrality measures are proposed to deal with this problem with a different conceptualization of what constitutes the 'importance' of an actor.

With closeness centrality, the importance of an actor is determined by relative distance to all other actors. The idea is that, if an actor is relatively close (in other words, a short distance) to all other actors through their direct or indirect ties, they can interact with other actors efficiently and thus can be more influential independent of how many direct ties they have.

Betweenness centrality introduces the concept of 'control.' For example, if an actor lies on a path between actor A and actor B in a communication network, information A sends may or may not get to B, depending on whether the actor between them passes on the information or not. In that sense, the actor between the other actors can control the flow of information. The betweenness centrality of actor *i* counts the number of shortest paths (called geodesic paths) between *j* and *k* (pairs of all the other actors) that actor *i* lies on.

Power centrality takes account not only of ties or paths in which an actor is involved but also of other actors connected to the actor. An actor is considered to be important if he/she has ties to other central actors.

2.3.3.2.3 Connectivity – a measure applied to networks

Connectivity is a graph-theoretic concept. The connectivity of a graph is defined by the reachability between pairs of nodes. Two nodes in a graph are said to be reachable if there is a 'path' connecting them. Network connectivity measures the extent to which actors in the network are connected to one another.

Connectivity together with density is used to measure cohesion or cohesiveness of the network and thus to detect subgroups within a network. A cohesive subgroup is one of the most important themes in network research, because it is provides a definition of the

fundamental sociological concept of group in network analytic terms. In social network analysis, groups emerge through a pattern of connections. Dense connections within a relatively bounded set of actors define cohesive groups – variously called cliques, components, circles, etc. - within a network. Technically, cohesive subgroup analysis is in effect partitioning the network into clusters. There is an array of techniques developed to detect patterns of connections and identify groups, including N-Clique, N-Clan, K-Plex, etc. (Scott, 2000; Wasserman & Faust, 1994).

2.3.3.3 Hypotheses and research questions

Many network studies are descriptive in nature. Social network analysts try to describe networks of relations as fully as possible, and to uncover the underlying patterns or regularities within the network. The description of relational patterns is regarded as of interest in and of itself in network studies, and often constitutes the first step toward further investigation of substantive questions. However, as Knoke and Kuklinski (1982) argue, "If network analysis were limited to a conceptual framework for identifying how a set of actors is linked together, it would not have excited much interest and effort among social researchers. But network analysis contains a further explicit premise of great consequences: The structure of relations among actors and the location of individual actors in the network have important behavioral, perceptual and attitudinal consequences both for the individual units and for the systems as a whole" (p. 13). In other words, rather than simple description, explanatory studies identifying and measuring the cause or effect of network components in a given setting are stronger in the sense that network theories can be applied and tested.

Hypotheses in explanatory network studies can fall into two broad categories.

Hypotheses in the first category examine influences on the formulation of a social network. In this case, the network is treated as a dependent variable. In the other category, the question is how the patterns of relations have effects on other variables such as behavior, attitude, or outcomes of the actors. The network here is an independent variable. As the levels of analysis can vary from actor to dyadic to whole network, network hypotheses can also be formulated at different levels: the individual level, the dyadic level, the whole network level, and mixed levels.

According to Borgatti and Foster (2003), 'direction of causality' is a fundamental dimension that distinguishes network studies in terms of whether the studies are about the causes of a network structure or its consequences. They suggest that the majority of network research has been focused on the consequences of networks, in part due to the impact of structuralism in sociology on the field of network analysis. However, they also acknowledge that network antecedent studies are increasing in number. For example, research on the effect of proximity/homophily addresses network causes. A recent body of contributions from physicists on the evolution of networks is also about factors causing changes on network structures.

Similarly, in their review of organizational studies using social network analysis, Brass and his colleges (2004) organized studies about antecedents and consequences of networks by levels of analysis. Antecedents of social networks include physical/temporal proximity, workflow and hierarchy, actor similarity (homophily), and personality. Consequences of social networks include attitude similarity, job satisfaction and commitment, power, leadership, getting a job, getting ahead, individual performance, and group performance.

A study conducted by Marwell and colleagues (1988) showed how network effect can be examined in empirical studies. In one of the series of studies regarding a theory of critical mass, the authors developed their model of collective action by using computer simulation. In this particular study, they featured some properties of social networks, and investigated how these properties affect the emergence of collective action. In the model of critical mass, each individual is not independent from each other, but each is affected by and affects the others. Individuals make their decisions based on others' decisions. Under such conditions, a few individuals' decisions can cause a "domino effect" that affects all others' decision. To investigate mechanisms of critical mass, the authors introduced social network theory and studied three properties of a network: density, centrality, and cost of communication. Their results showed that all three independent variables have a strong effect on the provision of public goods.

2.3.3.4 Issues

As in other research methodologies, general methodological issues such as accuracy, reliability and validity are to be considered. Methodological concerns arise in most phases of a study, including study design, data collection, and data analysis. Marsden's work on data quality and measurement presents a good overview of the range of issues (Marsden, 1990; Marsden, 2005).

2.3.3.4.1 Study Design – Whole Network vs. Ego Network

Broadly, there are two basic designs of network research, depending on how the network dataset is constructed (Marsden 1990; Marsden 2005). Whole-network studies examine all

the relations among actors within a population: a theoretically bounded collective. Whole-network research is also called complete-network research because complete (or quasi-complete) enumeration of the population (actors and ties linking them to one another) is done. Analytic techniques that employ information on indirect ties, such as subgroup analysis and position analysis, require whole-network data. Most network measures including centrality and connectivity discussed above also assume the availability of whole-network data.

An egocentric-network, on the other hand, is comprised of one focal actor (called ego) and actors (called alters) to which the ego is directly linked. Egocentric-network studies concern personal local networks surrounding individuals and usually do not attempt to link multiple local networks. Questions often studied with egocentric designs are: to compare personal networks of a set of sample actors (drawn from a larger population), to identify similarities/differences in their network composition, and to relate those similarities and differences to variation in outcome variables.

In general, whole-network and egocentric designs are regarded as disparate. However, Marsden (2005) suggested that they are interrelated and can be complementary in the sense that egocentric networks are embedded within larger networks, presenting local parts from the viewpoint of individual units, while the whole-network approach deals with the structural properties of networks at the global level.

2.3.3.4.2 Boundary Specification and Sampling

The problem of identifying the population to be studied or, in more network specific terms, of specifying boundaries on the set of actors to be included in the network, poses considerable challenges. Since network analyses draw on relations and interdependencies

among actors, as Marsden (1990) notes, "Omission of pertinent elements or arbitrary delineation of boundaries can lead to misleading or artifactual results" (p. 439).

Laumann et al. (1989) outlined two basic approaches to boundary specification: the realist approach and the normalist approach. The 'realist' approach defines a network as perceived by the actors themselves. In other words, actors who share an identity as belonging to a group comprise a 'realistic' network. In contrast, in the 'normalist' approach the researcher determines the boundaries based on their theoretical and/or analytical considerations. In this case, actors in the set might not recognize themselves as related (Wasserman & Faust 1994). For instance, a researcher can build an actor set for a citation study with scholars who published papers on a specific topic during a certain period.

In addition, Laumann et al. (1989) explained three general strategies for identifying actors to be included: positional, event-oriented, and relational. The positional strategy makes use of certain characteristics of actors or membership criteria. Examples include people employed in an organization, students attending a school, etc. The event-oriented strategy is based on involvement in particular events or activities. Freeman and Webster (1994) used an event-based approach defining 'regulars' at a beach as persons observed three or more times during a certain period. In cases where there are no relevant positions, reliable actor characteristics, or events to be used to define a comprehensive list of actors, the relational approach based on 'connectedness' can be used. Snowball sampling (Erickson 1978; Frank 1979) is the most well-known technique falling in this category. The basic process is to start out with a small number of sample actors (called the 'seed' sample) and expand the list such that actors who are linked from or nominated by one or more actors in the current sample (at first the seed sample, and later the sample built cumulatively over the rounds of recruitment)

are added, until few or no new names appear to be added. It should be noted that, by the nature of the technique, the resulting network is likely to be well-connected and actors in the seed sample tend to be relatively central (Scott 2000, p. 56).

Traditionally most network studies are focused on well-bounded social groups with a relatively small size. Therefore, among the above mentioned strategies for identifying actors, the positional method is most often used. In those cases, complete enumeration of actors within a group and measurement on ties among all the actors can be easily accomplished once the target is defined and boundaries are set. However, there are some other cases where boundaries are unknown (as in the above example of 'regulars at a beach') or the population is too large to be manageable. Sampling problems arise in those cases.

Even though it has been shown that some basic parameters, such as average number of degree or density (Granovetter 1976), can be reliably estimated from sample data, according to Scott (2000) network sampling is highly problematic because of the potential loss of relational information as well as the lack of a well-established model for assessing the reliability of the sampling method. Burt (1983) gives a rough estimation of the possible loss of relational information; when the sample size is k percent of the population, (100-k) percent of relational information is lost when actors are randomly selected from the population. The larger the population, the more likely the sample actors have many relations outside the sample network, and there is no reason to believe that their relations within the sample network are representative of their entire relations.

In order to deal with this problem, many network sampling techniques rely on what Wasserman and Faust (1994) refer to as "chain methods," which trace the links to acquire a 'connected segment' of the network and assume the sample segment is representative of all

the segments in the network (Scott 2000). Snowball sampling is one of them. Doreian and Woodard (1992) describe another link-tracing technique called 'expanding selection,' in which several connections (not just one) with actors in the prior list are required for a new actor to be added. In a later study, they apply this method to identify a network within a larger network. The expanding selection method is used to come up with a large list of candidate actors and then a dense segment is found using k-core clustering (Doreian & Woodard 1994). This group of methods falls into Laumann et al.'s (1989) category of 'relational approaches.'

More recently, probabilistic modeling approaches based on random graph models are being discussed. Even though traditionally small social group studies with well-defined boundaries have been dominant in network analysis, as the substantive domains adopting network analysis grow and larger networks are of interest, the need for more rigorous sampling methods increases. A good review of the body of research on network sampling is provided by Frank (2005), a leading researcher in the area.

2.3.3.4.3 Data collection

Network data can be collected in many ways including surveys, interviews, direct observations, archival records, and experiments. Survey questionnaires have been most commonly used. With the abundance of electronic data such as mailing-list archives or transaction logs, there is increased awareness of the usefulness of this kind of source. Each data collection method brings its own set of issues regarding accuracy, reliability, completeness, and so on.

Issues related to the survey method are well documented in Marsden (2005). Since it

is basically self-reported data, there is room for personal bias or subjective perception. For example, Kumbassar et al. (1994) reported that people tend to see themselves as more central in their relationships with others than they really are. Respondent accuracy is also a big concern. Bernard et al. (1981) found that the respondents' recall on their interaction frequencies with others within a specific time period did not correspond well with observations made by third parties.

In the case of archival data, the validity issue is more salient. Archival data have many advantages: they are relatively inexpensive and, thus, a much larger dataset can be built; they provide unobtrusive measures and can be especially useful when actors are not available for questioning; and it is much easier to obtain longitudinal data from archives than from any other sources. However, all these advantages are meaningful only when ties drawn from archival records are sufficiently close to the corresponding conceptual definition of relationships in the given study. Since the researcher does not have control over the creation of the source data, which was already done outside the study, it is important to understand the conditions under which the data were created and maintained (Marsden 2005).

2.3.4 Affiliation Networks

A typical social network consists of a set of actors and pairwise ties among them. As discussed in section 3, this network is represented as a case by case (actor by actor) one-mode matrix, called an 'adjacency' matrix.

There is a special kind of network, variously referred to as an affiliation network, a membership network, a two-mode network, or a dual network, which is different in many ways from the typical network data. An affiliation network consists of a set of actors and a

set of affiliations (often called 'events') (Wasserman & Faust, 1994; Faust 1997; Breiger 1974; McPherson 1982; Everett & Borgatti, 2005). Since there are two distinct sets of entities, actors and events, an affiliation network is a two-mode network. An affiliation is a broad concept and the term 'event' or 'group' is often used as a general term covering various sorts of affiliations. It may be some kind of event, activity, or issue that brings actors together either physically or conceptually. It may also be formal or informal groups or collectives of some sort. In fact, any common associations that allow us to define a subset of actors can be represented as affiliation relations. For example, in their study of social protest Bearman and Everett (1993) constructed a 'groups by issues' matrix, where protestant groups' involvements in various issues were represented as their 'affiliations' to those issues.

2.3.4.1 Duality of affiliation networks

An affiliation network is represented as a rectangular case by affiliation (actor by event) matrix, called an 'incidence' matrix. An incidence matrix is a two-way, two-mode matrix. In the matrix, each row is an actor and each column is an event (group). A positive value (usually 1) in row *i* and column *j* indicates that actor *i* participated in event *j* (or belongs to group *j*).

As we can see from the way an incidence matrix is constructed, in an affiliation network actors and events are connected by a relation such as membership or participation. Explicit relations represented in the network are between actors and events (i.e., between entities belonging to different sets or modes). Actors are not directly related to each other, neither are events. However, based on the affiliation relations we can derive relations among entities within each set: relations among actors and relations among events. Specifically,

actors are connected by participating in the same events (common memberships) and events are linked by the participants they share (overlapping members). The fact that we can construct both a network of actors and a network of events from an affiliation network explains why an affiliation network is called a 'dual network.'

This 'duality' of affiliation networks, in fact, has significant theoretical and analytical importance. In a seminal paper titled, "The duality of persons and groups," Breiger (1974) discussed theoretical ideas of early structural sociologists such as Simmel (1955) and Nadel (1957), and argued that those theories can be represented in and empirically studied with what he called a "membership network." The basic theoretical insight is that there are basically two different social ties: social relations between a pair of people and membership relations between a person and a collectivity (e.g., family, organization, etc.). Membership relations, however, imply yet other kinds of relations both at an individual level and at a group level, such that individual actors are linked 'through' their relations with groups, and groups are connected through multiple relations of actors. Based on Simmel's theory of intersecting social circles[5] and the dual perspectives embedded therein, Breiger (1974) argued that ties among actors can be defined as the intersection of their affiliations, and ties

[5] According to Simmel (1955), an individual is socially defined by the social circles to which he/she belongs. In a pre-modern society where individuals belong to a small number of tightly bonded social circles (such as kinship groups), their social experiences are confined to those groups. In a modern society, on the contrary, an individual is a member of many diverse social circles (groups), and each person occupies a unique social position in the "intersection" of many social circles. An individual may share membership with other individuals in one or more social circles, but it is not likely that any two have exactly the same memberships. The extent to which affiliations of two individuals overlap indicates how close they are in this social sphere. Conversely since individuals with multiple affiliations form the intersection of the social circles, groups are closer when they have more members in common (Coser, 1977, 189-193).

among groups as the intersection of their members. Beyond the metaphor, he used matrix algebra to show that a two-mode affiliation network can be transformed into a pair of one-mode matrices that are mathematically dual. Starting with an affiliation matrix A, we can define A^T as the transpose of A. The product of the original matrix A and its transpose (A x A^T) produces an actor by actor matrix, denoted by P, in which each cell gives the number of events in which both the row actor and the column actor participated. On the other hand, the product of the transpose and the original matrix (A^T x A) makes an event by event matrix, denoted by G (Freeman, 2003). In this matrix, each cell indicates the number of actors who attended both the row event and the column event. In these dual matrices P and G, as Breiger (1974) noted, "persons who are actors in one picture (the P matrix) are with equal legitimacy viewed as connections in the dual picture (the G matrix), and conversely for groups" (p.184).

From an analytic point of view, two points are noteworthy about the power of affiliation network approaches. First, an affiliation network depicts the pattern of relations among actors, beyond pairwise social relations. Since participations in events form subsets of actors of arbitrary size, ties in an affiliation network in effect link more than two actors at once (Freeman & White, 1993; Faust et al., 2002). Second, with an affiliation network it is possible to investigate the structural properties of the network both at individual and group levels, either separately or in conjunction with each other (McPherson, 1982, Emirbayer & Goodwin, 1994). Breiger (2003) further stated that the affiliation network approach can be extended to examine "linkage between different levels of structure" (p. 21), with an example of Mische and Pattison's (2000) three-level study in which the authors looked at the interrelation of social movement activists, their organizations, and projects.

2.3.4.2 One-mode analysis

Most network concepts/analyses, such as centrality and cohesive subgroups, are relevant to affiliation networks. As described above, a two-mode affiliation network can be transformed into a one-mode dataset. Once transformed, a full range of network analytic methods can be used for this dataset. One-mode matrices derived from a two-mode affiliation matrix contain in each cell a number indicating the 'proximity' of the row and the column entities. For example, in the actor by actor matrix, the number of events both the row actor and the column actor attended is given and indexes their similarity in terms of their event participation. What the number/value designating proximity or similarity specifically means depends on substantive matters and assumptions (Borgatti & Everett, 1997; Scott, 2000). In some cases, the existence of a certain relation (e.g., acquaintance) can be assumed. In other cases, co-attendance can mean increased opportunity or possibility of a relation. In yet other cases, it can indicate some kind of pre-existing commonalties between them, such as shared interests.

Substantive examples of affiliation networks are abundant including corporate boards and directors (Allen 1982; Davis & Greve, 1997), voluntary organizations and members (McPherson, 1982), movies and actors (Watts & Strogatz, 1998), and so on. Ennis (1992) conducted an analysis of the intellectual structure of sociology using an affiliation network consisting of the members of the American Sociological Association (ASA) and their areas of specialty. Based on the 1990 ASA membership directory, where each member reported up to four areas of interest from a list of 54 specialty areas, the author derived a one-mode specialty by specialty matrix. The matrix with proximity indexes based on overlapping members was then used for cluster analysis and multidimensional scaling to uncover the

pattern of linkage among specialty areas. Seven coherent clusters were identified and interpreted, combined with shared cognitive orientations. Newman (2000) studied networks of collaboration among scientists in physics, biomedical research, and computer science based on co-authorship of scientific papers. For each discipline, a one-mode actor by actor matrix in which scientific collaborations constitute ties between actors was constructed. A variety of network measures including distance, strength of ties, connectedness, etc., were applied to the network to analyze the structure of collaboration in the discipline. Barabási et al. (2002) also examined co-authorship structure using affiliation network approaches with different focuses. They took this network as a prototype of large-scale complex networks, and paid particular attention to the evolution of networks.

As seen in the above examples of the sociological specialty network and the collaboration network, it is common in affiliation network studies that once the original affiliation matrix is converted to one-mode matrices, only one of the two (either an actor network or an event network) is mainly used for subsequent analyses. In Ennis (1992) only the specialty matrix was used and the member matrix was not included in the main analysis. In collaboration network studies (Newman, 2000; Barabási, 2002), in contrast, only the actor matrix was used. In these cases, affiliation relations are collected "as an intermediary step toward the construction of a 1-mode network data set" (Borgatti & Everett, 1997, p. 245).

2.3.4.3 Two-mode analysis and representation

Most network analytic techniques require one-mode adjacency matrices as inputs for processing. Therefore, the conversion of two-mode networks to one-mode networks prior to the analysis is often a necessary step. However, it is important to note that there are

reductions of data with the conversion. As Faust (1997) put it, "In going from the affiliation relation to either the actor co-membership relation or the event overlap relation, one loses information about the patterns of affiliation between actors and events" (p. 189). In a derived one-mode actor network, for example, relations are defined in terms of the number of groups or events two actors have in common in their affiliations. Information on which groups or events are involved is lost in the process of conversion. Considering that groups or events themselves can have significantly different structural features, the loss of information can not be ignored depending on the substantive research problems. In fact, increasingly researchers are concerned that "important structural features of the relations between the elements of one mode can only be completely understood if one simultaneously considers the way in which these same elements form relations among the elements of the other mode" (Field et al., 2006, p.100). What is desirable in studying an affiliation network is to look at all three possible patterns of relations: actor-actor, event-event, and actor-event.

Faust (1997) points out that most affiliation network studies measuring centrality of actors or events overlooked the duality of the data. She argued that centrality measures developed for one-mode networks might not be appropriate for studying affiliation networks and a different conceptualization that takes into account "the relationship between centrality of actors and the centrality of events to which they belong, or the relationship between the centrality of events and the centrality of their members" (p.165) is necessary. She discussed five existing centrality measures – degree, eigenvector, closeness, betweenness, and flow betweenness – and the application and interpretation of those measures for affiliation network data.

In order to analyze two mode data without reducing the data, the idea of representing

an affiliation network as a bipartite graph was suggested by Wilson (1982). A graph is bipartite if its nodes can be partitioned into two mutually exclusive subsets, every tie links nodes from different subsets, and no tie is within a subset. Since an affiliation relation always links an actor and an event, a bipartite graph can be constructed such that a set of actors and a set of events comprise two different subsets in the graph. The number of nodes in this graph is the sum of the number of actors and the number of events (Faust 1997; Borgatti & Everett, 1997). A bipartite graph not only allows a representation of an affiliation network without losing any information but also enables direct extensions of existing network methods based on graph theoretic concepts, since it is a graph (Everett & Borgatti, 2005).

Using a bipartite graph approach, Borgatti and Everett (1997) discuss analyses of two-mode data with an emphasis on how to apply traditional network analytic concepts and techniques such as density, centrality, and subgroup analysis to affiliation networks and what concerns arise in the applications. For example, the maximum number of possible links between nodes, which is used as a standard denominator for normalizing the observed value when calculating density of a network, is different in two-mode data because there can be no links within a set, but only between sets. Everett and Borgatti (2005) further develop the discussion of two-mod data analysis and apply graph centrality measures directly to two-mode data with normalizations.

Another possible representation of an affiliation network is a Galois lattice (Wasserman & Faust, 1994). According to Freeman and White (1993), Galois lattices provide a better way to visualize an affiliation network, because whereas a bipartite graph shows only ties between different subsets, a Galois lattice can reveal relationships among actors or among events as well as between actors and events. As Mische and Pattison (2000)

put it, “Galois lattice analysis makes possible a simultaneous graphical representation of both the ‘between set’ and ‘within set’ relations implied by a two-mode data array” (p. 170). However, it is not useful for a large dataset because the picture quickly gets too complicated to see any pattern as the number of elements to be included increases.

2.3.5 Conclusion

We have reviewed theories and methods of social network analysis. Social network analysis has been used in a variety of substantive areas. While social network analysis was originally concerned with a relatively small dataset in sociological or anthropological studies, network concepts and measures such as connectivity have been successfully applied to very large datasets, including studies of internet link typology.

In many research areas, the social context of a problem at hand is of interest. However, social connections or structures are not always observable or salient in many datasets, especially if the data are generated or collected without the explicit intention to establish such connections. Affiliation networks can be a very useful analytical tool in such circumstances. They are also particularly useful in representing a situation where actors do not necessarily have a social tie but are involved in the same kinds of events or with the same artifacts. In fact, it is possible to induce social networks from any dataset that can be represented as a bipartite graph. For example, Perugini et al. (2004) noted the possibility of building an affiliation network for collaborative filtering.

Social bookmarking data can be represented as tripartite graph, because there are three distinct kinds of entities in a system: user, tag, and resource. Because of the complexity of computation required to process a tripartite graph, a tripartite graph is often transformed to

a bipartite or a unipartite graph. In the case of social bookmarking data, there can be three bipartite graphs: user-resource, user-tag, resource-tag. Mika (2005) used two bipartite graphs, user-tag and resource-tag, to induce networks of tags based on co-occurrences, in an attempt to derive semantic relationships among tags based on tagging data.

As reviewed above, once a one-mode network is induced from a bipartite graph (two-mode data), a full range of network analytic concepts and methods can be applied to the network. This means, for example, that we can find cohesive subgroups in a network of users, either based on their common items (possibly representing shared interests) or on their common usage of tags. Even though we have no data on direct social ties among users of a social bookmarking system, affiliation network analysis presents a theoretically and methodologically sound way to investigate the social dimensions of social bookmarking.

2.4 Social bookmarking studies

2.4.1 Background

Social bookmarking is a new phenomenon taking place in an open space on the Web where people can store and annotate information resources. There are a number of social bookmarking sites with increasing popularity. One of the most important characteristics of such sites is that, by default, all the resources and activities are open for everyone to see. While each user of a social bookmarking site constructs and maintains their own information space on the site, because all the bookmarking/tagging activities of individuals are recorded and made visible on an open area, social dynamics and collaborative effects are brought about. Information objects contributed by individual users are aggregated and form a kind of collective

knowledge of the community.

Often, in adoption of a groupware or collaborative system, the hardest part is to get everyone to use it. There is a well recognized problem of conflict between the self-interest of individual members and the interest of the collective (Kalman et al., 2002; Yuan et al., 2005). Social bookmarking effectively reverse the situation with its openness. In a social bookmarking site, the primary activity of bookmark posting is done by users for their own interests. Any social or collaborative effect is a by-product of doing this personal activity in a public space, without additional effort by individual users. With immediate personal benefit attracting people, it is a lot easier to achieve 'critical mass' for group benefits. As Grudin (1994) suggested, a collaborative system should provide its users with direct personal benefit in order for them to justify the effort they need to put in to use it. Social bookmarking systems seem to achieve the balance between the work and the benefit. With a simple bookmarking/tagging mechanism, the overall cost on the user side is relatively low, while personal and collective interests can be met without conflict.

The rapid adoption and growing popularity of social bookmarking have attracted attention from research communities. The unprecedented amount of end-user generated metadata given in the form of tags has been a powerful attraction, with a vision of building a bottom-up taxonomy, called a 'folksonomy'. The action of tagging an item implies that the person sees some value in the item, even though we do not know in which context or for which purpose he/she thinks it would be useful. The tags the person assigns to the item depict what he/she thinks it is about or the purposes for which it is useful. In aggregation, the number of people that tagged the item can be interpreted as an indicator of its value, and the collection of tags

assigned to the item can be seen as descriptors of the content from various perspectives. Another value is in the associations among the three axes of social bookmarking – people, information objects, and tags. The associative patterns that can be derived from the interaction of the three axes of social bookmarking data can be useful in organizing, finding, and evaluating people, information objects, and tags. This point was constantly emphasized from the outset. For example, Thomas Vander Wal, who coined the term 'folksonomy', suggested early on that, by combining any two data points, one can find instances for the third: "If you know the object and the tag you can find other individuals who use the same tag on that object, which may lead (if a little more investigation) to somebody who has the same interest and vocabulary as you do. That person can become a filter for items on which they use that tag. You then know an individual and a tag combination to follow" (Vander Wal, 2005).

2.4.2 Research on social bookmarking

Social bookmarking / social tagging sites feature extensive, naturally situated, uncontrolled information environments with massive user participation. Thus, they present a wide range of new research opportunities for understanding user information behaviors both at an individual and a social level, as well as natural patterns and tendencies in information and language.

Google has had remarkable success with its PageRank algorithm exploiting link structure. It leverages collective human intelligence (implicit and explicit metadata created by a large number of people) rather than solely depending on machine processes. This

approach appeals greatly to the research community as a promising way to solve complex problems. Social bookmarking and the notion of a folksonomy fit particularly well with this theme.

Discussions and debates on the social tagging phenomenon were first sparked in blog space with a great interest in its potential as an information organization mechanism. Clay Shirky's well-known blog article, "Ontology is overrated[6]," initiated lengthy debates on tradeoffs between traditional information organization devices, such as classification schemes or ontologies, and social tagging approaches by putting them into opposition. This issue has been taken on by many researchers. The first group of scholarly works on social tagging discussed the strengths and weaknesses of the tagging approach compared to more traditional approaches (Mathes, 2004; Quintarellin, 2005; Guy and Tonkin, 2006). Interestingly, both the limitations and the advantages of social tagging stem from the very same aspect of this approach: the lack of control. By allowing users to use terms of their own choice without imposing any rules, social tagging in a sense provides a solution to the 'vocabulary problem' (Furnas et al., 1987) by accommodating a broad spectrum of user vocabularies. However, there are abundant problems associated with interpretation/processing of these 'freely chosen' terms, including polysemy (a word that has multiple related meanings), synonymy (multiple words that have the same or equivalent meaning), different levels of specificity, acronyms, idiosyncratic choice of word combinations, etc. (Golder & Huberman, 2006). Macgreogor and McCulloch (2006) reviewed various arguments that had been made regarding the pros and cons of each approach. Similarly, Tennis (2006) compared and contrasted social tagging and subject

[6] http://www.shirky.com/writings/ontology_overrated.html

cataloging. Some researchers have suggested an 'ecological' view which focuses on the complementary functions these different approaches can provide. Christiaens (2006) claimed that there needs to be a continuous feedback mechanism between user-generated free-form metadata and more restricted metadata, such as an ontology, to enhance the overall quality of metadata provided to users. Campbell (2006) introduced Husserl's theory of phenomenology to argue that tagging systems and highly-structured systems have complementary relationships, especially in the ways each creates inter-subjectivities or consensus within a community for sharing and use of information resources.

Arguments for social tagging as an information organization mechanism are fundamentally based on the concept of self-organization and emergence (Johnson, 2001). As Campbell (2006) puts it, the notion is, "if you let users tag their own resources in their own ways, with their own words, patterns of order will emerge; these patterns will be truer, more convincing, more user-centered, and more useful than the pattern imposed by formal classification schemes. What's more, they will acquire greater accuracy and greater sophistication as more and more people use them." (p. 4). Not surprisingly, initial empirical research on social bookmarking / social tagging addressed whether a consistent pattern emerges from tagging data. Golder and Huberman (2005; 2006) analyzed data gathered from *delicious.com*[7] with an emphasis on the dynamics and the structure of social tagging. They examined tagging activities of users over time with various statistics including the number of postings and tags. More importantly, they analyzed the frequency distribution of tags assigned to specific resources, and found that tag distribution for a given resource (URL) follows a power-law distribution and presents a remarkably stable pattern in which the

[7] The site was originally called *del.icio.us* (http://del.icio.us). The official name of the site and its URL address were changed in July 2008 to *delicious.com*.

relative proportion of each tag, established after about 100 postings, remains fixed over time. More recently, Halpin et al. (2007), with a larger dataset drawn from *delicious.com*, reported the observed power-law distribution of tags applied to particular resources. This means that a small number of tags are extremely common, while a large number of tags are extremely rare, constituting the 'long tail' of the distribution. It is often conjectured that the long tail consists of either highly specific or personalized (idiosyncratic) tags.

In addition to the statistical characteristics of tagging data, more qualitative analysis of tag usage and tagging behavior (reflected in the use of tags) has been done in some studies. In the above mentioned study, Golder and Huberman (2006) categorized kinds of tags they observed in the dataset and discussed the tension between the shared and the personal purposes carried by different kinds of tags. They speculated that there are two underlying reasons for the stabilized proportions of tags: 'imitation' (people imitate the choice of terms previously made by other people) and 'shared knowledge.' Kipp and Campbell's (2006) preliminary analysis of tagging behavior indicated that there is a wide range of differences in the depth and specificity of tagging across users. For the same resources, topics people choose to represent vary. In addition, many of the tags appear to be related to tasks rather than to subject matter.

The influence of the 'social' nature of tagging systems on the resulting tag usage is another important research problem. Sen et al. (2006) studied how community influences and personal tendencies affect users' selection of tags. They built tagging features into the MovieLens recommender system and applied four different algorithms for selecting 'popular' tags to be displayed. Their analysis has shown that users' selection of tags is influenced by the displayed tags, indicating that people often conform to what others do.

Marlow et al. (2006) analyzed tag usages within the Flickr photo sharing site. They note the differences between Flickr and *delicious.com* in the types of tagging supported in each system. Tag usage patterns observed in this study are different from what Golder and Huberman (2006) found in their analysis of *delicious.com* data. For example, while Golder and Huberman found that there was little correlation between the number of tags used by a user and the number of items they have in their collection, Marlow et al. observed that, as the number of items increase in a user's collection, the number of unique tags tend to grow. The authors attribute the differences in tagging patterns to the distinct features of each system. More importantly, using the 'contact' feature available in Flickr, Marlow et al. found that there is a greater overlap of tags among people connected by a 'contact' network, suggesting a possible impact of a user's social network on his selection of tags.

While the regularities in user activities or in tag distribution bring insights into underlying dynamics of tagging systems, what is more important with respect to utility of social tagging as an information organization/subject access mechanism is whether a coherent semantic structure can be derived from tagging data. Some studies investigate this particular problem with various approaches. Brooks and Montanez (2006) examined tags assigned to blog postings by the author, to see whether documents clustered based on a shared tag are similar. A pairwise cosine similarity of documents within a cluster was calculated and compared with that of randomly constructed set of documents. They found that similarity of documents in a cluster with a shared tag is only a bit higher than documents in a random set. They argued that "tags are useful for grouping articles into broad categories, but less effective in indicating the particular content of an article." (p. 625). Begelman et al. (2006) introduced a more sophisticated clustering algorithm based on co-occurrences of tags.

An undirected weighted graph of tags was constructed to identified semantically related tags. The algorithm was tested with the database consisting of about 20,000 pages and 30,000 tags obtained from a social bookmarking site, RawSugar. Mika (2005) attempted to develop 'lightweight ontologies' from *delicious.com* data, based on the assumption that semantic relationships among tags can be inferred from the structure of tag network. Mika first suggested an abstract model of an ontology with a social dimension, in which actors as well as concepts and instances are represented with a tripartite graph[8]. In the case study with *delicious.com* data (a sample of 51,852 postings with 30,790 unique URLs, 10,198 users, and 29,476 unique tags), a tag is considered as a concept and a bookmarked resource as an instance. From the tripartite graph, two bipartite graphs of interest for the purpose of extracting ontologies were derived and subsequently transformed into two different networks of tags. One network is based on the co-occurrences of tags associated with resources, and the other is drawn from the common usage of tags among users. The resulting networks revealed different clusters of tags, but Mika argued that each showed evidence of semantic emergence to some extent. Mika's work (2005) is one of the earliest studies which adopted graph/network methodology.

While Mika's (2005) work primarily addressed the relationships among tags, the general model can be applied to draw a network of any of the three types of entities in a

[8] A tripartite graph is a graph whose vertices can be partitioned into three disjoint sets, such that no vertices in any one set are adjacent (directly linked). With three types of entities involved in tagging (user, tag, resource) each of which can be presented as a set, social tagging data can be modeled as a tripartite graph. Lambiotte and Ausloos (2006) introduce a tripartite graph for representing *delicious.com* data and present methods for projecting the tripartite graph on bipartite and unipartite graph, for reducing the complexity of the analysis. Other studies reviewed in this section, including Mika (2005), Paolillo and Penumarthy (2007), and Hortho et al. (2006) take similar approaches.

tagging system. Paolillo and Penumarthy (2007) collected and analyzed tagging data related to a particular type of resource, video, from *delicious.com*. They examined each of the three modes of the tripartite network (user, tag, resource), and reported that a relatively weak semantic structure was observed, casting doubts on the coherence of the emergent structure.

As the user population as well as the size of the database grows in a social tagging site, the issue of findability and navigability within the site becomes salient. Currently in most social tagging sites, a tag cloud, a visual representation of popularity of tags being used in the system, is a primary navigation mechanism. Researchers have started to point out the problem of limited search capability and the inefficiency of tag cloud-based navigation (e.g. Millen & Feinberg, 2006; Sinclair & Cardew-Hall, 2007; Begelman et al, 2006; Hortho et al., 2006). Chi and Mytkowicz (2007) introduced an evaluation metric based on Shannon's information theory. They analyzed the *delicious.com* site to evaluate the efficiency of tags as a navigation mechanism, and found that the retrieval/navigation efficiency drops over time. Given the rapid increase of tags and resources being tagged, the lack of effective search and exploration could have a direct impact on the overall usability of these sites. A large portion of the recent research in this area seeks a solution to this problem. For example, Hortho et al. (2006) proposed a ranking algorithm for presenting search results within a social tagging system, based on an adaptation of the PageRank algorithm. Choy and Lui (2006) evaluated similarity of tags by applying latent semantic analysis, and proposed to generate a graphical map using a self-organizing map (SOM) to represent the tag space.

2.4.3 Conclusion

Over the last two years, a large number of academic papers on social bookmarking have been

published, demonstrating the research community's interest in this phenomenon. Indeed, the unprecedented amount of user-generated metadata in a real-world situation (as opposed to a laboratory setting) presents a broad range of intriguing research problems, both theoretical and practical.

Perhaps the core value of a tagging system, from a research point of view, resides in the patterns of associations made by the distributed tagging activities of a large number of people. In fact, the vision of a 'folksonomy' as 'bottom-up' organization based on the emergent structure of tags has attracted great attention and debate. The major portion of current research seeks to understand tagging patterns, in an attempt to assess the potential utility of tagging data for information organization and access. With empirical observations of tag distributions and patterns, our overall understanding of the dynamics of tagging is improved. However, in terms of the coherence of the structure, studies adopting various approaches to derive the structure report mixed results.

There are three types of entities in a social bookmarking system: tag, user, and item. With the associations made by tagging, we can derive not only a network of tags, which is most commonly being investigated, but also a network of items and a network of people. Other than general discussions on their implication and possible applications such as recommender systems, there is little empirical research on the structure of these networks.

Chapter 3. Methods

3.1 Study design

The overarching purpose of this study is to improve our understanding of the information space of social bookmarking. The approach taken in this study is to conceptualize the space as the aggregation of personal information spaces, which is constructed by the bookmarking activities of individual users. The structure of the information space can then be studied in terms of unions and intersections of personal information spaces. In the framework of social network analysis there is a special kind of network, called an affiliation network, that is very suitable for representing this conceptual picture. An affiliation network provides a representation of the theoretical concept of intersecting social circles, and allows investigation of relations among people based on their joint participation in groups of a sort. Analogous to social circles, the bookmarked information objects comprise information spaces, and users are connected by intersecting information spaces. Together with this abstract representation, the methods of social network analysis provide the analytic framework for this study.

The main focus of the study is on the problem of identifying and characterizing shared interest space(s) within the large-scale information space of a social bookmarking site. The basic underlying assumption is that choices people made in the past can serve as implicit indicators of their interests or preferences, and that

non-random patterns emerge in aggregation which provide a basis for identifying similar people. That is, we can infer shared interests among users of a social bookmarking site, based on their past bookmarking behavior. The well-established research areas of citation analysis and collaborative filtering techniques, reviewed in the previous chapter (under the heading of social information space), provide theoretical and empirical support for the assumption.

In order to address the research problem of a shared interest space, this study is designed to be carried out in three phases, asking each of three separate yet closely related questions: First, to what extent are bookmarking activities accumulated and how much overlap is there in a social bookmarking site? Second, can users of a social bookmarking site be connected based on their shared interests (their common possession of bookmarks) and, if so, how? Third, is it possible to identify communities of interest within the network?

As a setting for the study, a popular social bookmarking site, *delicious.com*[9],

[9] *Delicious.com* (formerly del.icio.us; http://delicious.com) is a "social bookmark manager," where registered users save their bookmarks on the shared web site. When users add a bookmark, the URL and title of the web page as well as the creation-time of the bookmark are recorded. In addition, users can choose to "tag" the bookmark.

When a bookmark entry is created, it is immediately shown on the front page of the site, where several of the most recent posts are displayed. Here, not only the user who posted the bookmark, but anyone can see the entry. Each entry consists of the link to the web page with the title as link text, the list of tags, the username of the person who created it, the number of other people who have saved the same page (URL), and the time at which it was added. From the point of view of the user who added the entry, the moment he/she posts a bookmark, he/she can see how many other users bookmarked the same page and further how they tagged it and when they added it.

In *delicious.com*, three basic entities (user, item (URL), and tag) have a page per each and every instance of the entity within the system. Each user has a

was chosen. *Delicious.com* is known as the first and one of the most successful instances of social bookmarking. With its relatively long history and broad user base, the site can serve as a strong example of the aggregated information space of social bookmarking.

Finally, an important decision made in designing this study needs to be mentioned – the decision to draw a network based only on bookmark posting behavior and not tagging behavior. It may seem intuitively appealing to use all the available information, both the information objects bookmarked and the tags assigned to them, to build connections among users. Ideally if we find someone who is interested in the same material and also classifies that material in a way similar to our tagging, his/her interests are probably closely related to ours. In fact, there have been studies representing social bookmarking data as a tripartite graph (Lambiotte & Ausloos, 2006), which allows the presentation of all three entities (people, information objects, and tags) and their interconnections. However, dealing with a tripartite graph is computationally complex and demanding, and there is little tool support for studying tripartite graphs. Therefore, it is not feasible to explore a large scale dataset with a tripartite graph. The dominant practice is to reduce the complexity by transforming a tripartite graph into bipartite graphs, each of which consists of two

---

personal page, where all the bookmarks they have added are displayed in reverse-chronological order, along with the list of tags they have used. Compiled from individual user accounts, each and every tag in the system has a tag page where all the bookmarks tagged with that term by any user are listed. Similarly, for each unique item identified by a URL, there is a page listing all the bookmark entries made on the item.

Delicious.com was founded by Joshua Schachter in 2003 and acquired by Yahoo! in 2005. By the end of 2008, delicious.com claimed about 5.3 million users and more than 180 million unique URLs bookmarked.

distinct kinds of entities and connections between them. A bipartite graph is called an affiliation network in social network analysis. Given the necessity of choosing one entity – either information objects (URLs) or tags – to represent (in addition to people/users), the question is, which would represent the relationships between entities more reliably? Although tags have their own merits, relying on tags can introduce non-negligible noise due to a number of interrelated reasons. First, tags in social bookmarking systems are not controlled. The problems due to the uncontrolled nature of tagging systems, including polysemy and synonymy, have been pointed out and, indeed, reported to be abundant in social bookmarking data. This makes it challenging to process tags. Second, categorization research in cognitive science has documented strong empirical evidence that the categorization process is highly context dependent and subjective. Similarly, in the area of personal information management, it has been shown that, either in a physical environment or a digital environment, people's organization behavior is significantly influenced by various contextual factors. This means that tags can vary depending on specific tasks or situations and, thus, without knowing the context, there can be many cases where it is difficult to decide whether two instances of the same tag (the same string) used by different users (or even by the same user at different points in time) represent the same or a similar interest. Third, empirical studies collecting data from a social bookmarking site commonly report that a large portion of items are saved without any tags. This finding suggests that people reveal 'piling' behavior in this environment too. Considering these factors, it was decided that URLs bookmarked provide a better indicator of interests that will connect users.

The decision to exclude tags was a practical choice and is not, by any means, meant to refute the value of tags. In fact, one of the motivations for studying the shared interest structure within the network comes from the recognition that, given the highly subjective and variable nature of categorization, documented in both cognitive science and information science, it would be beneficial if we could identify homogeneous communities of interest first and study tagging behaviors within those communities. One of the potential contributions of this study would be laying the groundwork for a comparative study of tag usage within and across communities of interest within the broad information space of social bookmarking.

The remainder of this chapter describes the data collection and sampling methods, and measures and/or tools that were used for the analysis in each phase of the study. Note that each phase was designed to build upon the previous phase, with increasingly specific goals. The first phase evaluated the extent to which bookmark postings accumulate and overlap in relation to both entities of interest – users and information objects. Three separate datasets, each of which captures different portions of the information space, were used to properly represent the entire information space of *delicious.com*. In the second phase, the investigation was focused on a specific part of the information space, by creating and exploring a network of the *most active* users. By deriving relations (links) among users based on their intersecting personal information space (i.e., common bookmarks), the network represented users of shared interests. The overall structure of the network was examined with various network analytic measures. Finally, the third phase further narrowed down to a specific structure of the network, i.e. a community structure: a structure consisting of densely

connected sub-regions, possibly representing coherent areas of interest.

3.2 First phase

In this phase, the overall level of accumulation and overlap in *delicious.com* was assessed, analyzing sets of bookmark posting data. Although there are other activities that users of this site perform, including browsing other people's collections, the main activity is posting bookmarks. A user can post a bookmark to include it in their own collection of bookmarks, and optionally assign keywords, called tags, of their choice. The URL of the resource bookmarked, the user who posted the bookmark, and tags assigned to the resource are the three main entities represented in a bookmark posting. Among those three entities, the current study focused on URLs and users. In this phase, the basic statistics describing bookmarking activities were presented to characterize the information space in general. The level of accumulation and overlap with respect to URLs and users, respectively, were measured.

3.2.1 Data collection

Collecting data from *delicious.com* can be done using two different methods. One method is to use the Really Simple Syndications (RSS, a simple Web feed) feature offered by the site. *Delicious.com* provides a number of RSS feeds including *Recent* RSS (a feed of the bookmark postings made recently) and *Hotlist* RSS (a feed of the URLs that are most popular at a particular point in time). Among other methods, the *Recent* RSS feed was used for collecting data for this study[10]. Another method of

[10] The *Recent* RSS feed seems to be almost constantly updated as users post bookmarks. By fetching the feed regularly for a period of time, it is possible to

data collection is to crawl the pages on the site. On *delicious.com*, there are three kinds of pages corresponding to three distinct entities involved in bookmarking activities: information object (URL), user, and tag. Each user has their own page(s) including the entire list of their bookmarks, and there is a page for each information object (URL) including all the bookmark postings of the URL (made by different users). Each tag also has a page containing all the bookmark postings associated with it. Since all these pages are open to the public, one can crawl the pages as needed[11]. As will be described below, this method was used to get the entire history of sample users and URLs. Figure 3.1 and Figure 3.2 show an example of a user page and an example of a URL page, on *delicious.com* respectively.

---

collect a sample of bookmarking activities that occurred during the period. Although the collected data do not include all posting activities (because of the time interval of fetching), it can be assumed that no systematic bias would be involved in the data collection process.

[11] While *delicious.com* provides Application programming interfaces (APIs), they could not serve the purpose of this study because the APIs require authentication and allow access only to one's own account. Therefore, an alternative method, crawling and page scraping, was used as the primary data collection method. Note that web scraping necessarily relies on the consistency of the page structure, over which the researcher has no control. Major changes in the site design, for instance, may make scripts written for the previous version obsolete. In fact, at the end of July 2008, *delicious.com* changed the entire 'look and feel' of the site, and the underlying html document structure for each page was also changed. If the data collection for this study had not been completed by then, all the scripts would have had to be rewritten for the completion of data collection. Another limitation of the crawling approach is that many servers restrict the amount of data that can be crawled from a single IP address in a given time period. In fact, this was one of the factors that lengthened the data collection period in this study.

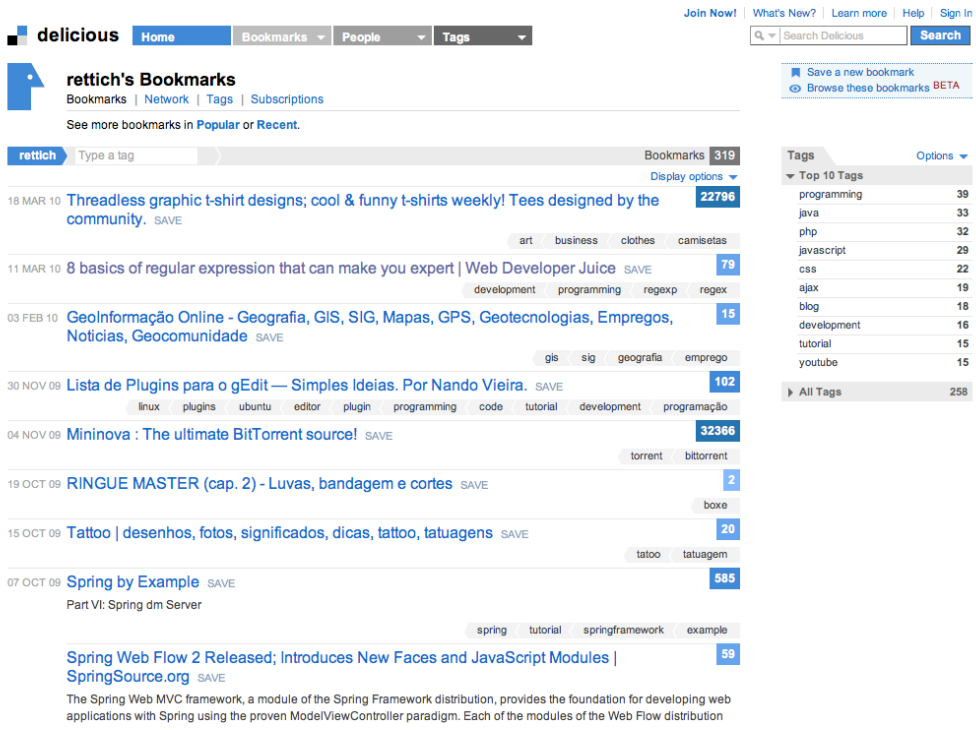


Figure 3.1 A user page on *delicious.com*

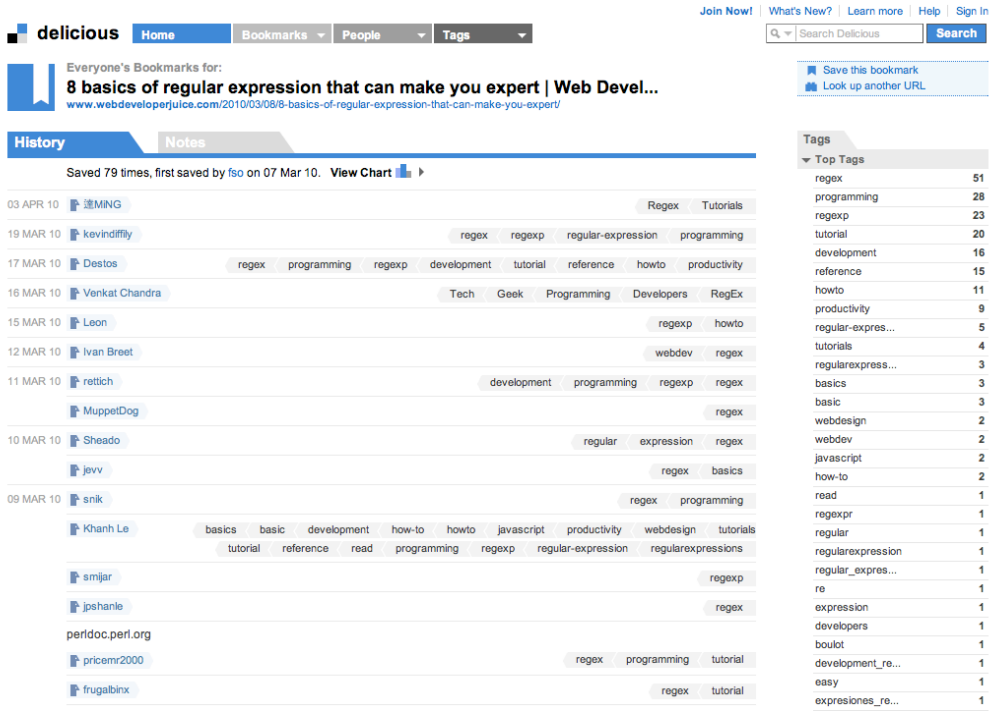


Figure 3.2 A URL page on *delicious.com*

Given the huge scale of the information space being studied, it was important to find a way to capture both the breadth and the depth of the space. To this end, two complementary methods of data collection were used. For capturing the breadth of the space, a large sample of recent bookmarking activities was collected into a dataset called *Recent*. The *Recent* dataset includes a sample of each of two main entities involved in bookmarking activities, users and information objects. The range of each entity in this dataset provides a sense of the breadth of the information space. For representing the depth of the space, two separate samples were drawn, based on the *Recent* dataset: one is a sample of users from the user population and the other is a sample from the population of information objects. For each sample set, the entire history of bookmarking activities associated with each sample element was collected. The resulting datasets are called *User History* dataset and *URL History* dataset, respectively. Figure 1 illustrates the range and coverage of each dataset.

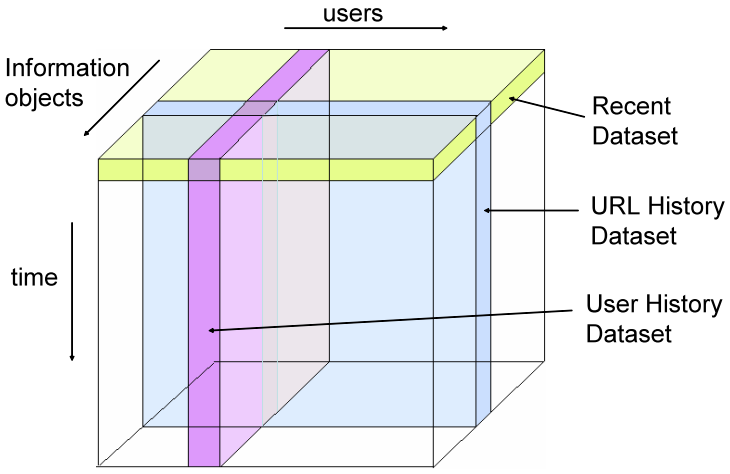


Figure 3.3 Datasets

The *Recent* dataset was collected from January 14, 2008 to April, 21, 2008 (for 14 weeks), using the Really Simple Syndications (RSS) feature provided by *delicious.com*. Through the subscript of the *Recent* RSS feed, a sample of the most recent bookmarking activities were collected. In total, 1,226,472 postings were collected with 999,835 distinct URLs saved by 288,727 distinct users. As described above, this dataset represents the current breadth of the activities on the *delicious.com* site.

In order to get data that accumulated over time, two additional datasets were collected: the *URL History* dataset and the *User History* dataset. The *URL History* dataset includes the entire set of postings associated with 10,000 sample URLs, and the *User History* dataset contains the entire set of postings ever made by each of 10,000 sample users. Sample URLs and users were randomly selected from the *Recent* dataset. The final *URL History* dataset has 1,733,178 postings (of 10,000 sample URLs) made by 484,034 users, and the final *User History* dataset has 3,521,843 postings of 2,451,711 distinct URLs (made by the 10,000 sample users). Table 1 summarizes the size of each dataset.

Table 3.1 The size of each dataset

Dataset	No. of postings	No. of users	No. of URLs
Recent	1,226,472	288,727	999,835
URL History	1,733,178	484,034	10,000
User History	3,521,843	10,000	2,451,711

Having these two *history* datasets, in addition to the *Recent* dataset, allows us to look at the question of accumulation and overlap from two views: a resource-centric view and a user-centric view. From the resource-centric view, we examined, for instance, what proportion of resources (represented by URLs) is shared by multiple users. From the user-centric view, on the other hand, we looked at how many users share one or more resources with other users.

3.2.2 Measures of accumulation and overlap

3.2.2.1 Resource-centric view

From the resource-centric view, the question of accumulation and overlap was primarily addressed by examining the distribution of popularity of the information objects. The popularity of an information object is defined as the number of users who have bookmarked the object over a given period of time. It is the size of the group of users who share interests in that object. The popularity distribution of information objects, therefore, depicts the overall spread/concentration of interests and the consequent groupings of users.

Then, two interrelated measures were used to characterize the entire dataset in terms of diversity and commonality of interests: the ratio of distinct URLs in the dataset to the total number of bookmarks, and the proportion of URLs that are shared by multiple users.

With regard to the effect of accumulation over time, another question arose: when and to what extent do common interests for certain information objects start to

accumulate in this information space? The rate that the proportion of distinct URLs decreases as the dataset grows can suggest an answer to that question. The proportion of URLs shared by multiple users, among the distinct URLs, provides a similar but slightly different measure of overlap. In order to see the changing level of accumulation and overlap, in the *Recent* dataset, cumulative statistics were calculated at a certain time interval among the recent postings collected (i.e., per 10,000 new postings). The same statistics were measured on the *URL History* dataset to show the current level of accumulation and overlap in *delicious.com*.

3.2.2.2 User-centric view

With the *Recent* and *User History* datasets, the volume of user activity was measured with the frequency of bookmark postings each user made. Similarly with the resource-centric view, the accumulated number of distinct users and the proportion of users who made multiple postings in each dataset were calculated. In the case of the *Recent* dataset, again, the accumulated number of users and their postings was calculated at an interval of every 10,000 new postings added to the dataset, to visualize the growth of the dataset.

In order to look at the level of overlap across users, as well as at their accumulation of activity, the number of bookmarks shared with other users, as well as the total number of bookmarks a user has, was calculated and compared. For instance, if a user had posted 100 bookmarks, each of those 100 bookmarks was checked to see whether other users had also posted it.

3.3 Second phase

In this phase of the study, we were interested in seeing whether and how users of a social bookmarking site, *delicious.com*, could be connected to form a network based on their shared interests. A sample drawn from the dataset collected in the first phase was used for building a dataset for this phase. The data were first represented as an affiliation network. An affiliation network is a two-mode network, meaning that it is comprised of two distinct types of entities. For the purpose of analysis, a two-mode network is often projected into a one-mode network, to which a full range of network analytic concepts and methods can be applied. In this study, from a two-mode network of information objects and users, a network of users was induced.

This study is an exploratory study. Instead of having a structural hypothesis or testing the fit of a specific network model, the goal of this phase was to explore the network for any meaningful patterns. The network analysis in the second phase was conducted in two parts. In the first part, with the network of users transformed from the affiliation network, a number of common global properties, useful in understanding and characterizing the structure of a network, were used to examine the structure of the network. In the second part, a technique called *m*-slice was used to further investigate the internal structure of the network.

3.3.1 Sampling strategy

Considering the scale and sparseness of the data, it would not be reasonable to expect a single study to collect data that can accurately reflect the entire space for a detailed analysis. Therefore, it was important to make the sampling criteria as clear

and systematic as possible, so that it is apparent exactly which part of the corpus was included in this study.

The basic principle of sampling was to choose the most *active* users in the *current* user population. The *Recent* dataset collected in the first phase reflected the current user population (at the point of data collection). It is most likely that any *current* user posted at least one bookmark during the period of 14 weeks when the *Recent* dataset was collected. This criterion allowed us to filter out inactive users who ceased to participate in the network. Two additional criteria were employed to define how *active* a user was in their recent activities: the number of different days they posted at least one bookmark and the number of recent postings. Tables 2 and 3 show the frequency distribution of users by these two criteria.

Table 3.2 The distribution of users in the *Recent* dataset by the number of postings

No. of postings	No. of users	Percentage	Cumulative %
201 or more	70	0.02%	0.02%
151-200	65	0.02%	0.05%
101-150	199	0.07%	0.12%
51-100	1107	0.38%	0.50%
41-50	803	0.28%	0.78%
31-40	1661	0.58%	1.35%
21-30	4142	1.43%	2.79%
11-20	14879	5.15%	7.94%
10	3533	1.22%	9.16%
9	4423	1.53%	10.70%
8	5525	1.91%	12.61%
7	7367	2.55%	15.16%
6	9879	3.42%	18.58%
5	13363	4.63%	23.21%
4	19444	6.73%	29.95%
3	29695	10.28%	40.23%
2	51570	17.86%	58.09%
1	121002	41.91%	100.00%
Total	288727	100.00%	

Table 3.3 The distribution of users by the number of days of posting one or more bookmarks

No. of days	No. of users	Percentage	Cumulative %
51 or more	25	0.01%	0.01%
41-50	62	0.02%	0.03%
31-40	207	0.07%	0.10%
21-30	1030	0.36%	0.46%
11-20	7417	2.57%	3.03%
10	2240	0.78%	3.80%
9	3056	1.06%	4.86%
8	4004	1.39%	6.25%
7	5498	1.90%	8.15%
6	8180	2.83%	10.99%
5	11757	4.07%	15.06%
4	18062	6.26%	21.31%
3	29358	10.17%	31.48%
2	54237	18.78%	50.27%
1	143594	49.73%	100.00%
Total	288727	100.00%	

By combining the top 10% on each criterion, the set of users to be included to construct a network was obtained. The set included 23,287 users (8.1%) with 6 or more active days AND 9 or more postings.

Having selected a set of currently active users, information objects in which they demonstrated interest by posting bookmarks were added to the dataset. The complete list of bookmarks ever posted by the 23,287 users contained 25,559,506 bookmarks posted on 13,633,750 distinct information objects. These information objects were used as the pool for the second mode of the affiliation network (with the above 23,287 users as the first mode). Due to technical and computational limitations, instead of using the entire list of information objects, three subsets of information objects were drawn applying three time windows (12 months, 6 months, and 3 months)[12]. Using each subset in turn, three affiliation networks were created. Each affiliation network was then transformed to obtain a one-mode social network of users. For the sake of computational efficiency, only the network built with the 3 month window was further analyzed.

3.3.2 Network properties

While the set of actors defines the boundary of the network to be analyzed, relations create the structure. Relations, in general, are characterized by content, direction, and strength. Depending on how the relation of interest is defined, a network can be undirected or directed (representing the direction of relations), and simple/binary or

[12] While there are network analysis tools that can handle a large scale dataset, and most of those tools, including the ones used in this study (Pajek and igraph package in R), support the transformation of a two-mode network (consisting of users and information objects, in this case) into an one-mode network, a transformation of this dataset could not be done using the existing tools due to the excessive number of entities in the second mode (information objects). Even when the number of information objects was reduced substantially by applying a shorter time-window (12 months, 6 months, and 3 months), this was still the case. Therefore, for each time window, a one-mode social network of users had to be constructed through a series of steps, involving separate database tables and scripts, outside those tools.

valued (denoting the strength of relations). When a social network is drawn from an affiliation network, it is typically a simple undirected network. Actors who participate in one or more common activities are connected. Since the content of the relation is a sort of commonality or group membership, there is no directionality. The strength of relations, however, may vary. The number of activities or groups each pair has in common, for instance, may be counted and used as a measure of the strength of their relation.

The boundary of the network studied in the second phase of this study was defined conceptually as the current active users and, as described above, the actual list of users was created by applying two criteria of recent activity in combination. For the network analysis, the affiliation network consisting of the chosen users and the information objects linked by bookmarking activities was constructed. This affiliation network then was used to induce the network of users to be explored. The content of the relations in this network was shared interests (assumed to be reflected in shared bookmarks), and the number of information objects (bookmarks) a pair had in common was used as the strength of the tie between two users.

Basic measures in network analysis are graph theoretic measures. In interpreting these measures, it is often useful to distinguish local measures that are applied to individual vertices or edges, and global measures that characterize the whole network. It is, however, also important to note that, in the network analytic framework, the global structure arises from the patterns of local connections. Therefore, many global properties are derived from local measures. Generally, the average value or the distribution of values of a local measure is used as a global

property. For instance, whereas *degree* is a property of a node, the degree distribution is a global property characterizing the overall connectivity of the whole network. It is also worth mentioning that, while typical social network studies have focused on the properties of individual vertices or edges, with a goal of, for instance, identifying central actors who take on important positions in the network, recent studies of large networks have shifted the focus, due to the scale, to global properties of networks. There are a number of non-trivial global patterns commonly found in real-world networks, including highly skewed degree distribution, short average pathlength, and high clustering coefficient. Each of these global properties, and the related local measures, was examined for the network of *delicious.com* users in this phase.

In many network studies, finding the typological structure of the given network constitutes an important empirical understanding of the network. An analysis of network components depicts the basic typological structure of a network. In the last part of the second phase, components of the network were located and studied. Components divide the network of interests into partitions, and thus, are often used as a basis for further analysis.

The following describes the network measures and properties used in the second phase of this study. Local measures and related global properties will be discussed together. For most of the analysis, two network analysis tools that are well known for their capacity for processing a large dataset, *Pajek*[13] and the *igraph*[14]

[13] *Pajek* (the Slovenian word for Spider) is a Windows program developed by Vladimir Batagelj and Andrej Mrvar for analysis and visualization of large networks (de Nooy, Mrvar, & Batagelj, 2005). The program and other resources, including

package in *R*, were used in combination.

3.3.2.1 Degree, density, and degree distribution

A (undirected) graph G (*V, E*) consists of a set of vertices (nodes) *V* representing actors and a set of edges (ties) *E* representing relations. The most basic statistics describing a graph G are, of course, the number of vertices (nodes) $n = |V|$ and the number of edges (links) $m = |E|$. Given the number of vertices, the number of edges indicates the overall volume of relations or connections in the system.

The *degree* k_i of a vertex n_i is defined as the number of edges adjacent to the vertex, in other words, the number of neighbors of the node. While the *degree* is a fundamental property of an individual node, the average degree *k* of a graph, the average of degrees over all its nodes, depicts the overall connectivity of the whole network. One of the most widely used measures of network structure, *density* δ is the number of existing links in the graph divided by the total number of possible links ($\delta = 2 \cdot m/n \cdot (n\text{-}1)$). It is a measure of how fully the graph is connected. The *density* of a network, in effect, is the probability that two randomly chosen nodes of the network are connected.

Another important notion related to degrees is the degree distribution. The

---

published papers, presentations, and tutorials, can be downloaded from its Wiki (http://pajek.imfm.si/doku.php). Pajek is one of the main software packages used by social network analysts (Huisman & Duijn, 2005). The key strengths of this program are known to be its ability to handle large datasets and its visualization capabilities.

[14] *igraph* is a free software package available under GNU General Public License (http://igraph.sourceforge.net/). It was developed specifically for the analysis of large networks, and includes implementations for a broad range of network analyses. It can be installed as a C library, as an R package, or as an extension module in Python or Ruby. In this study, *igraph* was used in *R* (http://www.r-project.org/)

degree distribution of a graph is the proportion p_k of vertices in the graph that have degree k, for all k. In other words, it gives the probability that the degree of a randomly chosen vertex equals k. In recent studies of real-world networks, it has been found that the degree distribution of a large empirical network is typically highly right-skewed with a long tail, meaning that the majority of the vertices in the network have low degree while only a small number of vertices have high degree. The distribution is often observed to approximately follow a power law $p_k \sim k^{-\alpha}$, for a constant exponent α, which is called the degree exponent. The highly skewed distribution reveals that the degree distribution is heterogeneous and there is high variability in nodes in terms of their degrees. In this case, the average degree has little value in describing the characteristics of the network structure. Rather, a measure of heterogeneity is more useful. The *degree exponent* α is considered to provide such a measure. Therefore, many recent network studies report the degree distribution, and estimate the degree exponent if the distribution follows a power law or exponential form[15].

In this study, the above basic measures of the overall network cohesion and connectivity were used. In the induced network of *delicious.com* users, links between two users were created if they had one or more items (i.e., bookmarks) in common. Therefore, the degree of a user means, in this context, the number of other users who have one or more shared items with the user. The application and interpretation

[15] The most frequently used approach to estimating the degree exponent α is to rely on graphical methods, such as fitting the slope of the line in the log-log plot of the histogram of the degree distribution (Newman, 2003). Recently, it has been pointed out that graphical methods based on linear fitting introduce biases (Goldstein et al., 2004). As an alternative, a method for extracting the degree exponent using maximum likelihood estimation is suggested (Newman, 2005).

of degree related measures in this context is straightforward. Due to the projection of a two-mode affiliation network, however, one precaution needs to be taken in looking at high-degree nodes. By definition, all the people who belong to or participated in the same group or event are completely connected in the induced one-mode network. It means that, for instance, a person who has one popular item in his/her collection may have more connections than another person who has a number of unpopular items shared with only a few other users.

3.3.2.2 Distance, diameter, and average pathlength

The *distance* between two vertices in a network is the length of the *shortest path* between the vertices, that is, the minimum number of edges that need to be followed to go from one vertex to the other. The average or characteristic pathlengh of a network is the average distance over all pairs of vertices, while the diameter of a network is defined as the largest distance between any pair of vertices in the network. These concepts and measures based on *distance*, of course, can only be applied if there is a path between two vertices. The diameter, therefore, is often measured on the largest component of the network of interest. In addition, since measuring distance of all pairs of vertices is computationally expensive, the average pathlength is often calculated over a random sample of vertices (Watts, 1999).

An important global feature of a network related to the average pathlength is the small world effect. The term ‘small world’ refers to the empirical finding that nodes in a network can be connected through a small number of intermediaries. In fact, for most real networks, regardless of the nature of the system they represent, it

is found that the average pathlength is surprisingly small, even when the network has a huge number of vertices with sparse connections among them. It is further found that the average pathlength often scales with the natural logarithm of the number of vertices in the network.

3.3.2.3 Clustering coefficient (transitivity)

While the degree is a property of a vertex and the distance is a property pertaining to a pair of vertices (a dyad), the clustering coefficient is based on the concept of triadic closure or transitivity. A triad (or a triple) of vertices A, B, and C is called transitive when the three vertices have 'balanced' relations. In an undirected graph, it means that if A connects to B and B connects to C, then A connects to C. In social network analysis, triadic relations have been emphasized as a meaningful building block of a group structure. The number of ties (edges) among three actors (vertices) defines different patterns of triadic relations: no tie means isolates, one tie means a couple and an isolate, two ties means one actor bridging the others, and all three ties means a cluster. By counting the instances of the different triadic patterns, one can get a picture of whether and how actors in the network are scattered or clustered. The clustering coefficient is a measure quantifying this feature.

There are two different ways to calculate the clustering coefficient of a network. One way is to directly calculate the transitivity ratio, by defining a clustering coefficient C as:

$$C = \frac{3\times \text{ number of triangles in the network}}{\text{number of connected triples of vertices}}$$

where a triangle is a set of three vertices each of which is connected to both others

and a connected triple of vertices refers to a set of three vertices at least one of which is connected to the other two. In other words, a triangle is a triad with three edges and a connected triple is a triad with at least two edges. Therefore, C measures the portion of triples that have the third edge to make them transitive. In effect, C is the probability that two neighbors of a randomly selected vertex are connected.

An alternative way of obtaining the clustering coefficient C of a graph is to calculate the clustering coefficient of each vertex i in the graph and get the average over all vertices. A definition of the clustering coefficient of a vertex i, proposed by Watts and Strogatz (1998), is:

$$C_i = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered on vertex } i}$$

Note that the denominator of the above equation, the number of the triples centered on a vertex, is in effect the number of pairs among the neighbors of the vertex. If there is an edge connecting a pair of its neighbors, that makes a triangle connected to the vertex. Therefore, the clustering coefficient of a vertex i can be equivalently defined as:

$$C_i = \frac{2 \cdot E_i}{k_i \cdot (k_i - 1)}$$

where E_i is the number of edges connecting the neighbors of vertex i, and k_i is the degree of vertex i. In other words, the clustering coefficient of a vertex is the ratio of the number of existing edges connecting its neighbors to the maximum possible number of edges between the neighbors. Note that the above definition of C_i is, in effect, the density of the neighborhood of vertex i. Therefore, it provides a local

measure of cohesion among neighboring vertices.

The clustering coefficient of the graph can then be obtained by averaging the local clustering coefficient of all vertices (of degree at least two) in the network.

$$C = \frac{1}{n}\sum_{i} C_i$$

The two definitions of clustering coefficient of a graph are both widely used. As Newman (2003) points out, these definitions reverse the order of operations. The later approach has the advantage of having a local value for each vertex, and, hence, of being able to observe the distribution of the local clustering coefficient. However, it should be noted that, since the clustering coefficient of a vertex is in effect a local density measure, it is affected by the size of its neighborhood (i.e, low degree vertices tend to have higher density due to the small denominator). As a consequence, the global measure of the clustering coefficient of a graph, averaging local values, weight low-degree vertices more.

Regardless of which definition is used, however, many networks are found to exhibit a high clustering coefficient. That is, if vertex A is connected to vertex B and vertex B to vertex C, the probability that vertex A and C are also connected is much higher than the probability that two randomly selected vertices are connected. This finding indicates that vertices in a real network tend to cluster to form a cohesive structure.

In the induced network of *delicious.com* users, transitivity means that if two users are neighbors of the same third user, then the two users are likely to be connected to each other. In other words, if A and B, and B and C have one or more item in common, then A and C have a high probability of having a shared

item. A highly clustered group of users in this network indicates cohesive interests with a dense intersection of their bookmark collections. It should be noted, however, that a network induced from an affiliation network inherently has a higher clustering coefficient than a typical one-mode network, due to the fact that all members in a group are to be completely connected in the process of projecting the two-mode graph into a one-mode graph. If users A and B and C bookmarked the same item, the transitive relations among A, B and C are given by definition.

The high clustering coefficient of a network, together with the short average pathlength discussed above, contributes to the small-world effect. A theoretical model for small-world networks, suggested by Watts and Strogatz (1998), characterizes a small world network as a loosely connected set of highly clustered subgroups.

3.3.2.4 Components

Actors in a social network are embedded in a number of relational structures, ranging from dyads, triads, subgroups, to the network. The measures and properties discussed above are, in a sense, concerned with how the actors are embedded in and form a part of the local structures, including dyads, triads, and their neighborhood. By aggregating these local features, one can characterize the overall global structure of the network of interest.

Another way of studying the structure of the network is to shift the focus to larger constructs of the network. One of the most common problems in social network analysis is to identify substructures or subgroup structures that may be present in a network. Identifying subgroups within a network and their patterns of

grouping provides a high-level picture of how the network is constructed. There are numerous ways to define and characterize subgroups or substructures in a network, based on various structural patterns. The analysis of network components is a commonly used way to find topological substructures of a network.

A *component* is a partition of a network where each pair of nodes is connected by a path[16]. It is a maximal set of connected nodes, meaning that all nodes in the set are connected and no other nodes can be added to the set without compromising the definition of the set. Each node in a network belongs to one and only one component (an isolated node is itself a component). When there are multiple components in a network, they represent the regions of the network that are completely disconnected from one another.

A small social network may be comprised of a single component, but it is common for a network, especially for a large network, to have a number of components with varying sizes. The number of components in a network together with the distribution of their sizes is an important property representing the macro structure of the network. A network consisting of two components of comparable sizes, for instance, has a clearly different substructure from a network of one large component and several isolated nodes. In fact, it has been observed that real-world networks typically consist of a large component, often called a giant component, and

[16] For directed graphs, two different kinds of components can be defined. A weak component is a set of nodes each pair of which is connected regardless of the direction of links. A strong component, on the other hand, considers direction, and each pair in the strong component should have a directed path. For undirected graphs, by definition, the distinction of weak and strong components does not apply. A component in an undirected graph is equivalent to a weak component in a directed graph.

a number of small components, each containing only a small percentage of the nodes.

Investigating the composition and distribution of components was relevant and useful to understanding the interest space of *delicious.com* users. Component analysis addressed a set of typological questions. How do components divide the network? How large is each component? Are there a few large components, or a number of smaller components? Regardless of the specific characteristics of each component, the existence and relative size of the components within the network can characterize the overall configuration of the interest space.

3.3.3 Network decomposition: m-slice analysis

While the various network properties and measures discussed in the previous section characterize the overall typological structure of the network, different substructures may exist within the higher order structure. In general, substructures emerge when there are subsets of nodes and edges that share certain characteristics in a way that distinguishes them from the rest of the network. Therefore, in order to identify such subsets or substructures, a number of techniques for decomposing a network into smaller parts have been proposed based on different defining characteristics.

In this study, a technique called *m*-core (Scott, 2000) or *m*-slice (Nooy et al., 2005)[17] was used for network decomposition. An m-slice is defined as "a maximal

[17] Scott (2000) introduced this technique as a variation or an extension of *k*-core analysis. In fact, as Scott (2000) pointed out, *k*-core analysis and *m*-core analysis are based on essentially the same procedure, while each adopts a different definition or criterion of cohesion to draw substructures from a given network. While Scott (2000) named this technique *m*-core and emphasized its basic similarity to the

sub-graph in which each line has a multiplicity greater than or equal to *m*" and shows "a chain of points connected by lines of the specified multiplicity" (Scott 2000, p. 112). When the two-mode affiliation network, consisting of information objects and users, is transformed into a one-mode social network of users, each information object shared by a pair of users produces a link between them. Therefore, if a pair of users shares multiple information objects, at some point during the transformation process, there are multiple lines (links) between them. This phenomenon is called line multiplicity. At the end of the transformation process, any multiple lines between two users are merged into a single edge, to which the number of lines (the line multiplicity) is then assigned as a weight. Since the weight on an edge reflects the number of shared information objects between the connected pair, it is in fact a useful indicator of the level of shared interests between the two users. The weight, however, was ignored in the first part of the analysis where network properties were measured, because all of them are defined assuming an unweighted network. The *m*-slice technique was chosen for the substructure analysis, among a number of alternatives, because it allows us to take different levels of shared interests (represented by the number of shared information objects) into account.

By definition, an *m*-slice consists of edges that have a value of *m* or higher and nodes that are incident on those edges. In order to obtain an *m*-slice from a network at a given *m* value, therefore, all the edges (connections) with a value less than *m* and any isolated nodes are removed. The basic procedure for *m*-slice analysis involves iterative removal of edges and nodes. Starting from the original network,

---

k-core approach, Nooy et al. (2005) chose to call it *m*-slice instead of *m*-core in order to avoid confusion. We use the term *m*-slice for the same reason.

edges and nodes are progressively removed as the value of *m* increases, and the original network is iteratively broken down into smaller sub-networks. It is, in effect, filtering out the weakest ties at each step so that areas with stronger connections are brought forth. A result of the procedure is a nested structure where lower *m*-slices contain higher *m*-slices. It discloses the overall shape of the network created by different levels of tie strengths (different degrees of interest sharing in our network) like contours on a map.

In the network of *delicious.com* users, each *m*-slice represents the sub-network where each connected pair of users share *m* or more information objects. The nested structure of *m*-slices, therefore, depicts the internal structure of the network with varying degrees of shared interests. In order to see the effect of increasingly stricter conditions for making connections among users on the resulting structure of the network, the same set of network properties was measured in each *m*-slice as in the original network. In addition, the number and sizes of its components were examined to show how the overall composition of the network changes as *m* changes.

The *m*-slice analysis was also used to identify cohesive subgroups or communities within the network. In principle, detecting a cohesive region within the network relies on the assumption that naturally existing subgroups would be reflected in some non-random pattern of connections. The most basic and obvious subgroups that appear in the connective structure of a network are network components, since by definition each node in the network belongs to only one component and components are separated from one another. As mentioned above, network components, therefore, can be regarded as communities. In a large scale network,

however, it is often the case that a large component contains multiple subgroups, especially if there is a giant component covering the majority of the network. The *m*-slice analysis provides a way to detect subgroups or communities by repeating the removal of weak ties until a large component in a lower slice is broken down into separate components. Note that components in any given *m*-slice can be regarded as communities defined by the minimum strength of connections with the threshold value of *m*. The method in fact falls into the category of hierarchical clustering techniques[18] which have been widely used in the tradition of social network analysis in an attempt to locate cohesive subgroups within a social network.

3.4. The third phase

The third phase was a preliminary exploration of communities found in a specific *m*-slice. From a methodological point of view, this phase explored ways to investigate the content and coherence of shared interests within communities that are identified by structural features. While subgroups or communities within a network can be located based on certain structural features, such as connected components within an

[18] Hierarchical clustering divides actors in a network into groups, based on their pairwise similarity or strength of relations. Unlike general clustering applications where the definition of similarity is typically based on some attribute(s) of the elements, in the context of network analysis, elements (vertices) are grouped based on structural criteria (e.g., the number of paths between vertices) rather than their inherent attributes. Hierarchical clustering can be done in an agglomerative way (grouping similar vertices into increasingly larger units) or in a divisive way (iteratively breaking down the network into smaller subsets). In any case, the first step is to measure, for every connected pair of vertices in the network, how close or strong the connection is. There are a number of ways to assign the weight to each connection, based on various structural characteristics. In the case of a one-mode projection of an affiliation network, one can use the number of joint memberships as a measure of the strength of a relation.

m-slice as in this case, further investigation is needed to determine whether those components of the network indeed constitute coherent communities with shared interests.

Within the network of *delicious.com* users, where users are linked by their common possession of bookmarks, a component consisting of users who are connected to one another but separated from the rest of the network may represent an area of interest or a characteristic pattern of bookmarking behavior. Although there is no specific information regarding individual nodes (users) available to understand the traits of a component they comprise, the links connecting the nodes in a component may provide such information. Given the fact that the network of users is induced from the affiliation network of users and information objects, it is possible to trace each link back to a specific shared information object(s). In other words, exploiting the duality of affiliation relations, a component can be characterized by examining what kinds of information objects constitute links among nodes within the component.

Every link in a community was traced back to the information objects involved in creating the link[19], and the union set of information objects for the community was constructed. For each information object, its contribution to the connectivity and coherence of the community was measured with an index, called the

[19] When a one-mode network is created by transforming an affiliation network, the information on the other mode is lost. To the author's best knowledge, there is no support in either of the network analysis applications used in this study (Pajek and igraph package in R) for tracing a link in the resulting one-mode network back to the affiliation network. In order to find out which information objects constituted each link in the network of users, therefore, a separate database was created to keep track of entities in both modes (users and information objects) and their relationships.

contribution index. In addition, the extent to which the users who bookmarked the given information objects were brought together into the community in question was measured by another index, called the aggregation index. (The method for calculating each of these indices is described with the results from the third phase.) The set of information objects for a community was then sorted by the contribution index, and the content and cohesiveness of shared interests defining the community was examined by looking at the URLs and titles of the information objects, especially the highly ranked ones, and by averaging over the aggregation index within the set.

Chapter 4. Results of Phase I

This study attempts to understand the structure of the information space of a social bookmarking site, *delicious.com*, in terms of the intersecting interest spaces of its users. In this study, the term information space is used to denote the entirety of information available in a social bookmarking system, along with the dimensions of information objects, users, and tags. One of the main characteristics of social bookmarking, to which this study pays particular attention, is that its information space is the aggregate of personal information spaces of individual users. As individual activities of bookmark posting accrue on each user's account, not only each user's personal information space, but also the information space of the site as a whole, is shaped and expanded. The first phase of this study, therefore, starts by examining bookmark postings, with an emphasis on how bookmarking activities accumulate at the individual level and the community level. This entails analyzing basic statistics of bookmarking activities including the number of bookmark postings per user and per information object, and measuring the extent to which bookmark postings of individuals intersect and overlap.

The analysis was performed on three different datasets, collected from *delicious.com*, which allows us to look at the question of accumulation and overlap from both a resource-centric view and a user-centric view. The first set, the *Recent* dataset, was constructed by fetching a RSS feed from *delicious.com* regularly during

a fourteen week period (from January 14, 2008 to April 21, 2008). The *recent* dataset contains 1,226,472 bookmark postings made by 288,727 users on 999,835 information objects (URLs). The second and the third datasets are called the *History* datasets and contain the entire history of bookmark postings associated with a sample of 10,000 users and a sample of 10,000 URLs respectively. More specifically, the second dataset, *User History*, consists of 10,000 users randomly selected from the *Recent* dataset and all the bookmark postings made by those users. In other words, it contains the entire bookmark collection of each of 10,000 users. A total of 3,521,843 postings were included in the second dataset. Likewise, the third dataset, *URL History*, consists of 10,000 URLs randomly selected from the *Recent* dataset and all the bookmark postings made to those URLs. A total of 1,733,178 postings comprise the third dataset.

In the following, basic statistics of bookmarking activities shown in the *Recent* and the two *History* datasets will be presented and the level of accumulation and overlap with respect to URLs and users, respectively, will be measured.

4.1 Resource-centric view

From a resource centric view, a set of bookmark postings can be seen as a set of information objects, each of which has a subset of users who bookmarked the given information object. In this section, a number of basic statistics of bookmarking activities in our datasets, focusing on the axis of information object, will be presented.

4.1.1 The Recent Dataset

1) Popularity of information objects

As reported above, the recent dataset contains a total of 1,226,472 bookmark postings. The number of distinct information objects in this set is 999,835. For each information object in the *Recent* dataset, the number of bookmark postings made on the information object and collected into the *Recent* dataset during the fourteen-week period of data collection[20] was calculated as a measure of popularity. The popularity distribution is important because it shows how bookmarking activities, and by extension interested users[21], scatter and gather around particular information objects.

Figure 4.1 shows the frequency distribution of the 999,835 URLs in the *Recent* dataset by the number of bookmark postings. Each point in the figure represents the number of URLs that have the given number of postings. For instance, the left-most point on the top shows the number of URLs with 1 posting. The figure is plotted on logarithmic scales on both axes. The most notable fact about the frequency distribution of bookmark postings in the *Recent* dataset is the overwhelmingly large proportion of URLs with only one posting. Of the 999,835 URLs in the recent dataset, 897,145 URLs (89.73%) occurred only once.[22] The

[20] It should be noted that, since the data collection involves time intervals between RSS fetches, the number may not be the same as the total number of bookmarks added to the information objects on the *delicious.com* in that period.

[21] Although there are a small number of cases where the same URL is posted multiple times by the same user, the number of bookmarking postings on an URL is approximately the same as the number of different users who bookmarked the URL, in the dataset at hand.

[22] The fact that a URL appeared once in the *Recent* dataset does not necessarily mean that the URL was posted only once during the fourteen week period (from January 14, 2008 to April, 21, 2008), because of the time interval in data collection.

average number of bookmark postings per URL in the *Recent* dataset is 1.23 with the 95th percentile value of 2. It means that, except for a very small portion of URLs, the vast majority of URLs were of interest to only a few users, at least in the given time window. In other words, most of the 288,727 users in the *Recent* dataset showed little overlap in their recent bookmarking choices.

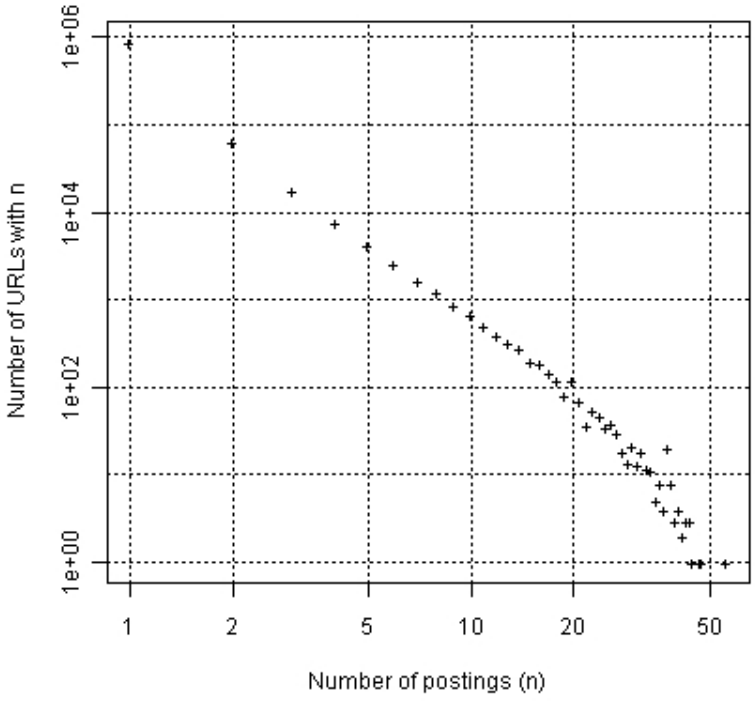


Figure 4.1 The number of URLs by the number of bookmark postings in the *Recent* dataset

2) Accumulation over time

While the frequency distribution in Figure 4.1 presents a summary description of the overlap in recent bookmarking activities, we are also interested in examining the accumulation of bookmark postings on information objects over time.

Since the *Recent* dataset was built up incrementally over a fourteen-week period by collecting new postings on delicious, we can observe the level and the pattern of accumulation as the number of postings increased in the dataset. Calculated at an interval of every 10,000 new postings added to the dataset, the accumulation is measured by 1) the ratio of distinct URLs in the dataset to the total number of bookmarks and 2) the proportion of URLs whose frequency (*n*) in the dataset is greater than 1 (URLs that have been posted by more than one user). Table 4.1 shows these two statistics for the *Recent* dataset.

Table 4.1 Distinct URLs and the proportion of overlaps

Postings	Distinct URLs	URLs with n > 1	
		Number	Percent
10,000	9,982	18	0.18%
20,000	19,839	160	0.81%
30,000	29,512	470	1.59%
40,000	39,108	829	2.12%
50,000	48,660	1,233	2.53%
100,000	95,003	4,051	4.26%
150,000	139,937	7,515	5.37%
200,000	183,522	11,415	6.22%
250,000	226,005	15,632	6.92%
300,000	268,589	19,649	7.32%
400,000	351,556	28,392	8.08%
500,000	432,076	37,565	8.69%
600,000	510,665	46,753	9.16%
700,000	589,635	56,143	9.52%
800,000	666,338	65,968	9.90%
900,000	742,297	75,974	10.23%
1,000,000	820,638	84,415	10.29%
1,100,000	899,833	92,503	10.28%
1,200,000	978,990	100,632	10.28%

Table 4.1 shows that the number of distinct URLs increases rapidly while little accumulation per URL occurs. The URLs posted multiple times indicates that there were users who shared interests in those information objects. Up until 10,000 postings there were only 18 URLs that were posted more than once; the rest of the 9,982 URLs were posted only once. However, the proportion of URLs that were posted multiple times slowly yet steadily increased as the *Recent* dataset grew. In addition, as will be described below, examination of the *URL History* dataset, which contains the entire history of a random sample of 10,000 URLs, revealed that the majority of those URLs were shared by multiple users.

The large portion of distinct URLs in the *Recent* dataset indicates the diversity of user interests and the broad range of resources being bookmarked in this site. On the other hand, the huge increase in the proportion of repeatedly posted URLs in the *URL History* dataset demonstrates the effect of accumulation over time.

4.1.2 The *URL History* Dataset

The URL history dataset was constructed by randomly sampling 10,000 URLs from the *Recent* dataset and crawling all the pages associated with each of those 10,000 URLs. The crawled pages were parsed to extract information about bookmark postings, including the user and the time of the posting. As the result, all the postings on each URL from the point when it was first posted to *delicious.com* to the time of crawling were collected. The total number of bookmark postings collected into this dataset is 1,733,178. The average number of bookmark postings per URL, therefore, is 173.3.

1) Popularity distribution

As in the analysis of *Recent* activity, it is of interest to see how bookmarking activities are distributed over different information objects. Figure 4.2 presents the frequency distribution of 10,000 sample URLs in the *URL History* dataset by the number of bookmarking postings. Each point in the figure represents the number of URLs that have the given number of postings. The graph is plotted on a logarithmic scale on both axes.

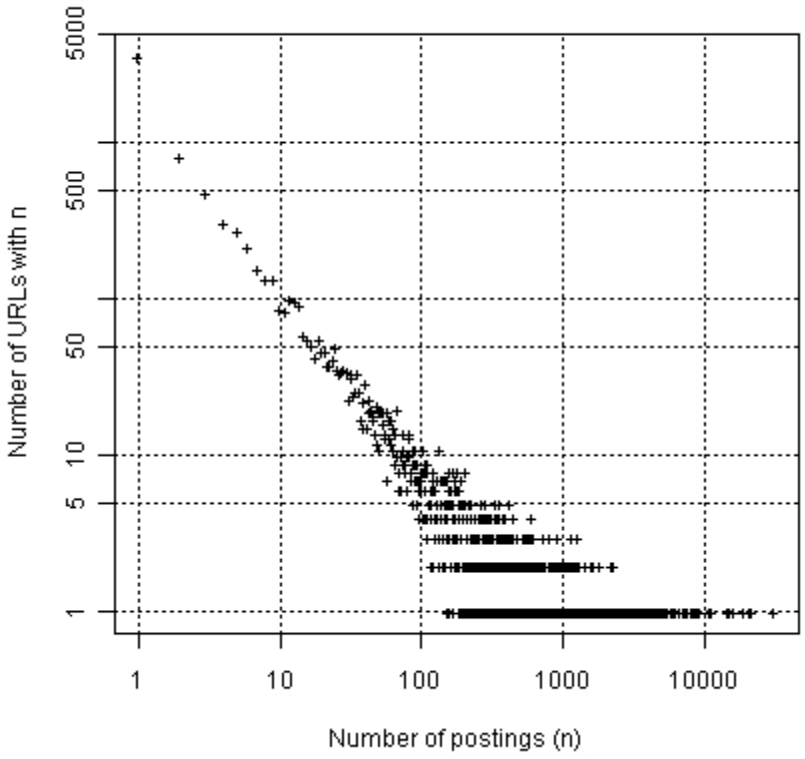


Figure 4.2 The number of URLs by the number of postings in the *URL History* dataset

In Figure 4.2 the points are nearly in a straight line especially at the low end of less popular items, showing a signature of a power-law like distribution. The vast

majority of the information objects were bookmarked only a few times, while a small number of information objects were extremely popular. As shown in Table 4.2, among the 10,000 URLs in the *URL History* dataset, 3,645 (36.45%) URLs were posted only once, and 6,273 (62.73%) URLs were posted 10 times or less.

Table 4.2 Frequency of URL posting

No. of postings	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
1	3645	36.45	3,645	36.45
2	833	8.33	4,478	44.78
3	479	4.79	4,957	49.57
4	313	3.13	5,270	52.70
5	275	2.75	5,545	55.45
6	216	2.16	5,761	57.61
7	156	1.56	5,917	59.17
8	135	1.35	6,052	60.52
9	134	1.34	6,186	61.86
10	87	0.87	6,273	62.73
11-100	2056	20.56	8,329	83.29
101 - 1000	1265	12.65	9,594	95.94
1001-10000	395	3.95	9,989	99.89
10001 -	11	0.11	10,000	100.00

While 83.29% of URLs have less than 100 postings, the mean number of postings in this set is 173.3. It is clearly attributed to the small number of the extremely popular URLs. There are 406 URLs that occur more than 1,000 times in the collection. Eleven of those have more than 10,000 postings each, with an average of 17,646 postings. The number of postings associated with these 11 URLs is 194,111, accounting for 11% of the total number of postings (1,733,178) in the *URL History* dataset. Table 4.2 lists those 11 URLs in descending order of the number of occurrences.

Table 4.3 Top ranked URLs in the *URL History* dataset

URL	Count	No. of months
http://slashdot.org/	31187	68
http://www.digg.com/	21949	38
http://www.mininova.org/	21398	37
http://www.zamzar.com/	19145	15
http://www.imdb.com/	19125	54
http://vectormagic.stanford.edu/	16596	5
http://www.smashingmagazine.com/2007/01/19/ 53-css-techniques-you-couldnt-live-without/	15374	13
http://www.techcrunch.com/	15103	32
http://www.boingboing.net/	11613	69
http://www.scribd.com/	11336	12
http://www.opensourcemac.org/	11285	27

The last column of Table 4.3 shows the number of months that had passed since the URL was first posted to *delicious.com*[23]. Most of the extremely popular items shown in Table 4.3 had a relatively long history, in terms of the time that passed from its first addition to the site, but some achieved a comparable level of popularity within a short period of time.

In order to look at the accumulation of bookmark postings in the information space over time, Table 4.4 shows the number of URLs grouped by their age within *delicious.com* (the number of months that had passed since the URL was first posted to *delicious.com*), and the average number of postings by age group. The URLs that were first posted in the same month of the data crawling were marked as 0 months old. Figure 4.3 plots URLs by the number of postings and the age of the URL within *delicious.com*. Each point represents an individual URL and shows the relationship between the number of postings it has accrued and its age (the number of months). Note that the figure is a semi-log plot, with a logarithmic scale on the y axis.

[23] *Delicious.com* was found by Joshua Schachter in 2003. However, it was found that, for a number of bookmarks in the *URL History* dataset, their history goes further back, to before 2003. Since *delicious.com* provides a feature to import old bookmarks stored in browsers, those URLs with a longer history than *delicious.com* itself are most likely imported ones. Some of them may also be carried over from the precursor of *Delicious.com*, which was called Muxway (Souce: http://en.wikipedia.org/wiki/Delicious_(website)).

Table 4.4 Average number of postings by age of URLs

Age (in months)	No. of URLs	Avg. No. of postings per URL	Max. No. of postings per URL	URLs with postings > 1000
0	1784	7.51	1994	1
1	2883	18.17	5972	7
2	302	90.14	1691	8
3	261	83.04	1842	5
4	229	118.68	4183	4
5	207	183.84	16596	8
6	181	104.26	2510	6
7	153	124.47	2386	5
8	156	143.97	3677	4
9	157	192.57	4273	7
10	207	180.70	7731	6
11	128	163.23	1869	6
12	168	188.15	11336	5
13-24	1389	278.84	19145	95
25-36	1050	318.20	15103	80
37-	745	2037.48	31187	159

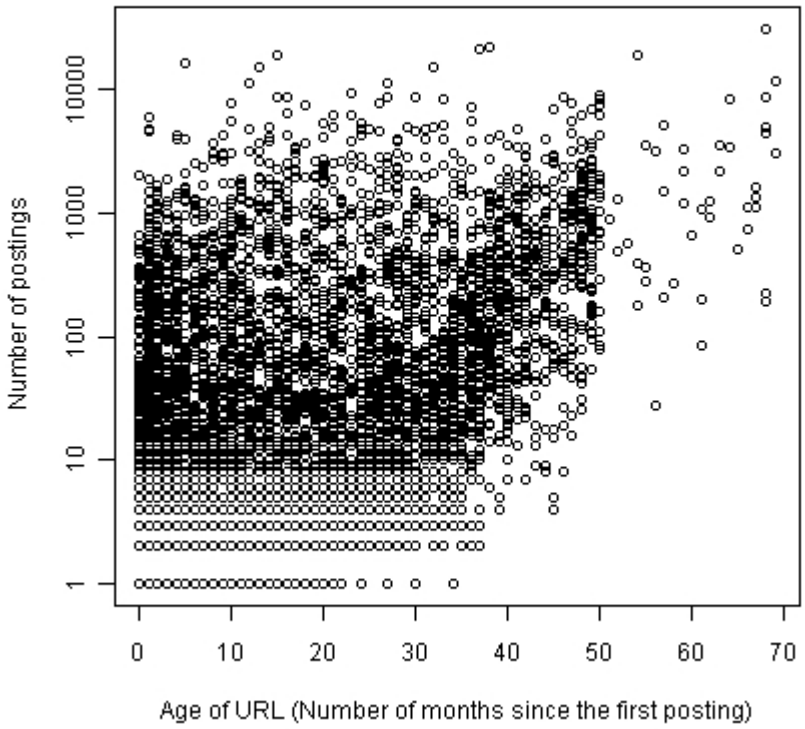


Figure 4.3 Relationship between the age of each URL and the number of postings

In any given period, there exists a great deal of variation in the popularity of information objects. Some bookmarks as old as 2 years or more still had only one posting, while some bookmarks gained the highest level of popularity within less than a month. When all bookmarks of the same age are aggregated, the average number of postings tends to increase as time goes on. Highly popular information objects with more than 1,000 postings are rather evenly distributed over ages, except that the last group (i.e., the oldest ones) has a disproportionately large number of highly popular items. This group consists mostly of those sites that are widely used by general Internet users. Examples include well-known commercial sites such as amazon.com and general news sites such as cnn.com and msnbc.msn.com. On the other hand, among the 8 URLs on which more than 1,000 postings were made

within about a month from its first appearance in *delicious.com*, half of them are blog or webzine articles newly published and the other half of them appear to be fairly new sites or services at the time of posting. An example is youtorrent.com, a meta search engine for torrent sites, which launched their service in January 2008 and had been bookmarked 5,972 times within a month.

4.2. User-centric view

In contrast to the above analysis of bookmarking patterns from a resource-centric view, the analysis in the following adopts a user-centric view. From a user-centric view a set of bookmark postings is seen as a set of users, each of whom has a collection of information objects.

In the following section, a number of statistics characterizing bookmarking activities of users, including the number and frequency of bookmark postings and the distributions of those quantities, will be examined. Each of these statistics was calculated in both the *Recent* dataset and the *User History* dataset. In addition, with the *Recent* dataset, the rate of distinct users to the number of postings will be compared with the same statistics involving information objects.

While the basic statistics depict bookmarking/usage patterns of individual users in *delicious.com* and suggest how individual activities accumulate in the information space, they do not show how much overlap exists among bookmark collections of different users. Whereas the distribution of information objects by the number of postings in and of itself shows the level of interest sharing (since the number of postings on a particular information object is in effect the number of users who share

interests in that information object), the number of bookmark postings of a user pertains only to the individual. Since one of the main questions addressed in this phase is concerned with shared interests among users, for each user the number of bookmarks he/she shares with one or more other users and the distribution of the quantity over all users were calculated in both the *Recent* dataset and the *User History* dataset.

In the following, the distribution of measures of individual bookmarking activities in each dataset will be presented first. Then the discussion will be moved on to the level of shared interests in terms of the proportion of shared bookmarks in bookmark collections of users.

4.2.1 The *Recent* Dataset

Figure 4.4 shows the number of users in the *Recent* dataset by the number of bookmarks they posted. Each point in the plot represents a set of users who have posted the given number of bookmarks.

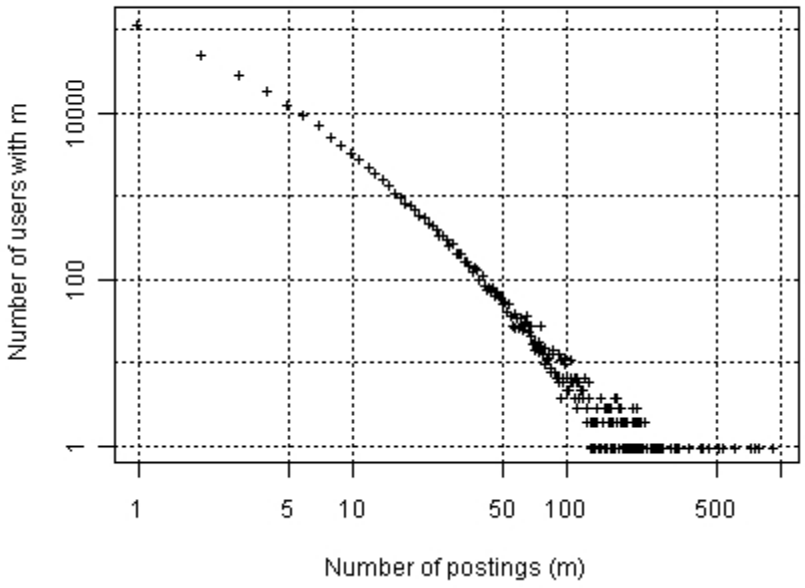


Figure 4.4 The number of users by the number of postings in the *Recent* dataset

The mean number of postings per user in the *Recent* dataset is 4.25. The median is 2 and the third quartile is 4. As can be seen in Figure 4.4, the distribution is highly skewed, indicating largely heterogeneous behaviors of users in their use of *delicious.com*.

Among the total number of 288,727 users in the *Recent* dataset, 121,002 users (41.91%) have only one bookmark, and 235,074 users (81.41%) have five or less bookmarks. It should be mentioned again that the data collection for the *Recent* dataset involved intervals between RSS fetches, and therefore it is likely that the proportion of users who posted more than once would be higher if the data collection process had been continuous. The figures should be interpreted as the overall pattern of distribution, not as the exact values. The important point here is the fact that there exist a large number of users who are less active and a small

number of users who are highly active. On the other hand, as shown in Figure 4.4, there are a small number of users who were very active during the period that the *Recent* dataset was collected. More specifically, 334 users have more than 100 bookmarks, with a mean of 169 bookmark postings each. This 0.1% of users account for 4.6% of the total number of postings in the *Recent* dataset. The highest number of postings is 935, which means that, on average, that user posted 9.5 bookmarks per day during the 14-week period.

In order to look at the accumulation of bookmarking activities of users over time, we examine the number of distinct users and the proportion of users who made multiple postings as the bookmark postings in the *Recent* dataset increased, in the same way that the accumulation rate associated with information objects was calculated (See Table 4.1). Table 4.5 shows the growth of distinct users in the *Recent* dataset and their accumulated activities, in terms of the proportion of users with multiple postings.

Table 4.5 Distinct users and users with multiple postings

Postings	Distinct users	Users with multiple postings (m > 1)	
		Number	Percent
10,000	6,941	1,424	20.52%
20,000	12,753	3,117	24.44%
30,000	17,828	4,884	27.40%
40,000	23,323	6,673	28.61%
50,000	28,582	8,517	29.80%
100,000	50,254	18,222	36.26%
200,000	82,455	35,657	43.24%
300,000	108,849	52,091	47.86%
400,000	131,712	67,341	51.13%
500,000	149,515	80,402	53.78%
600,000	167,541	93,247	55.66%
700,000	183,969	105,497	57.34%
800,000	197,020	115,806	58.78%
900,000	210,899	126,604	60.03%
1,000,000	234,093	138,506	59.17%
1,100,000	258,303	151,357	58.60%
1,200,000	282,447	164,358	58.19%

Tables 4.1 and 4.5 together show that the number of distinct URLs grows a lot faster than the number of distinct users. As the *Recent* dataset grew to contain

1,226,472 bookmark postings, the proportion of distinct URLs remained extremely high, barely dipping below 90% (978,990 distinct URLs). However, the total number of distinct user IDs in the *Recent* dataset is the much smaller number of 282,447. As the *Recent* dataset grew, more of the growth was brought about by existing users coming back to add new resources than by new users added to the set. Even within the relatively small window of time in which the *Recent* dataset was collected, users frequently came back to the system and added new resources. As the number of postings increased 100 times from 10,000 to 1,000,000, the number of users in the dataset increased about 34 times and the number of postings per user, accordingly, went up from 1.44 to 4.27.

Another measure of the level of user activity is the number of days on which a user posted one or more bookmarks. Figure 4.5 shows the frequency distribution of the number of different days on which each user posted one or more bookmarks. Each point in the figure shows the number of users by their number of posting days. This distribution, not surprisingly, is also highly skewed. The mean is 2.7, the third quartile is 3, and the highest value is 94.

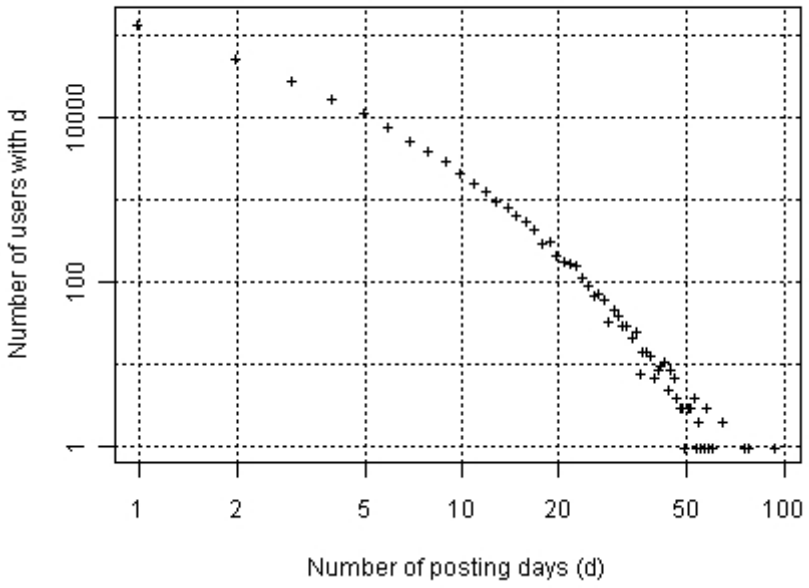


Figure 4.5 The number of users by the number of posting days

4.2.2 The *User History* Dataset

The *User History* dataset contains the entire bookmarking history of 10,000 users randomly selected from the *Recent* dataset. In total, there are 3,521,843 bookmark postings on 2,451,711 different information objects in this set. In order to look at the intensity and frequency of their use of *delicious.com*, the total number of bookmarks a user has and the number of days when one or more bookmark posting was made by the user was examined. In addition, since this dataset includes the entire bookmarking history of each user since their first use of the *delicious.com*, the age of each user account, that is, the duration of each user's membership with *delicious.com,* can be inferred from the history data.

Figure 4.6 shows the frequency distribution of users in the *User History* dataset by their number of postings. Each point in the plot represents a set of users

who have posted the given number of bookmarks in the dataset. On average, each user has 352.3 bookmarks in his/her collection. The first quartile, the median, and the third quartile values for the number of bookmarks per user are 32, 117, and 342.2, respectively. Again, users vary greatly in their degree of activity in terms of their total number of bookmark postings (the size of their bookmark collection), ranging from 1 and 20,859. Although this distribution is clearly skewed, with a small number of highly active users and a large number of less active users, compared with the distribution of URLs shown in Figure 4.2, it has far more 'middle' points placed in between the two extremes of the ones who rarely used the system and the ones who used it very heavily. While the vast majority of URLs (62.73%) in the *URL History* dataset have 10 or less postings, with 36.45% having only one posting, 88.55% of the users in the *User History* dataset have more than 10 bookmarks, and 53.45% of the 10,000 users have more than 100 bookmarks in their collection. Nearly half (45.64%) of the users in the dataset have between 100 and 1,000 bookmarks in their collections.

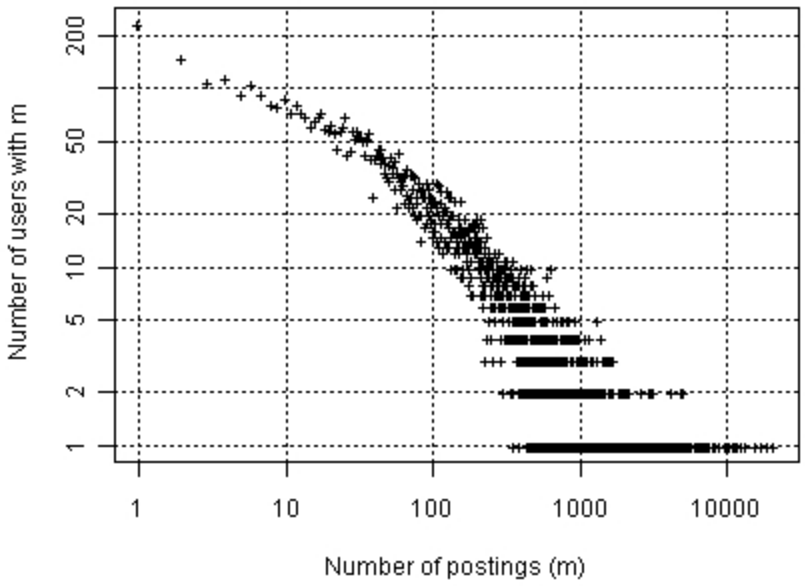


Figure 4.6 The number of users by the number of postings in the *User History* dataset

Since this dataset contains the entire history of each user, regardless of the duration of their membership, some variation in the number of postings may be due to the difference in membership duration. The mean is 388 days[24]. The first quartile, the median, and the third quartile of the membership duration in days are 73, 288, and 628 days, respectively. Each data point in Figure 4.7 represents an individual

[24] There are 15 cases where the first bookmark record does not have valid date information. There are 48 other cases where the first date goes back before 2003, the year *Delicious.com* (then Del.icio.us) launched their service. These cases were not caused by any error in data collection, but appear to be due to the importing feature of *delicious.com* which allows users to import old bookmarks from their browsers. Since it is obvious that, in these 61 cases, the number of days, obtained by calculating the difference between the date the account was crawled and the date of the first bookmark in the account, does not represent the duration of membership with *delicious.com*, these 61 accounts were excluded when calculating the statistics related to membership duration and drawing Figure 4.7.

user and shows the relationship between the number of days passed since the user made his/her first posting on *delicious.com* and the size of his/her collection (the number of postings he/she made ever since the first posting).

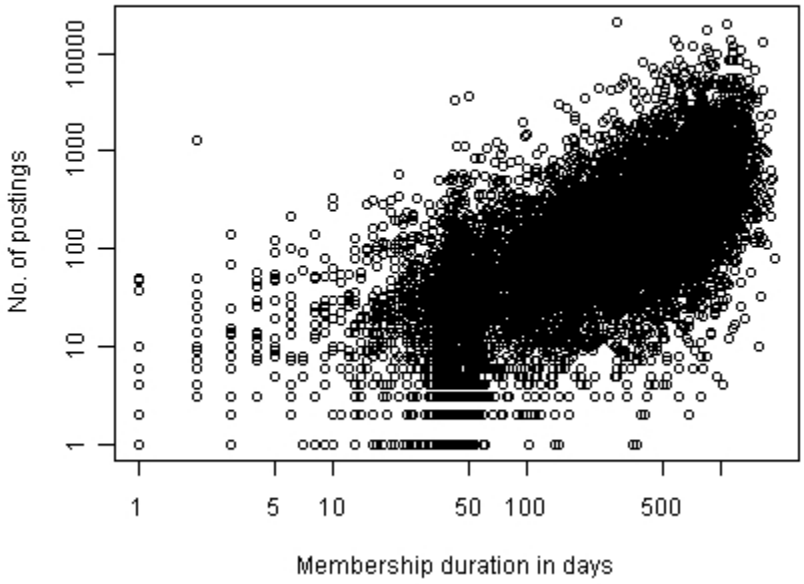


Figure 4.7 Relationship between the duration of membership and the size of bookmark collection

Although there exist considerable variations in the number of postings among users with equivalent membership duration and there are users who joined long ago yet have only a small number of postings, one can also see the general trends that users with longer membership tend to have larger collections. Spearman rank correlation indicates a positive correlation between the number of postings and the number of membership days (Spearman $\rho = 0.72$, $p < 0.00001$). Although such correlation seems intuitive, it is worth mentioning that Golder and Huberman (2006), one of earliest

empirical studies of social bookmarking using *delicious.com*, reported that there was no strong relationship between the length of membership and the number of bookmark posting.

4.2.3 Shared bookmarks

The above analyses looked at how individual users use the site and how their personal information space is built and expanded with continuing bookmarking activities. In order to look at the level of overlap across users, as well as at their accumulation of activity, the total number of bookmarks a user has and the number of bookmarks that he/she shares with other users were calculated and compared. For instance, if a user has posted 10 bookmarks, each of those 10 bookmarks was checked to see whether other users had also posted it. For the *Recent* dataset, only those users who are in the *Recent* dataset were considered. That is, a URL is considered shared if there exist two or more postings on that URL made by two or more users. For the *User History* dataset a bookmark is considered 'shared' if two or more *delicious.com* users among the 10,000 users in the sample bookmarked the same URL. Figures 4.8 and 4.9 show the scatter plot of users, from the *Recent* dataset and the *User History* dataset, respectively, by their number of bookmarks and the number of shared bookmarks. Those users who do not have any shared bookmark (151,456 users in the *Recent* dataset and 373 users in the *User History* dataset) were omitted from the plots. Both figures are log-log plots. In the *Recent* dataset, almost half of the users (47.54%; 137271 users out of 288727 users) have one or more shared URLs. In the *User History* dataset, 9627 users (96.27%) of the users have one or more shared URLs.

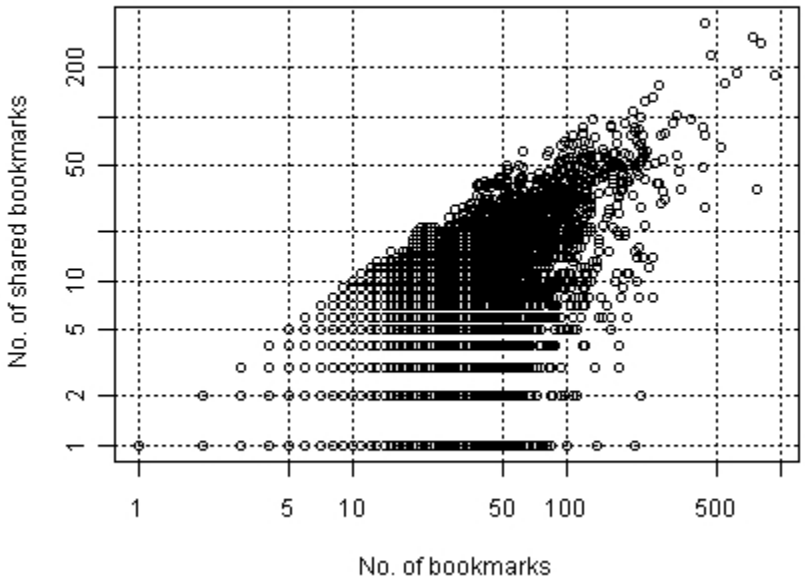


Figure 4.8 Relationship between number of bookmarks and number of shared bookmarks among users in the *Recent* dataset

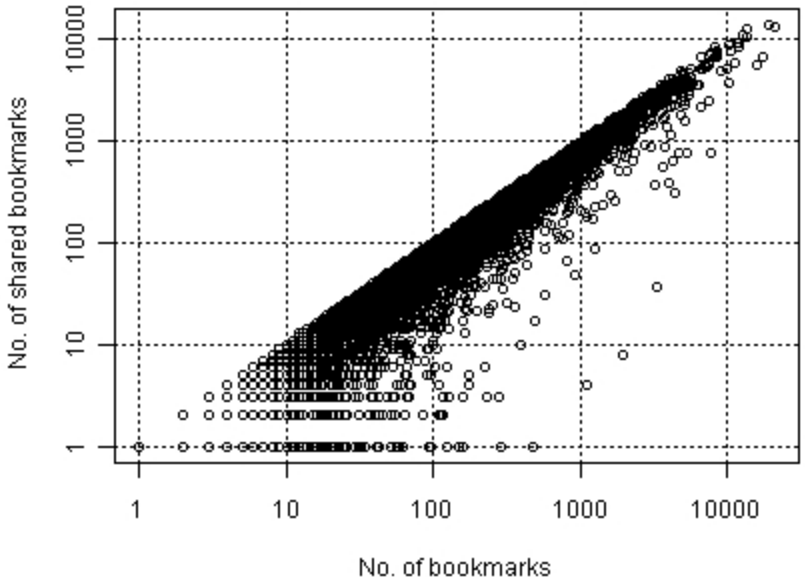


Figure 4.9 Relationship between number of bookmarks and number of shared bookmarks among users in the *User History* dataset

Note that the points in the plot representing users are relatively spread in Figure 4.8, while points in Figure 4.9 tend to converge on a rather straight line, except for a small number of outliers. In the *User History* dataset, the size of the collection (the total number of bookmarks) and the number of shared bookmarks are strongly correlated (Spearman $\rho = 0.965$, $p < 0.00001$). The patterns shown in Figures 4.8 and 4.9 suggest that, while many users have rather unique selections, with a large number of bookmarks not shared with others, the proportion of shared bookmarks grows over time and as the size of bookmark collection increases, in most cases. Whereas the proportion of shared bookmarks averaged over all users in the *Recent* dataset is 26.67%, the average proportion of shared bookmarks in the *User History* dataset is 66.65%. Moreover, within the *User History* dataset, the proportion of shared bookmarks is consistently higher for those users with larger collections. Table 4.6 presents statistics of the *User History* dataset grouped by the size of their bookmark collections, *m*.

Table 4.6 User groups in the *User History* dataset by the size of bookmark collection

No. of bookmarks (collection size)	No. of users	Avg. no. of bookmarks	Avg. proportion of shared bookmarks
$m <= 10$	1145	4.66	55.14%
$10< m <= 100$	3510	45.64	63.82%
$100 < m <= 1000$	4564	341.54	70.82%
$m > 1000$	781	2303.24	71.81%

In addition, Figure 4.9 shows a tendency that not only the proportion of shared bookmarks increases but also that it more or less stabilizes across users in the *User History* dataset. It should be mentioned, however, that the number of users in the *User History* dataset (Figure 4.9) is considerably smaller than in the *Recent* dataset, and therefore the different patterns may be in part ascribed to the difference in sample size.

Chapter 5. Results of Phase II

The second phase of this study is concerned with the question of whether and how users of a social bookmarking site can be connected based on their shared interests. The findings of the previous phase suggest that there exists a broad range of distinct information objects reflecting diverse user interests within the information space and that bookmarking activities accumulate on a significant portion of those information objects, making implicit connections among users possible. In this phase, we focus on taking advantage of such implicit connections to induce a social network of users.

While the first phase described the overall characteristics of the information space by looking at the basic statistics of bookmarking activities with three datasets representing different portions of the information space, in this phase we concentrate on the specific part of the information space where the most active users are involved and draw a network representing that part.

Network representation allows us to investigate intricate potential relations among users and uncover emerging patterns. When a large network is being analyzed, there can be two broad approaches to studying it. One approach is to examine various network measures that characterize the overall typological structure of the network. There is a set of well defined network measures that have been used for this purpose, both in the long tradition of social network analysis and in the more recent literature on complex network analysis. Another approach is to decompose a

large network into smaller parts, in an attempt to identify and extract important or interesting parts of the network based on particular features. In social network analysis, the umbrella concept of 'cohesive subgroups' has been used to represent the idea that actors within a social network tend to show patterns of clustering which allow the identification of different groups. Similarly, in the more recent field of complex network analysis, the term 'community' is often used to denote certain regions in the network that have tighter connections among the nodes in those regions. In this study, both of the above approaches were used.

In the following sections, a description of the dataset collected and used for this phase will first be presented. The basic statistics, including the average number of bookmarks per user and the average number of postings per information object, demonstrate that the region of the information space being analyzed in this phase indeed shows a different level of activity compared to the overall level of activity in the entire space examined in the first phase. We then proceed to the main focus of this phase, network analysis, starting with the discussion of the construction of the network, to the analysis of network properties, and to the iterative process of decomposing this network for identification of subgroups.

5.1 Dataset

For this phase, a subset of users from the *Recent* dataset was selected based on their level of activities, in an attempt to capture an *active* part of the information space. More specifically, users who made 9 or more bookmarks in 6 or more days during the period in which the *Recent* dataset was collected were included. The initial set of

users contained 23,387 users, filtered by the two inclusion criteria. For each of the users in the dataset, all the pages under their account, containing their bookmarks in reverse chronological order, were crawled and parsed to extract the bookmarks and the time of each posting. The crawling took place in May-June 2008. Among the 23,387 user accounts, problems were encountered with 169 accounts while crawling their pages. Manual examination revealed that there were some cases where the accounts did not exist anymore. In some other cases, the accounts themselves were not deleted, but no pages were stored for that account. In the latter cases, it may be that all the bookmarks had been removed or changed to be private. In the end, the entire usable dataset included 23,238 users whose crawled pages retained one or more bookmarks regardless of their posting time. The total number of bookmark postings collected is 25,559,506. The number of distinct information objects, represented by their respective URLs, is 13,633,750.

Note that the dataset contains each user's entire history of bookmark posting up until his/her pages were crawled. Therefore, the time span of collected activity varies by user due to two factors: 1) different users started using the service at different times, and 2) since the data collection took place over more than a month, some user accounts were accessed later in time and thus their accounts may contain recent bookmarks added after others users' accounts were collected. Since we have the time of posting for each bookmark posting, it is possible to draw subsets of bookmarking activities applying different time windows. Table 5.1 shows some basic descriptive statistics for the entire dataset and the subsets of three different time windows: the 12-month period from May 2007 to April 2008, the 6-month period

from November 2007 to April 2008, the 3-month period from February 2008 to April 2008. The second row of the table shows the number of users who have at least one bookmark posted within the corresponding timeframe. The third row presents the number of distinct information objects that have ever been bookmarked during the period.

Table 5.1 Basic statistics of datasets

Time window	Entire history	12 month (2007.05-2008.04)	6 month (2007.11-2008.04)	3 month (2008.02-2008.04)
Total no. of postings	25,559,506	14,438,954	8,834,495	4,644,558
No. of users	23,238	23,218	23,217	23,172
No. of URLs	13,633,750	8,349,081	5,401,733	3,036,359
Avg. no. of postings per user	1095.82	621.89	380.52	200.44
Avg. no. of postings per URL	1.87	1.73	1.64	1.53

It is worth mentioning that this set of users indeed show a higher volume of activity than the random set of users and their history data analyzed in the first phase. With the set of 10,000 randomly selected users, the average total number of postings per user was 352.2, and if different time windows are applied in a similar fashion for the random *User History* dataset, the average numbers are 215.1, 132.4, and 78.7 for 12-month, 6-month, and 3-month periods, respectively. In general, users in this Phase 2 dataset show almost three times the volume of activity in any given period.

Not only is the average level of user activity higher than that of the random users (in the *User History* dataset), the shape of the distribution is different. Figure 5.1 shows the distribution of the size of users' bookmark collections (i.e., the number

of bookmarks) with the 3 month-period data (See Figure 4.4 for the distribution in the *User History* dataset). Although there is still a great deal of variation in the collection sizes, unlike the random dataset where the most frequent values are found at the lower end of the scale (below 10), the users included in the Phase 2 analysis have a unimodal distribution with a peak around 100 and an exponential tail. With 90% of the users having 45 or more bookmarks, the values in this distribution are less spread. The Inter Quartile Range (IQR) of the Phase 2 distribution is 152 (234 – 78), while the IQR of the *User History* (Phase 1) data is 310.2 (342.2 – 32).

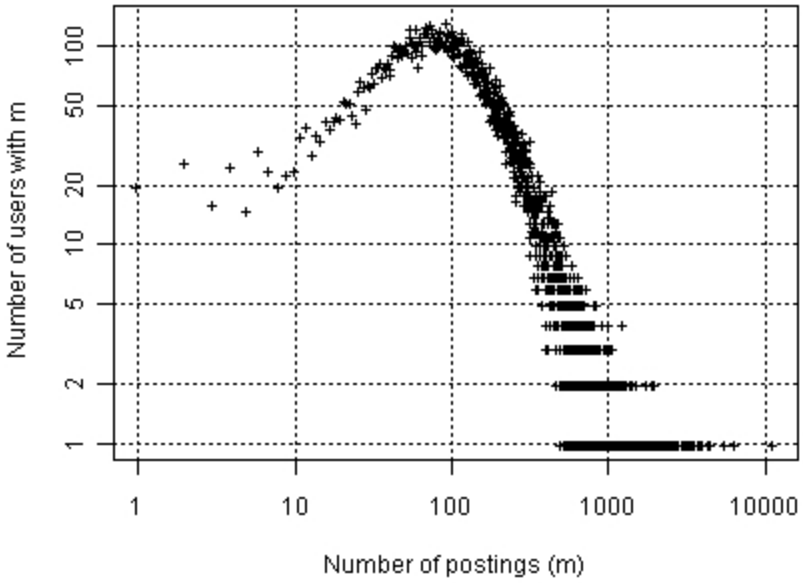


Figure 5.1 The distribution of the number of bookmarks posted (3 month data)

If we compare the last two rows in Table 5.1, it is notable that, while the average number of postings per user increases almost directly proportional to the length of the time window, the average number of postings per URL does not change substantially. Note that the number of postings for each URL here is measured within

this particular dataset. Instead of counting all the postings made by any *delicious.com* users, only the postings made by users included in this dataset were counted. Our interest here is not in the general popularity of an information object in *delicious.com* but its popularity among the subset of users within this dataset who share interests in it. Figure 5.2 shows the distribution of the URLs by the number of postings in this set.

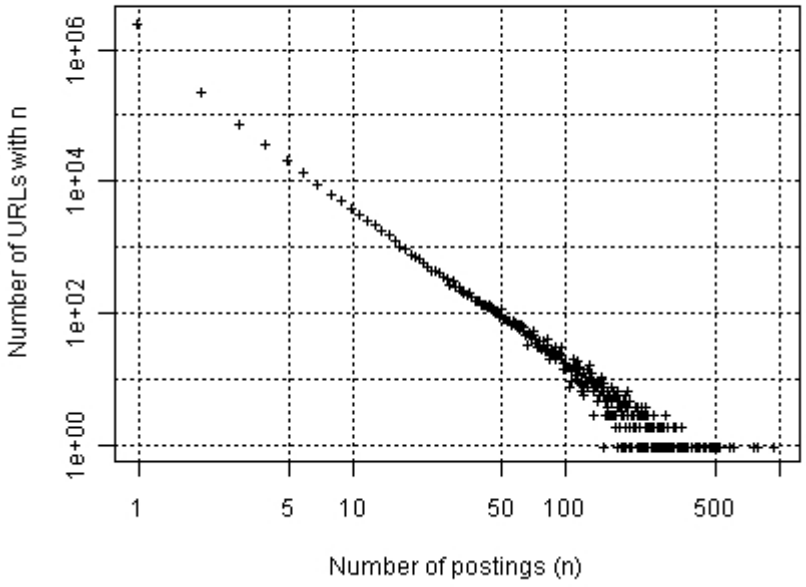


Figure 5.2 Frequency distribution of URLs by the number of postings

During the 3 month period, among 3,036,359 URLs, 2,599,557 (85.6%) were bookmarked only once by the users in our Phase 2 dataset. Although the vast majority of the URLs do not contribute to making connections between users, there are still almost 15% of the URLs (436,802) that will build one or more connections among users. The highest value for a single URL is 948 (4% of the 23,172 total users bookmarked this particular information object).

5.2 Network construction

A network is comprised of a set of nodes and links between pairs of nodes. In a social network, nodes represent actors or subjects and links represent relations of shared interests among them. The problem of the selection of actors to be included in the analysis of social networks is known as boundary specification (Marsden, 2005; Doreian and Woodard, 1994). For the network of *delicious.com* users in this study, network boundaries were specified by inclusion criteria based on two measures of recent user activity – the number of postings and the number of days on which one or more postings were made. How to define relations among users is another key problem in instantiating a network. As will be shown below, the relations between pairs of users were drawn by projecting their affiliation relations, which were derived from their bookmarking records. In effect, we define a relation of shared interests between a pair of users based on the bookmark(s) they have in common.

More specifically, the derivation of the network of *delicious.com* users based on their bookmarking records involved two main steps. First an affiliation network consisting of users and information objects was built. The affiliation network in this case has two modes, one with information objects and one with users. Second, the two-mode affiliation network is transformed into a one-mode network of users, where relations among users are defined by their common possession of one or more information objects. As a result, the network is constructed such that two users who have one or more bookmarks in common are connected.

5.2.1 Affiliation networks and one-mode projection

In the framework of social network analysis, an affiliation network represents and allows empirical investigation of a structure interwoven by various social groups and individual members thereof. In an affiliation network, each individual is connected to groups to which he/she belongs and each group has links to its members. Individual actors in this network are not directly linked to one another, but are connected through their common membership in one or more different groups.

A group in an affiliation network can be defined broadly to include any cluster of individuals with a certain commonality. A group in the context of this study is defined as the set of users who have bookmarked a particular information object. Conceptually, by bookmarking an information object, a user joins the group of users who are gathered by their shared interest in that information object.

Since there are two modes, actors and groups, constructing an affiliation network involves defining each mode. In this study, actors to be included were filtered by their level of recent activity. Groups (information objects in this case) were drawn from bookmarking records of those users. The initial set of information objects was the aggregate set of the entire bookmark collections of the users, regardless of the time of bookmark posting. In addition, three subsets based on different time windows (12 months, 6 months, and 3 months) were formed. The size of each set is shown in Table 5.1, above. For each time window, an affiliation network can be constructed by drawing an edge between each user and each information object he/she bookmarked.

As most network analytic techniques assume a one-mode, simple, undirected

network, a two-mode affiliation network is often transformed into a one-mode network for further analysis[25]. Figure 5.3 shows an example of such transformation. On the left, a two-mode network sampled from our dataset is shown. For the purpose of demonstration, seven URLs/bookmarks were selected first and, for each URL, users who bookmarked the URL were added. Other URLs bookmarked by those users were omitted from this simplified example. The nodes belong to the first mode (users) are colored blue, and the nodes in the second mode (information objects) are colored red. Note that edges exist only between users and information objects, and a path between a pair of users is created by information objects they have in common. On the right (Figure 5.3-b), the one-mode network of users drawn from the affiliation network is presented. The implicit relations among users based on their shared bookmarks are now projected to create direct links between users. In this transformed network, two users are connected if they have at least one bookmark in common. Note that the network has a large connected component and a few isolated nodes.

[25] Since there are two modes, it is possible to derive two separate one-mode networks from an affiliation network: a network of actors and a network of groups. A general practice is to choose one, depending on the main question of interest. In this study, since we are mainly interested in how users are connected by shared interests, the actors' network was chosen.

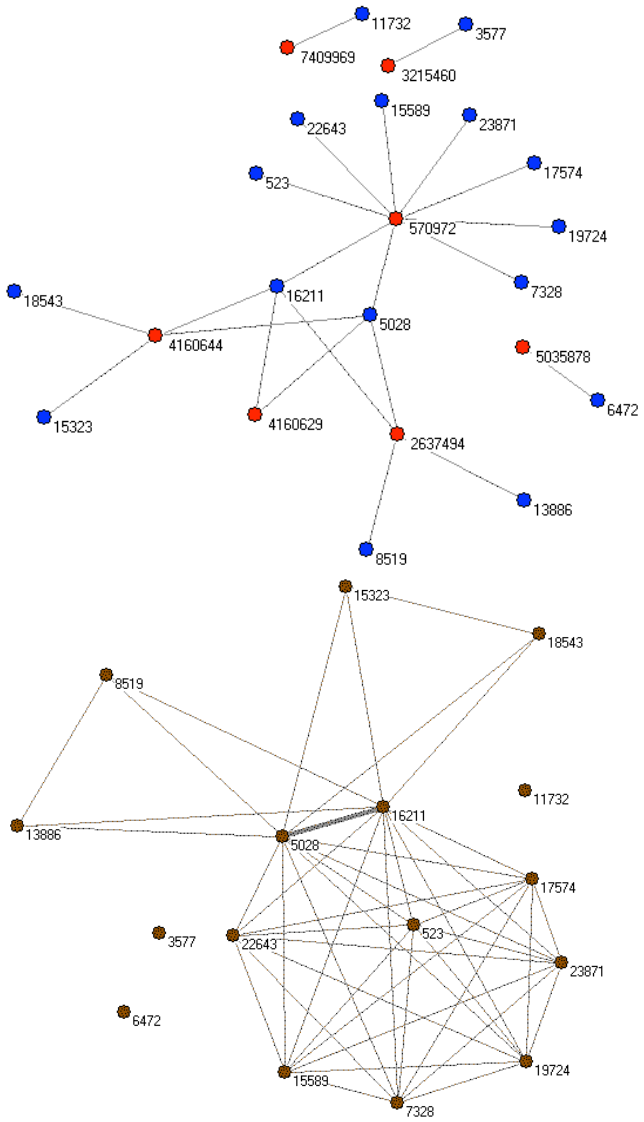


Figure 5.3 Transformation of a two-mode network into a one-mode network. (left) Two mode network; (right) One mode projection of the left

In order to better understand the figure and the one-mode projection of an affiliation network in general, a few points need to be noted. First, in the one-mode network, all the users who bookmarked a particular item now form a complete clique, being fully connected to one another. For instance, we can see in the two-mode network on

the left that an information object, numbered 570972, was bookmarked by 9 different users. These 9 users are considered to have a group relation by the virtue of their common possession of the information object (#57092). The clique comprised of nine nodes at the bottom of the one-mode network in Figure 5.3-b represents this group relation. Note that one information object (#570972) in the two mode network has generated 36 (9*8/2) edges in the one-mode projection. This explains why a network derived from a two-mode affiliation network tends to be denser than a social network built on pair-wise relations in the first place. Second, information objects that were bookmarked by only one user do not create any connections, and those users are left isolated in the one-mode projection. Therefore, the number of neighbors a user has in this network, a node degree, depends on the popularity of each information object he or she has bookmarked, as well as how many shared bookmarks he or she has. Finally, the transformation entails loss of information. The one-mode network shows whether two users share a bookmark or not, but in this network there is no information as to which information object(s) they share[26]. Note, however, that user 5028 and user 16211 have four information objects in common (on the left), and the link between them in the transformed network (on the right) is thicker. In the process of transformation, when edges are added between each pair of users, it is possible to assign weights to each edge indicating the number of shared connections (shared bookmarks in this case) that the two users have in the two-mode network.

[26] Since the information about information objects involved in making connections among users is lost in transformation and either of network analysis tools (*Pajek* or the *igraph* package in *R*) used in this study does not support the reverse operation, in order to be able to trace an edge between a pair of users back to the information objects that they share in later analysis, a set of database tables were created to maintain detailed information about links before/after the transformation.

Since the weight corresponds to the multiple connections transformed into an edge, it is called line-multiplicity. It is in fact a measure of the strength of the tie between a pair of users. As mentioned above, however, most network analytic measures and techniques take a simple unweighted network, and therefore the weights are ignored in the analysis.

As stated above, with the dataset containing the entire history of 23,238 active users, three different affiliation networks were constructed with subsets of information objects applying different time windows (12 months, 6 months, and 3 months). In other words, while the first mode, users, remained basically same, the second mode containing information objects was iteratively reduced with shorter time windows. The number of users and the number of URLs for each time window in Table 5.1 are equivalent to the number of nodes in the first mode and the second mode, respectively. Each of the three affiliation networks was then converted to create a one-mode network of users. Table 5.2 below shows the basic statistics of each of the derived one-mode networks.

Table 5.2 Basic statistics of one-mode projection of users

Statistics	12 months	6 months	3 months
No. of nodes (users)	23,218	23,217	23,172
No. of edges	65,574,879	40,294,283	17,918,940
No. of isolated nodes (d = 0)	140	162	279
No. of users with neighbor(s) (d >= 1)	23,078	23055	22,893
Average no. of neighbors (degree)	5,648.62	3,471.10	1,546.60

Overall, the users in the Phase 2 dataset are very well connected to one another. As can be seen in Table 5.2, when the 12-month window is applied the average number of neighbors (the average degree) per node is 5,648. A user, on average, is connected to a quarter of all the users in the set. The level of connectivity in this dataset turned out to be much higher than anticipated[27].

Not surprisingly, the longer is the time window, the greater is the number of edges connecting users. What is interesting is that, even with the short time window of three months, most users in this set are connected with some other users in this set. The percentage of isolated users (i.e., users who have no connection with other users) is only 1.2% in the network drawn from the 3-month data.

[27] Due to the large number of edges, the network analysis software, *Pajek*, which is known for its capacity to handle a large dataset, failed to read the 12-month data (on a Windows machine with with a 3.0GHz Pentium 4 processor and 8G RAM). The 6 month data was read into the program as a two-mode network, but the conversion to a one-mode network was aborted due to an out-of-memory error. Another network analysis tool, the *igraph* network package in R (Windows version), which handles fairly large datasets, also failed to process the entire dataset due to the volume. Therefore, the statistics in Table 5.2 were calculated through a number of steps, by running scripts and storing intermediate results in a database.

Given the fact that about 99% of users are still connected to the network when the 3-month dataset is used, while the computational cost of drawing a network with a longer time window is preventatively high, the 3-month window was used for all further analysis. Another advantage of using a short time window, other than the efficiency of computation, is that it can reduce the effect of the variability in different membership duration on the resulting network. All but 503 of the users made their first *delicious.com* posting before February 1, 2008, but there are 9,480 users whose first posting appeared after May 1, 2007. That is, if we take the 3-month period, most of users had already been using *delicious.com* before the period started, but with the 12-month period, a large portion of users in the set had not started using *delicious.com* at the beginning point of the period.

5.3 Network analysis

By representing bookmarking activities as affiliations, a social network of *delicious.com* users was derived. In this network, social relations were defined by the existence of shared interests, inferred from bookmarking choices made on the same information object(s) in the given period, between February 2008 and April 2008. As shown in Table 5.2, the resulting network has 23,172 nodes representing users and 17,918,940 edges created from shared information objects.

As discussed before, there are broadly two approaches one can use to analyze a large network. The first approach relies on statistical measures. While a graphic representation or visualization of the network is an intuitive and informative way of examining the structure of a network, especially for a network of small to moderate

size, when the size of a network is larger than a few thousand it is hard to draw a network, and even harder to depict any meaningful information about its structure. Thus, with large networks, statistical measures or indices are used to understand and characterize a network. There are global measures, such as density or characteristic path length of a network, that characterize a network as a whole. On the other hand, local measures such as degree describe structural properties of individual nodes, pairs, or subsections within a network. The distributions of the local metrics, for instance the degree distribution, provide a way to visually examine the network topology.

The second approach is to decompose the large network into smaller networks that can be further analyzed. The goal of this approach is often to identify and extract important or interesting parts/regions within the large network.

In this section, both approaches are adopted. The first part of the analysis will examine the network that is induced from the affiliation network based on 3 months of bookmarking activity. In order to understand the characteristics of this network, basic structural properties including connectivity and clustering are analyzed. As discussed above, in the process of transforming a two-mode affiliation network into an one-mode social network of users, the data on which information objects users have in common are lost, but we can retain the information on how many bookmarks a pair shares, as the weight of the edge connecting them. However, since most of the network measures assume simple binary relations, the weights of edges are ignored in the analysis of structural properties.

In the second part of the network analysis, the weights play an important role. In an attempt to decompose this network and identify subgroups, a technique called

m-slice (or m-core) is applied. This technique is based on the idea of the tie strength or the intensity of the connection and defines a subgroup/subgraph based on a minimum weight for each edge among members of the subgroup.

5.3.1 Network properties

Table 5.3 shows the basic statistics and properties of the network obtained with the 3-month period data. The number of nodes in this network (that is, the number of users included in this network) is 23,172. As stated before, the users are, from among the initial set of 23,238 active users filtered from the *Recent* dataset, those who have one or more bookmarks posted between February 2008 and April 2008. Out of the 23,172 users, there are 279 users who have no shared information objects with any other users in the dataset, and so are isolated nodes in the network. (The number of bookmarks averaged over the 279 isolated users is 117.9.) The number of users with one or more neighbors in this network is, therefore, 22,893. The measures shown in Table 5.3 are the most commonly used network analytic measures that have to do with the overall connectivity and the level of cohesion in the given network. They are described and discussed here.

Table 5.3 Basic properties of the network

Properties	Values
Total number of nodes *n*	23,172
Total number of edges *m*	17,918,940
Number of isolated nodes	279
Density	0.0667
Average degree *z*	1546.60
Diameter	6
Average shortest path length (distance) *l*	2.01
Global clustering coefficient C_1	0.32
Average node clustering coefficient C_2	0.43

5.3.1.1 Density and Degree

Density and average degree measure how dense or tight the connections are within a network. Density is a global metric characterizing the overall cohesion of the network in terms of the volume of the interactions. It is defined as the ratio of the edges present in the network to the maximum possible edges that may exist between the given number of nodes. Since this network is an undirected network, the number of possible edges would be $n(n-1)/2$, in this case (23,172 * 23,171) / 2 = 268,459,206. That is, if the 23,172 users in this network were fully connected to one another, there would be 268,459,206 edges. The actual edges between users in the network, established by their common bookmarks, amount to 17,918,940 which is about 6.7% of the maximum possible edges. It should be noted that a network with a large number of nodes tends to have a small density figure because, while the number of possible edges grows quickly as the number of nodes increases, the actual connections a new node adds to the network are hardly proportional to the size of

the network. The density of 0.067 in a network with as many nodes as this network is remarkably high (for a comparison with other empirical networks, see Table 7.1 in Chapter 7).

The degree of a node is the most basic property of a network node, and accounts for the structural cohesion of a network. In an undirected network, node degree is measured by the number of neighbor nodes to which a node is connected. When the node degree is averaged over every node in the network, the average degree serves as an indicator of the level of connectivity for the network as a whole. As shown in Table 5.3, the average degree is a high value of 1546.60. In other words, on average, a users in our dataset is connected to about 1546 other users in the dataset (of 23,172 users). However, for this dataset, the average does not represent a typical case because there is a great deal of variability in the distribution, with the standard deviation being 1687.77. The large average number of neighbors suggests that users in this network are highly interconnected, but the high standard deviation indicates a high degree of heterogeneity. This can be further examined with the distribution of degrees.

5.3.1.2 Degree Distribution

The degree distribution of a network is one of the most prominent network properties, being frequently used to understand the internal structure and the topology of a network in terms of connectivity. The degree distribution shows the number of nodes or the proportion of nodes in the network with a given degree.

Figure 5.4 shows the degree distribution of the network. In the plot, the

horizontal axis represents node degree *k*. The plot shows the number of nodes that have degree *k*.

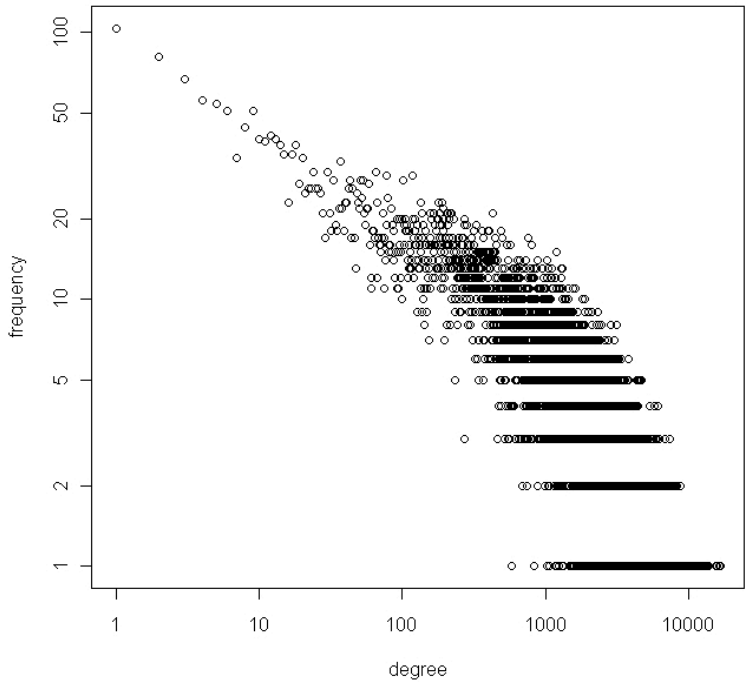


Figure 5.4 Degree distribution of the network.

The distribution is inhomogeneous and individual nodes (users) vary substantially in their connectedness. There are 279 users with degree 0 (excluded from graph) and 104 users with 1, while there are 42 users who have a degree greater than 10,000 being connected to about the half of all the users in the network. What is remarkable in the degree distribution in this network is that the majority of the users are well connected. 88% of users have a degree greater than 100 and 50% of users have more than 1,000 neighbors. About 5% of users have degree higher than 5,000. The highest degree in this network is 16,500; in other words, this particular user has

connections to about 71% of all other users in this network by sharing at least one bookmark with each of them.

5.3.1.3 Distance

Average path length and diameter are measures related to distances between nodes. The distance between a pair of nodes is given by the shortest path length between the two nodes, which is the minimum number of edges that need to be traversed to move from one node to the other. The average path length of a network is the average of the shortest path lengths between all pairs of nodes in the network; the diameter of a network is the maximum distance between any pair of nodes in the network. These measures look at patterns of connections in terms of how far apart users in this network are or, on the flip side, how close and accessible the users are to one another. As shown in Table 5.3 both the average path length and the diameter take small values. The average path length is only about 2, and the diameter is 6. This means that in this network, a pair of users is typically only two links away, and it takes a maximum of six links for a user to reach any other user.

Figure 5.5 shows the distance distribution of the network. Each point in the plot shows the number of pairs of nodes at a distance x. As shown in the figure the majority of users are reachable to one another within the distance of 2.

Figure 5.5 Number of pairs by the length of the shortest path (distance)

5.3.1.4 Clustering coefficient

The next measure examined is the clustering coefficient. The clustering coefficient assesses the local cohesiveness of a network and the extent to which nodes tend to form local clusters in the network. There are two different yet equivalent measures of clustering that have been widely used in empirical network studies: the global clustering coefficient and the average of the node clustering coefficients. Table 5.3 shows those two measures for the network of *delicious.com* users.

The global clustering coefficient is defined as the proportion of triangles with respect to the number of connected triples in a network, where a triangle is a set of three vertices each of which is connected to both others and a connected triple refers to a set of three vertices at least one of which is connected to the other. It gives

the probability that two nodes with a common neighbor are connected to each other. As shown in Table 5.3, the global clustering coefficient of our network is 0.32. When two users with a common neighbor are chosen, there is a 32% chance that those two, themselves, are connected to each other.

The second measure is the average of the local clustering coefficients. The local clustering coefficient, sometimes called the node clustering coefficient, is measured for each node. Instead of counting triadic connections in the whole network, the local measure examines triadic relations surrounding a particular node. For instance, if node *A* has 3 neighbors (*B, C,* and *D*), there are three possible triadic relations with *A* in the center (*B*-*A*-*C*; *B*-*A*-*D*; *C*-*A*-*D*). Note that the number of triples centered on a node equals to the possible number of pairs among its neighbors. If there is actually an edge connecting a pair of its neighbors, the three nodes (the node in question and the connected pair of its neighbors) form a triangle. The local clustering coefficient, therefore, is defined as the ratio of the number of existing edges connecting its neighbors to the maximum possible number of edges (possible pairs) between the neighbors[28]. It is, in effect, the number of triangles including a node over the number of triples centered on the node. From the above example, if there is an edge between *B* and *C*, and one between *B* and *D* (but not between *C* and *D*), then the number of triangles among *A*'s neighbors is 2. The local clustering coefficient of *A* is 2/3. Since it is a local measure, the clustering coefficient for the whole network is obtained by averaging the node clustering coefficients over all the nodes in the network. The clustering coefficient of this

[28] In an undirected network, for a node with a degree k, the possible number of edges between its neighbors (i.e, the possible number of pairs) is $k(k-1)/2$.

network, the average of local clustering coefficients, is 0.43. It means that, when a node is randomly selected from this network, the probability that a pair of its neighbors is connected is 43%.

It is often useful to see whether the observed clustering coefficient for a network is higher than expected at random. When an Erdös Rényi (ER) random network (Erdös & Rényi, 1959) of the same size (with the same number of nodes and the same number of edges) was created, both the global clustering coefficient and the average local coefficient were calculated to be about 0.07. For both clustering measures, this network reveals a much higher tendency of local clustering.

For the local clustering coefficients, since each node has a measured value, it is possible to see the distribution of values over all the nodes. Figure 5.6 shows a histogram of the local clustering coefficients in this network.

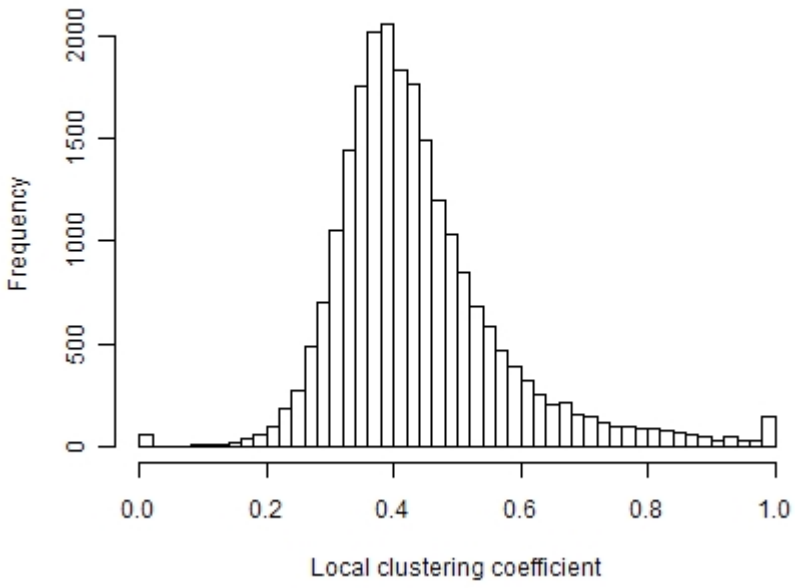


Figure 5.6 Histogram of local clustering coefficient

With local clustering coefficients, one can also look at the correlation between the node clustering coefficients and other properties of individual nodes. For instance, in order to examine the correlation between local clustering coefficient and node degree, we can define and plot a function that gives the average clustering coefficient of all nodes with degree *k*. Figure 5.7 shows the average local clustering of nodes with a given degree *k*. As in many other empirical networks, this network exhibits a negative correlation between the clustering coefficients and the degrees, that is, nodes with lower degrees have higher local clustering than those with higher degrees.

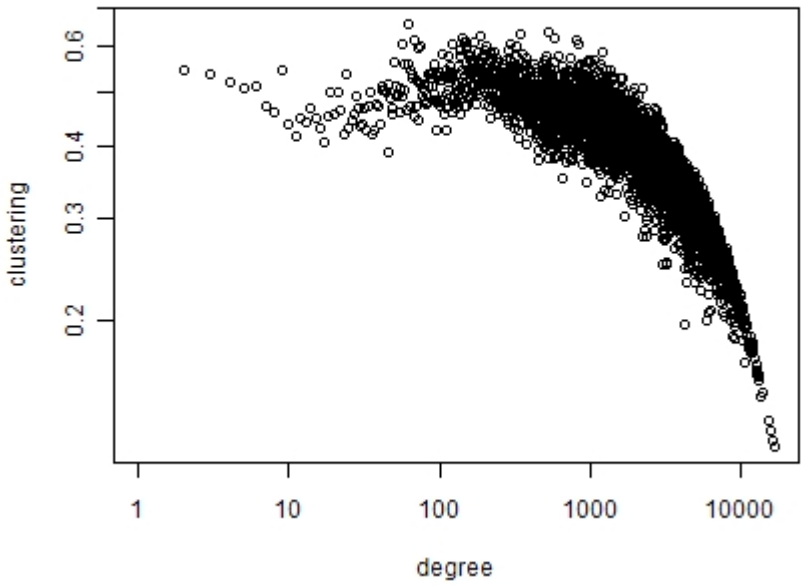


Figure 5.7 Local clustering coefficients averaged over nodes with a given degree

5.3.1.5 Components

While the clustering coefficient measures how much local structures exist in the network, the macro structure of the network can be examined by looking at the components that comprise the network. A *component* is a partition of a network

where each pair of nodes is connected by a path. All the nodes belonging to a component are reachable to one another, while any nodes belonging to different components are completely disconnected from the nodes in the component of interest. The number and sizes of components within a network, therefore, outline how the network is segmented.

In many empirical networks, a significantly large portion of the network belongs to the largest component, called the giant component. We observe a similar, yet far more extreme, situation with the network of *delicious.com* users. In this network, partly due to the high density, the giant component contains the vast majority of the nodes (98.8%) in the network. Only 1.2% of users are disconnected from the giant component.

While the number of components in the network of *delicious.com* users is 283, 279 components are comprised of only one user, that is, an isolated node. Therefore, there are only 4 components with a size of two or more. Moreover, while the giant component includes 22,887 nodes, 98.8 % of the entire network, the second largest component has only two nodes.

It is worth mentioning that not only are nearly all users located in a single giant component, being connected to one another through at least one path, but the paths connecting pairs of users in the component are typically very short, as shown in Figure 5.5 above.

5.3.2 Network decomposition

In the previous section, we looked at some of the global metrics that carry

information about the structural characteristics of the whole graph (the entire network), and some local measures that describe the properties of nodes and edges. While those metrics and the distributions of network measures characterize the underlying structure of the network, it is often useful to decompose a network into smaller regions to better understand how it is constructed.

The basic segment of a network is a component since each component by definition makes an isolated section of the network. Therefore, the component analysis often constitutes the first step in examining the substructures of a network. In a large network with several thousands of nodes or more, components that have a considerable number of nodes (e.g., a few hundreds), besides the giant component, can be regarded as separate modules or communities within the network. As described above, however, the network that we construct based on the 3-month bookmarking activities of *delicious.com* users has a remarkably large giant component comprised of 98.8% of the entire network, leaving the second largest component with only two nodes. Although it is interesting that the network comes in a shape of one big lump and almost all users in this network belong to one community (in a loose sense), this does not tell us much about a subgroup structure that may exist within the network.

In order to further divide nodes into groups of higher cohesion, we used a technique called *m*-core (Scott, 2000) or *m*-slice (Nooy et al., 2005)[29]. An m-slice is

[29] Scott (2000) introduced this technique as a variation or an extension of *k*-core analysis and named this technique as *m*-core and emphasized its basic similarity to *k*-core approach. However, Nooy et al. (2005) chose to call it *m*-slice instead of *m*-core in order to avoid confusion. We use the term *m*-slice for the same reason.

defined as “a maximal sub-graph in which each line has a multiplicity greater than or equal to *m*” and shows “a chain of points connected by lines of the specified multiplicity” (Scott 2000, p. 112). Whereas other techniques such as clique analysis or *k*-core analysis make use of the volume of connections as a basis for identifying subgroups, *m*-slice analysis uses the strength of connections as the defining characteristic for a group. The technique is applicable to this network because there exists an intuitive measure of tie strength built into the network. As a product of the transformation from a two-mode affiliation network to an one-mode social network, each edge in this network has a weight, called line multiplicity, which is the number of information objects involved in making the particular connection between two nodes. In other words, each edge in the network of *delicious.com* users carries the number of shared bookmarks between the pair of users being connected as a weight or a line value. Since the network analytic measures and techniques we have used so far consider only an unweighted network, the line values (weights) have been ignored in all the analyses presented above. With *m*-slice analysis, the number of shared bookmarks of two users can now be used as a measure of the intensity of connections, which arguably reflects the degree of similarity of their interests.

5.3.2.1 Distribution of line values

In the network of the *delicious.com* users constructed in this phase, two users are connected by an edge if they have one or more information objects in common. The total number of edges (17,918,940) in this network, therefore, is the number of pairs of users who share one or more bookmarks. The information on exactly how many

bookmarks each pair of users has in common is represented as the line value or the weight of the edge connecting the two users.

Since the *m*-slice technique will be based on the line values, the distribution of the line values was first examined. Figure 5.8 shows the number of edges for a given line value, in other words, the number of user pairs connected by the given number of shared bookmarks. Note that the points in this log-log plot appear on a straight line, which is the signature of a power-law distribution. It shows that the vast majority of the users have only a few shared information objects with other users, while a small number of users share a significantly large number of information objects. More specifically, among the total of 17,918,940 edges, 13,267,132 edges (74%) take a value of 1. That is, 74% of all the pairs of users in this network have only one information object in common. Table 5.4 shows the number of edges with low values ranging from one to ten. As can be seen, 93.8% of all the edges have a value of 3 or less and 99.2% have 10 or less. On the other end, there are eight edges with a value higher than 600 with the maximum value of 2,018.

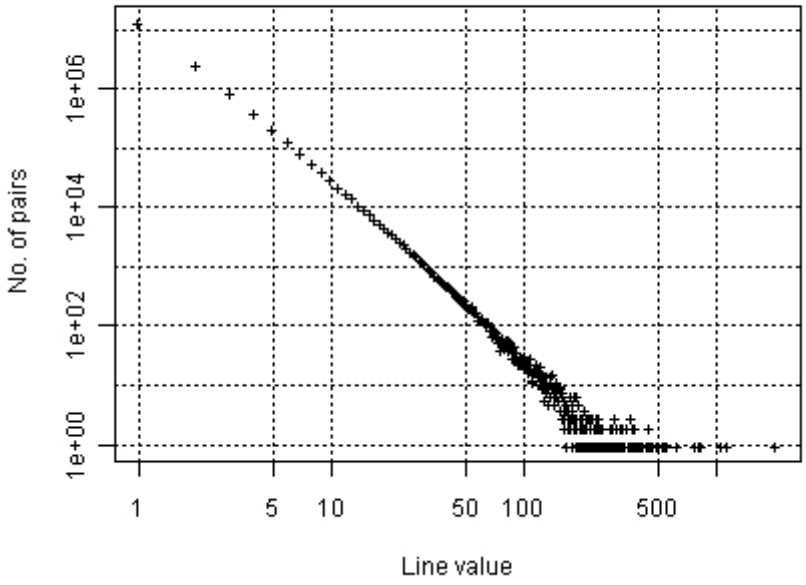


Figure 5.8 The distribution of edges by line values (weights)

Table 5.4 The number of edges with low line values

Line value	No. of edges (frequency)	Cumulative frequency	Percentage	Cumulative percentage
1	13,267,132	13,267,132	74.04	74.04
2	2,637,937	15,905,069	14.72	88.76
3	898,567	16,803,636	5.01	93.78
4	407,245	17,210,881	2.27	96.05
5	218,484	17,429,365	1.22	97.27
6	131,142	17,560,507	0.73	98.00
7	84,599	17,645,106	0.47	98.47
8	57,429	17,702,535	0.32	98.79
9	41,358	17,743,893	0.23	99.02
10	30,779	17,774,672	0.17	99.19

5.3.2.2 *m*-slice analysis procedure

An *m*-slice is a sub-network defined by the line multiplicity values. For a given *m* value, the *m*-slice consists of edges that have a value of *m* or higher and nodes that are incident on those edges. In other words, every node in an *m*-slice has a connection of the strength (line value) equal to or greater than the given value *m*, with at least one other node.

The basic procedure of *m*-slice analysis is similar to that of hierarchical clustering using a divisive method. Starting from the original network, edges and nodes are progressively removed as the value of *m* increases, and the original network is iteratively broken down into smaller sub-networks, each of which is characterized by the minimum tie strength of respective *m*. More specifically, for a given value *m*:

- First, the edges with a value less than m are located and removed. Those edges are the weakest connections within the (sub)-network at hand.
- Second, any isolated nodes are removed. After the removal of the edges in step one, some nodes may become isolated. If a node does not have any connection that is as strong as *m*, the node would be isolated after the edge-removal process. Since every node in an *m*-slice, by definition, should have at least one connection with *m* or a higher line value, isolated nodes are removed.

The resulting network after the above two-step removal constitutes the *m*-slice network at the given value of *m*. In a nutshell, an *m* slice is obtained by removing the weakest ties and thereby isolated nodes from the (*m*-1)-slice. By repeating the

above process, increasing *m* by one at each iteration, one can obtain a set of nested sub-networks. The nodes and edges in an *m*-slice are a subset of those in the (*m*-1)-slice and the superset of those in the (*m*+1) slice.

Another product of this procedure is an additional attribute of the nodes in the network, which was named the *m*-index. In the nested structure of *m*-slices, a node may belong to multiple slices depending on the strength of its connections. If a node has one connection of which the strength (line value) is 3, for instance, it will appear in 1-slice, 2-slice, and 3-slice, but not in 4-slice. The *m*-index of a node is defined by the highest *m*-slice to which it belongs. Moving from the lower slices to the higher ones, each node that belongs to an *m*-slice but not to the (*m* + 1)-slice is given the *m*-index value of *m*. The analysis using this *m*-index value will be presented later in section 5.4.2.4.

For the network of *delicious.com* users, we repeated the slicing process from *m*=1 to *m*=600. At *m*=1, two users are connected if they share one or more bookmarks, while at *m*=600, an edge connecting two users indicates they have 600 or more bookmarks in common. By the time *m* reached 600, there remained 8 nodes and 8 edges[30].

Figure 5.9 shows that number of nodes and the number of edges in an *m*-slice of our network up to the 600-slice. Note that there are two different vertical axes: one for the number of nodes which scales up to 22893 and the other for the number of edges which scales up to 17,918,940.

[30] The values for those 8 edges are 634, 782, 817, 819, 834, 1,055, 1,142 and 2,018.

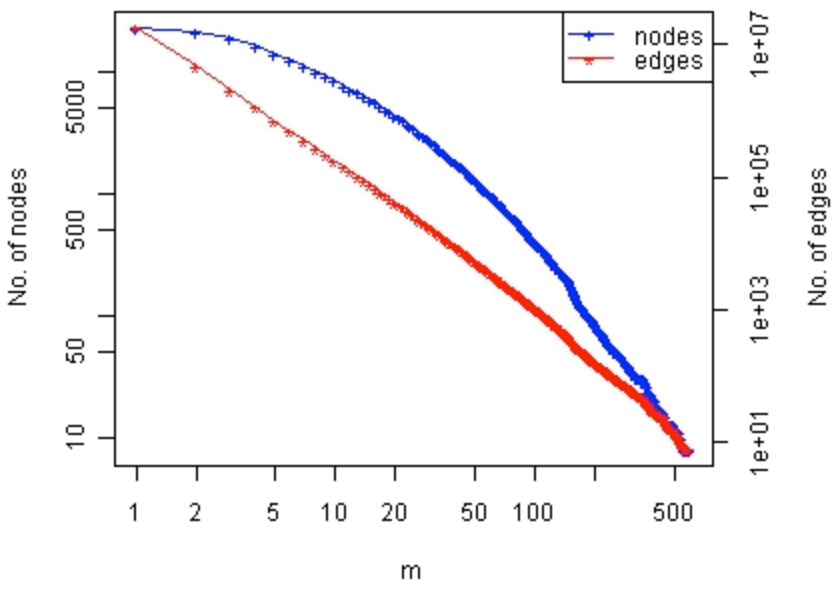


Figure 5.9 Number of nodes and edges in a given *m* slice

While the number of edges decreased exponentially, creating a straight line on the log-log plot, the number of nodes reduced relatively slowly at the beginning making a downward curve. Note that the number of edges removed at each level of *m*, shown in this graph, is the number of edges with the line value, *m*-1.

Table 5.5 shows the number of removed vs. remaining edges and nodes at each *m* value. The last two columns of Table 5.5 show the cumulative percentage of the removed edges with respect to the total number of edges in the original network, and the proportion of remaining nodes at each level of *m* to the total number of nodes in the original network. Notice that at *m*=2, while 74% of the edges were removed, about 92.8% of the nodes were still connected.

Table 5.5 Number of removed vs. remained edges and nodes

m	edges removed	edges remaining	isolated (removed) nodes	connected (remaining) nodes	cum. % of removed edges	% of remaining nodes
1	0	17,918,940	279	22,893	0.00%	98.80%
2	13,267,132	4,651,808	1,393	21,500	74.04%	92.78%
3	2,637,937	2,013,871	2,806	18,694	88.76%	80.67%
4	898,567	1,115,304	2,622	16,072	93.78%	69.36%
5	407,245	708,059	2,072	14,000	96.05%	60.42%
6	218,484	489,575	1,559	12,441	97.27%	53.69%
7	131,142	358,433	1,333	11,108	98.00%	47.94%
8	84,599	273,834	1,077	10,031	98.47%	43.29%
9	57,429	216,405	846	9,185	98.79%	39.64%
10	41,358	175,047	762	8,423	99.02%	36.35%
20	4,768	42,445	266	4,384	99.76%	18.92%
30	1,359	17,724	101	2,731	99.90%	11.79%
40	543	9,437	73	1,837	99.95%	7.93%

5.3.2.3 *m*-slice sub-networks

As discussed above, the *m*-slice technique produces a nested set of subgraphs in a recursive process. As *m* progressively increases, more and more edges and newly isolated nodes are removed from the network. At each step, the *m*-slice represents a sub-network of the original network, where each and every edge has a weight equal to or greater than the given *m* value. In the context of this study, it means that each and every connected pair of users in a given *m*-slice has *m* or more bookmarks in common. Starting from the initial network (the 1-slice [31]) where two users are connected if they have one or more shared bookmarks, we obtained all the nested sub-networks up to the 600-slice. For each sub-network, the same set of network

[31] The 1-slice is also a sub-network of the original network which consists of 23,172 users, because it does not include the 274 isolated users in the original network. By definition, an *m*-slice eliminates any isolated nodes.

properties reported in Section 5.3.1 was measured. Table 5.6 presents the network measures for lower *m*-slices, every 10^{th} slice, and every 100^{th} slice.

The rapid drop-off of the average degree as well as the density at the first few iterations (especially at the 2- and 3-slices) is no surprise at all, given that almost 94% of all the edges have a value of three or less. As discussed in the previous section, however, the reduction of the number of nodes in those sub-networks was not as drastic. We can also see that the distance between users, in terms of both the maximum distance (diameter) and the average distance (average shortest path length), did not increase to a great extent even after the vast majority of the edges were removed. In the 2-slice, after removing 74% of all the edges that had been in the 1-slice, the diameter increased only by 1 while about 93% of nodes remained in the network. Likewise, the increase in the average path length was not substantial. This means that a large portion of the removed edges had provided rather redundant paths among users in this network.

Note that the network density kept falling as *m* increased from 1 to 20, but from after the 30-slice, the network density started to increase again. The distance measures also turned around at the same point. The diameter and average path length reached their maximum at m=20, and started to decrease after that. While the increase in the density may be ascribed to the smaller number of nodes since the density measure tends to be higher in a small network, the shorter distance as well as the higher density shown in the sub-networks of a large *m* value suggest that some of the nodes that are strongly connected (with a high line value) are also tightly knit with one another, forming a core community in this network.

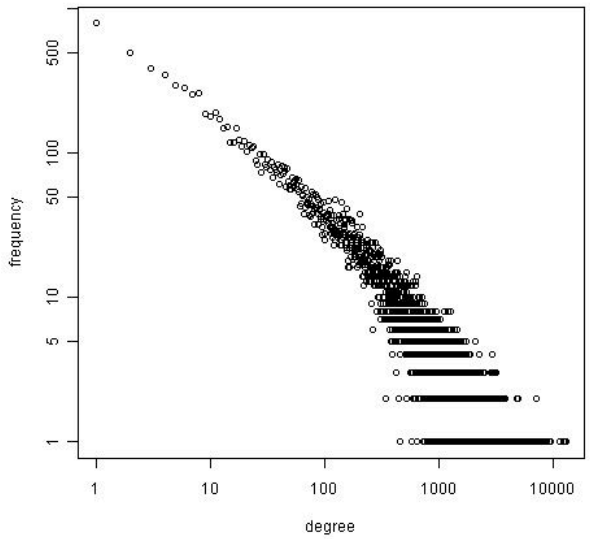
Table 5.6 Network properties of *m*-slice sub-networks

m	No. of nodes	No. of edges	Density	Average degree	Diameter	Average path length	Global clustering coefficient	Avg. node clustering coefficient
1	22,893	17,918,940	0.0684	1,565.45	6	2.01	0.32	0.44
2	21,500	4,651,808	0.0201	432.73	7	2.33	0.27	0.51
3	18,694	2,013,871	0.0115	215.46	8	2.48	0.24	0.61
4	16,072	1,115,304	0.0086	138.79	9	2.54	0.22	0.68
5	14,000	708,059	0.0072	101.15	10	2.58	0.21	0.73
6	12,441	489,575	0.0063	78.70	10	2.62	0.19	0.76
7	11,108	358,433	0.0058	64.54	10	2.65	0.18	0.77
8	10,031	273,834	0.0054	54.60	11	2.67	0.17	0.79
9	9,185	216,405	0.0051	47.12	11	2.69	0.16	0.80
10	8,423	175,047	0.0049	41.56	11	2.72	0.15	0.80
20	4,384	42,445	0.0044	19.36	13	2.99	0.11	0.84
30	2,731	17,724	0.0048	12.98	10	2.58	0.09	0.85
40	1,837	9,437	0.0056	10.27	7	2.50	0.09	0.86
50	1,337	5,645	0.0063	8.44	9	2.50	0.08	0.87
60	988	3,702	0.0076	7.49	6	2.47	0.08	0.87
70	770	2,552	0.0086	6.63	6	2.47	0.09	0.87
80	622	1,900	0.0098	6.11	7	2.51	0.09	0.86
90	501	1,425	0.0114	5.69	7	2.51	0.10	0.86
100	408	1,128	0.0136	5.53	7	2.56	0.10	0.83
200	86	167	0.0457	3.88	6	2.47	0.23	0.75
300	37	67	0.1006	3.62	3	1.77	0.37	0.79
400	20	31	0.1632	3.10	3	1.81	0.45	0.76
500	12	15	0.2273	2.50	3	1.60	0.50	0.76
600	8	8	0.2857	2.00	1	1.00	1.00	1.00

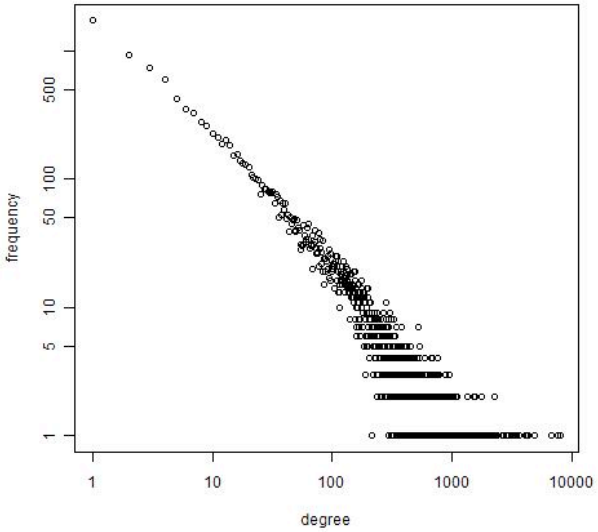
Another interesting pattern that shows up as *m* increases is that the discrepancy between the global clustering coefficient and the average of the local clustering coefficients gets larger. A possible explanation for this can be found in the variations in node degrees. Newman (2003) explained that lower degree nodes tend to have higher local clustering coefficients. Since the local clustering coefficient of a node is defined as the ratio of the actual connections to the possible connections among neighbors of the node, the measure is sensitive to the node degree. Whereas

the denominator of the equation increases proportionally to the degree of the given node, it is not likely in general that the numerator, the actual observed connections among its neighbors, increases at a similar rate. As Newman (2003) pointed out, by averaging local coefficients over all the nodes regardless of their respective degrees, the average clustering coefficient obtained for the whole network tends to give more weight to lower degree nodes. In a network where the majority of nodes have a small number of neighbors, the effect may be large. The degree distributions of a selection of *m*-slices, shown below, confirm that, in each of those sub-networks, the nodes with few neighbors occupy a very large portion of the network. In Section 5.3.1.4 we also observed that there is a negative correlation between the node degree and the local clustering coefficient, meaning that nodes with lower degrees tend to have higher local clustering than those with higher degrees. Therefore, we may conjecture that the large number of nodes with lower degrees and relatively higher local clustering coefficients brought about the observed discrepancy between the two clustering measures.

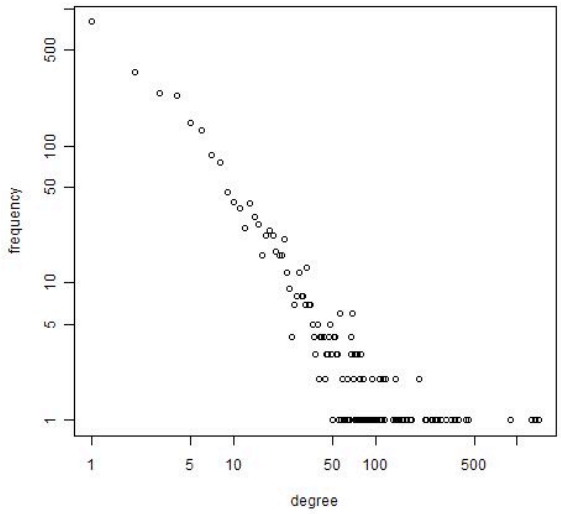
Figure 5.10 shows the degree distribution of 2-slice, 5-slice, 30-slice, and 50-slice sub-networks, respectively. Each distribution shows the number of nodes in that sub-network that has a given degree.



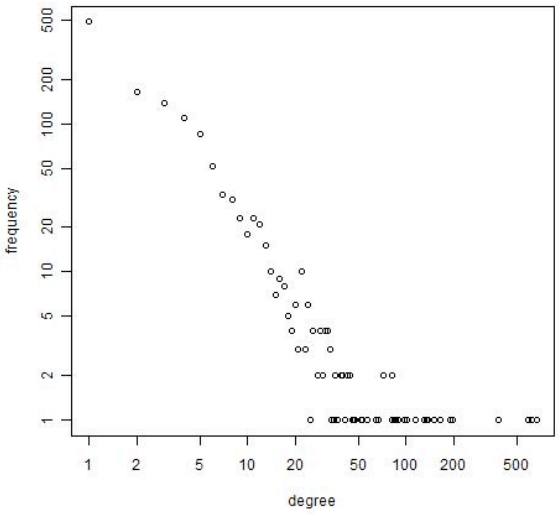
a) 2-slice



b) 5-slice



c) 30-slice



d) 50-slice

Figure 5.10 Degree distributions of different *m*-slices

Compared with the degree distribution of the original network shown in Figure 5.4 (in Section 5.3.1.2), the degree distribution of the 2-slice sub-network is clearly

different in that the proportion of nodes with a small number of neighbors increased drastically while the thick middle layer consisting of the nodes with a modestly high degree that had been in the original network was trimmed out substantially. In the 5-slice sub-network, the distribution has an even larger number of nodes with degree 1 while the total number of nodes was reduced. Unlike the degree distribution of the original network (which was far from a power-law distribution), the degree distribution of the 5-slice sub-network appears to be closer to a power-law distribution, indicated by the nearly straight line on the log-log plot, especially in the lower end where the nodes with a small degree were plotted. The distribution of 30-slice and 50-slice sub-networks shows that, while the size of the network is significantly smaller, the degree distribution more or less sustained the characteristic pattern, with the majority of sparsely connected nodes and a small number of extremely well connected nodes

5.4.2.4 *m*-index

As a result of the *m*-slice procedure, each node in the original network is given an index value that we call *m*-index. In his introduction to *k*-core techniques, Seidman (1983) introduced the concept of '*k*-remainder'. Due to the nested nature of *k*-cores, the nodes at a given level of *k* can be divided into two sets: those that move up to (*k* + 1) core and those that would 'remain' at the given level without belonging to any higher *k*-core. He argued that the sequence of these complement sets, which he called the 'core collapse sequence,' also disclose the internal structure of the network. Introducing the idea of core collapse sequence, Scott (2000) suggested the idea can

be applied to the *m*-core (*m*-slice) technique to depict the overall 'texture' of the network.

During the procedure of progressively obtaining higher *m*-slices, we also assigned *m*-index values to corresponding nodes incrementally, such that a node has an index value *m*, if it belongs to *m*-slice but not to (*m*+1)-slice. At the beginning, the 274 isolated nodes in the original network were given the index value 0, since they are not even included in the 1-slice. At *m*=1, therefore, there were 22,893 nodes excluding the isolated nodes. When *m* is increased to 2, 13,267,132 edges with the line value of 1 were removed, and as the result of the removal of edges, 1,393 nodes became isolated and therefore removed. The *m*-index value of 1 is assigned to these 1,393 nodes that were removed at *m*=2. Similarly, 2,806 nodes were removed at *m*=3 and were given the *m*-index value 2. As *m* increases, more and more nodes were removed and the corresponding *m*-index values were assigned. At *m*=600, where we stopped the procedure, there were 8 nodes left. Their *m*-index values were manually calculated so that every node in the original network has an *m*-index.

Figure 5.11 shows the distribution of *m*-index values. Each point in this plot represents the number of nodes that have a given *m*-index value. For instance, the point at the left most position represents the 1,393 nodes that have the *m*-index of 1.

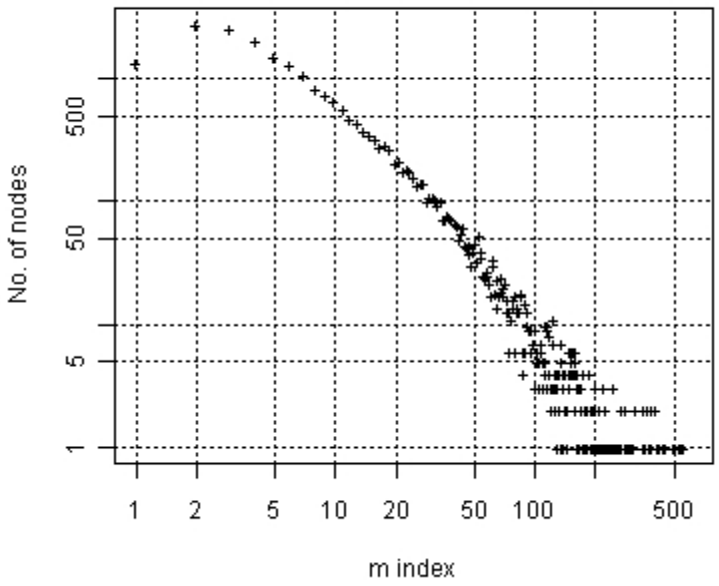


Figure 5.11 *m*-index distribution

The *m*-index value of a node represents that highest *m*-slice to which it belongs. Note that the highest *m*-slice that a node can be a part of is determined by the strongest connection that the node has. As long as a node has at least one connection that is as strong as a given value *m*, the node would appear in the *m*-slice. Note also that the strength of a connection between two users in this network is defined by the number of shared information objects. Take the above example of the 1,393 users (nodes) with *m*-index of 1. The fact that they were removed from the network when *m* is increased to 2 means that they had only weak connections, regardless of the number of connections, with other users. They might have connections with many other users in the network, but they don't have two or more common information objects with any of them. Figure 5.12 shows the *m*-index and degree *k* of each node in this network. As shown in Figure 5.12, there exists a large variation in degrees among the nodes with the same *m*-index value.

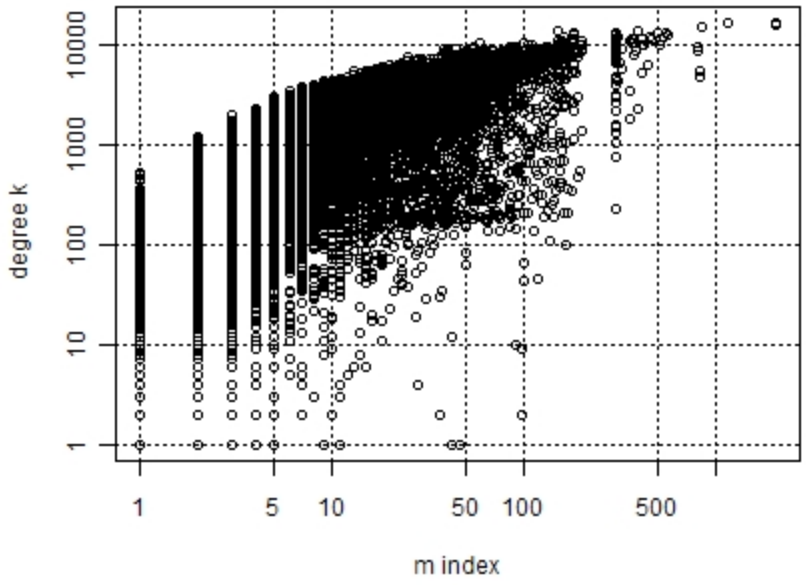


Figure 5.12 *m*-index versus degree of the network nodes

5.3.2.5 Components in *m*-slices

The last part of the analysis in the second phase is concerned with the question of communities of shared interests within the network. The *m*-slice technique provides a way to look at the community structure of a given network, since in essence it identifies strongly connected regions in the network. As shown in section 5.3.2.3, a sub-network of the initial network is created at each *m*. While the initial network was constructed by connecting two users who have one or more bookmarks in common, in an *m*-slice sub-network two users are connected if they have at least *m* bookmarks in common. Therefore, this technique allows us find partitions of the network based on the similarity of interests defined by the number of shared bookmarks.

An *m*-slice is not necessarily a connected network. It is usually the case that an *m*-slice contains multiple components of varying sizes. As the slicing process goes on, a large component in a previous step (at a lower value of *m*) may break into smaller components. Components that emerge at any point in this divisive clustering procedure can be regarded as subgroups or communities of varying cohesiveness. In other words, a component in an *m*-slice can be taken as a community within which nodes are connected by the minimum strength of *m*. The *m*-slice technique provides a straightforward method of finding stronger segments of the network by iteratively removing the weakest ties.

As reported in section 5.3.1.5, the network of *delicious.com* users, when constructed based on one or more shared bookmarks, has only four components of size two or more. The network is virtually a single piece, because not only is the number of components small but also all the three components other than the giant component consist of only two nodes. In order to see how the network splits off as *m* increases, we looked at both the number and the size distribution of components.

Figure 5.13 shows the number of components of size two or more in the *m*-slice of a given *m*. The number of components was very small at first, but it grew quickly, reaching the maximum of 105 at *m*=10. Right after that the number drops to 89. Until *m*=34, the number of components stayed within the range of 87 and 96. The number started to fall down at *m*=35, mainly due to the decreasing size of the network and the increasing proportion of isolated nodes.

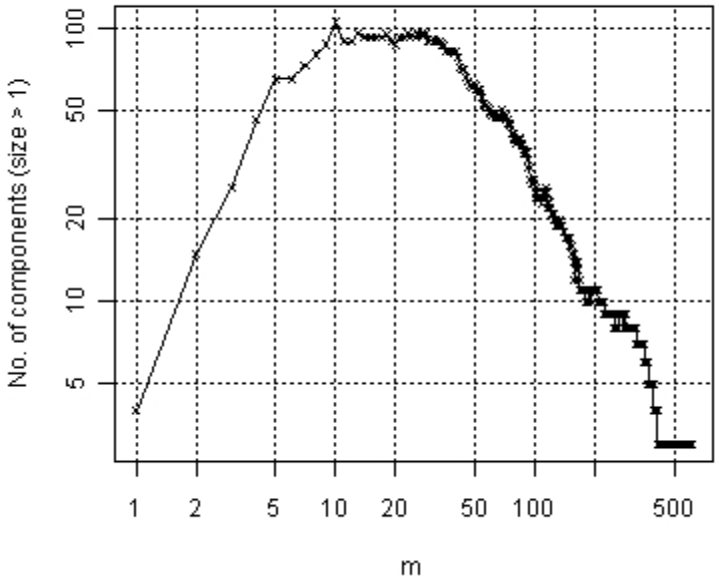


Figure 5.13 The number of components of size two or more.

Figure 5.14 plots the size of the giant component (the number of nodes belonging to the component) and the size of the second largest component for different m values from 1 up to 600. Note that there are two separate y axes in this figure because the scales of the two components were different. From Figure 5.14, we can see that, while the size of the giant component decreased steadily, the second largest component remained small even after 20 iterations. Considering Figure 5.14 in conjunction with Figure 5.13, which shows the steep increase in the number of components of size two or more, it means that the components taken apart from the giant component were mostly pairs or clusters of only a few nodes. The size of the second largest component first exceeded ten at m=15, and stayed between 10 and 20 for a while. The size finally jumped to 248 at m=28, and it is the first major break-down of the giant component into segments of substantial sizes.

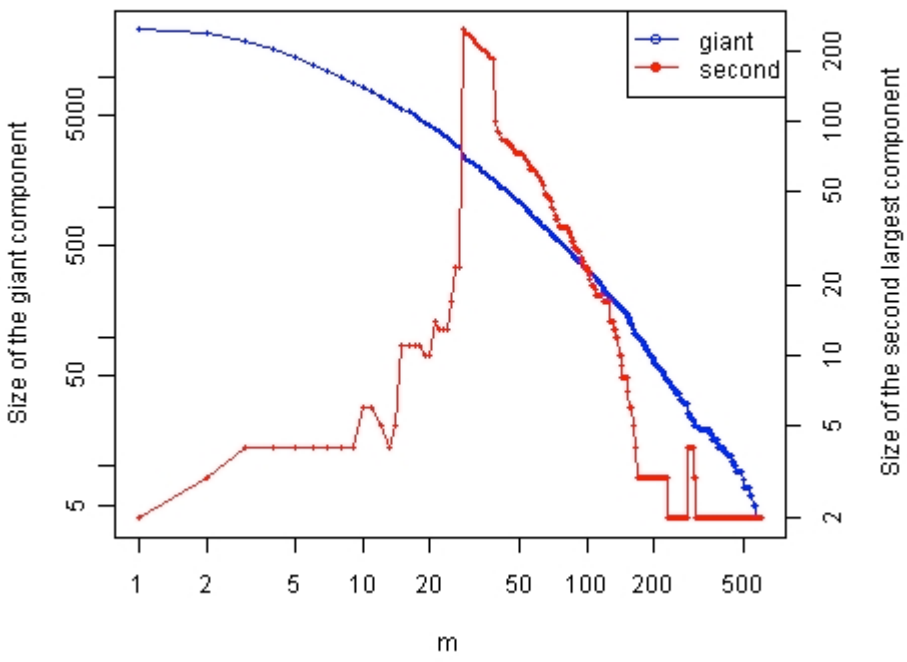


Figure 5.14 Sizes of the giant component and the second largest component

Chapter 6. Results of Phase III

In the second phase, the *m*-slice technique was introduced as a method for scrutinizing the internal structure of a network based on successive levels of tie strength. As explained in the previous section, with the nested sub-networks at varying levels of *m*, the *m*-slice technique also allows an examination of the community structure that may exist in the network. Overall, the results suggested that there exists a tightly connected core in this network and that the network does not break down easily. As the threshold value of the minimum tie strength (*m*) was increased, peripheral nodes were taken off from the network, either becoming isolated nodes or forming small groups consisting of a few nodes. However, only a few groups of nodes, other than the giant component, that could be considered as separate communities emerged early in the *m*-slicing process. More specifically, the second largest component remained small, and a component containing more than 1% of the remaining nodes first appeared when *m* was raised to the high value of 28. In this section, we take the 28-slice of the network and examine the communities (the components) identified in that analysis.

At *m*=28, the number of remaining nodes (excluding isolated notes) is 2,975. The sizes of the giant component, the second largest component, and the third largest components are 2,458, 248, and 24, respectively. The giant component still contains 82.6% of the remaining nodes, but the second largest component accounts for 8.3%

of the network[32]. In the following, each of the three largest components in the 28-slice network will be examined to see whether each indeed constitutes a distinct community of shared interests. Figures 6.1, 6.2, and 6.3 show the three components in order.

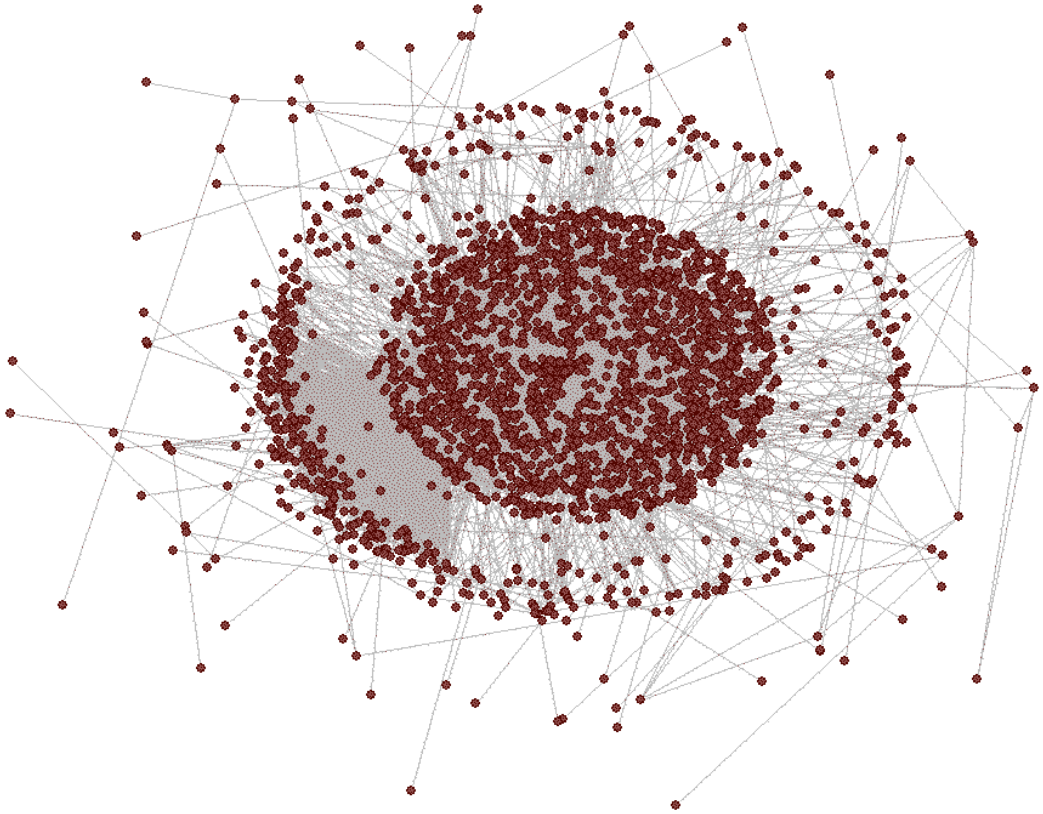


Figure 6.1 The largest (giant) component of the 28-slice sub-network.

[32] For a comparison, the proportions of the giant component and the second largest component in the 27-slice, the network produced immediately before the 28-slice, were 91.3% and 0.8%.

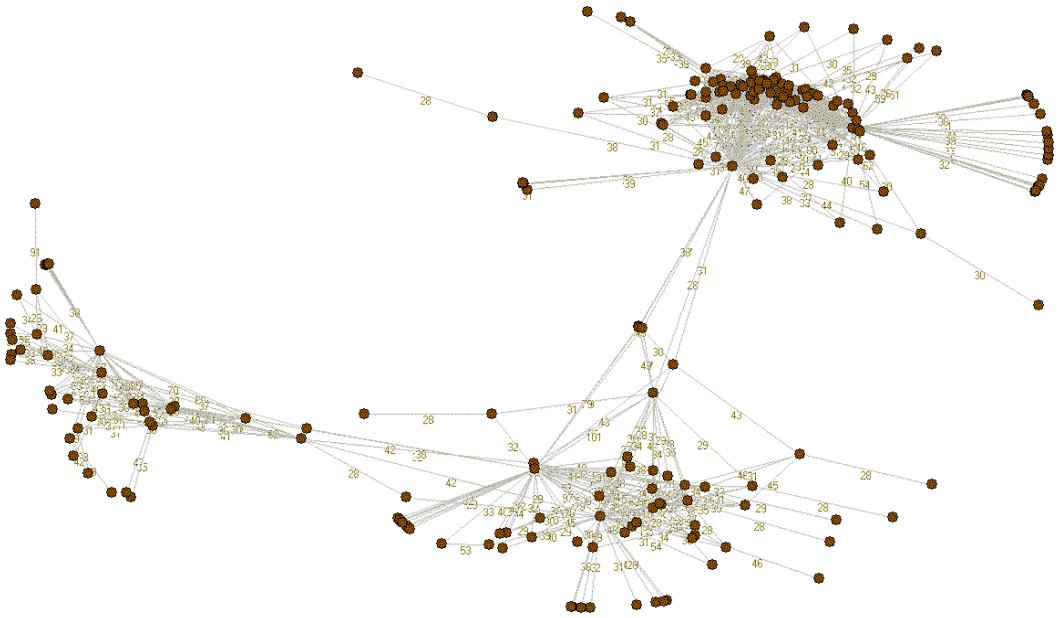


Figure 6.2 The second largest component of the 28-slice sub-network.

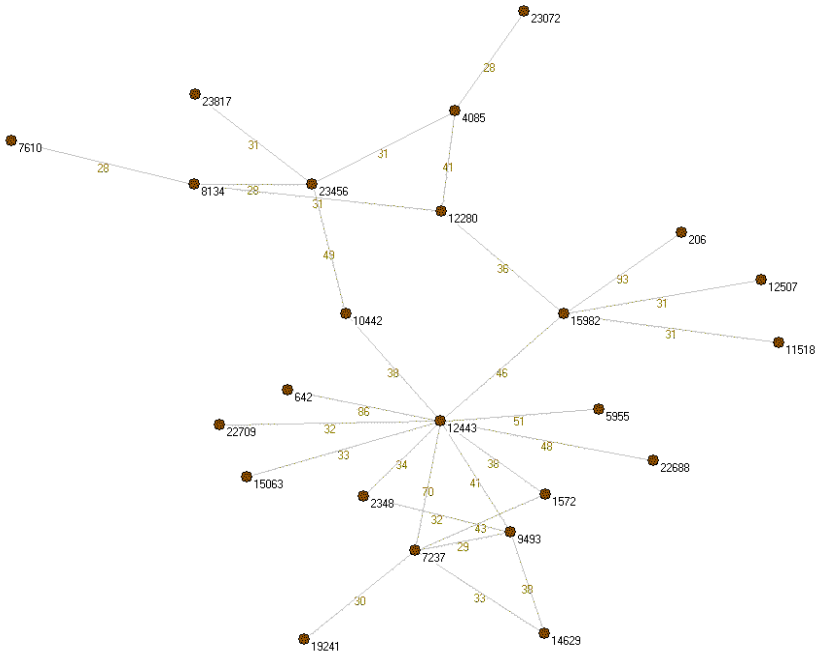


Figure 6.3 The third largest component of the 28-slice sub-network.

Note that, in the network analytic framework, communities or cohesive subgroups are identified solely based on structural characteristics, regardless of the specifics of the technique being used. The basic assumption is that the patterns of connections among users, whether in the volume of connections or in the intensity of connections, would reveal how individuals are clustered within a social structure. Whether the members of a group identified based on structural features indeed share some properties or attributes of interest (e.g. race, income, religion), other than the structural ones, is often treated as a separate question. With the communities identified using the *m*-slice technique, we now try to move on to the next question of whether these communities represent communities of shared interests.

In order to examine the communities, we leverage the duality of the affiliation network. Recall that, when we constructed the network of users by transforming an affiliation network of information objects and users, the links in the induced network were created by the information objects shared by each pair of users. When two users share more than one information object, multiple lines are created between those two users, which are then represented as a line value (weight) when multiple lines are merged into one at the end of the transformation process. While the *m*-slice technique used the multiplicity of the shared information objects involved in making connections between users, in this section we trace each link back to those information objects. In other words, when examining a specific component, we gather the actual information objects (shared bookmarks) creating each connection, for all the links within the given component.

Note again that the *m*-slice technique only takes into account the number of

shared bookmarks. If A and B are connected and A and C are connected in the 28-slice, all it means is that A and B have 28 or more shared information objects, regardless of what those objects are, and A and C have 28 or more shared information objects. The connection between A and B and the connection between A and C may be comprised of totally different sets of information objects. That is, there may or may not be any overlap or similarity between the two sets of information objects that created the A-B connection and the A-C connection, respectively. The fact that each pair has a certain number of bookmarks in common does not necessarily mean that the connected component as a whole has a coherent set of interests. Therefore, one way of looking at the actual degree of shared interests within the group as a whole is to examine the entire list of shared information objects with their frequencies and see whether the union set represents any coherent theme and, if so, to what extent. More specifically, the entire list of shared information objects delineates the range of shared interests in the community, as well as their characteristics; the frequency distribution of the information objects depicts the degree of overlap among the shared interests, that is, the extent to which different pairs of users within the group converge on particular information objects.

For each of the three largest components (with sizes of 2,458, 248, and 24, respectively), the list of distinct information objects along with a number of statistics and measures was constructed. First, for each connected pair within the given component (community), we obtained the list of information objects involved in the connection (link). Since the components are identified in the 28-slice, there are 28 or more information objects corresponding to each connection. After compiling all the

information objects for every connected pair within the component, a list of distinct information objects with their frequencies was constructed.

Second, two different indexes were defined and calculated for each information object in the list and for the entire list. One index, named the contribution index, measures the extent to which each information object contributes to the cohesiveness as well as the connectivity of the community. The contribution to connectivity was measured by the number of connections within the community that the given information object was involved in creating, that is, its frequency of the occurrence in the connections within the community. The contribution to the cohesiveness of the community, on the other hand, needs to be considered because a generic and popular information object may create a large volume of connections but not only within the community but also outside the community. Intuitively, the more connections an information object makes within the community relative to those outside the community, in the more it contributes to defining the distinct characteristics of the community. In order to account for this, the proportion of connections within the community was calculated, in relation to all the connections of which an information object is a part in the entire original network. A weight for each information object, then, was given by multiplying its frequency within the community (connectivity measure) by the within-community proportion (cohesiveness measure). The list of information objects for a community was sorted in the descending order of this measure. Tables 6.2, 6.3, and 6.4 show the top 15 information objects (ranked by the contribution index) from the first, second, and the third community, respectively. The index values per se are not included in the Tables.

The second index, named the aggregation index, also considers statistics with regard to an information object both within the specific community and in the entire original network. For this index, however, the number of users is considered, instead of the number of connections. For each information object in the list, we checked the popularity of the information object (the number of users who bookmarked the information object) both in the entire dataset and in the given community. Next, the proportion of users who were included in the community, among all those who were interested in the given information object, was calculated. While the first index (the contribution index) described above represents the importance of individual information objects for binding the community together, this index (the aggregation index) shows the capacity of the community to bring together the users who shared an interest in those information objects. By the nature of the original network being transformed from the affiliation network of users and information objects, all the users who bookmarked a particular information objects were to be connected to form a completely connected clique in the original network. As the value of *m* (the minimum number of shared bookmarks) increases, however, the clique formed by the single information object loses connections, more or less so depending on whether the users have additional shared information objects. If, for example, eight users who share an information object are all still connected when *m* increases to two, it means that the each of the eight users has at least one additional bookmark shared with other user(s) in the group. Therefore, the proportion of users who bookmarked an information object and are connected at a given *m*-slice shows the degree of shared interests of the users *beyond* the particular information object. Note that, although the

index was calculated per information object, it does not represent the capacity of that information object per se, but a collective strength of a set of information objects -- the intersection of circles of shared interests to which the information object belongs. The average value of this index, therefore, can serve as an overall indicator of the cohesiveness of user interests within the community. In Tables 6.2, 6.3, and 6.4, the aggregation index for each top-ranked information object is shown in the fifth column. In addition, the average of this index for each community is shown in the last row of Table 6.1.

Finally, the URL and the title of each information object was added to the list, in order to assess the similarity of the information objects themselves. Since this phase of the study was designed as a preliminary exploration of the communities, only the URLs and titles were considered without further crawling or inspection of the actual documents. As will be described below, a basic text analysis was conducted on titles in order to extract high-frequency title words that appeared in the set.

Table 6.1 shows some basic statistics with regard to the compiled list of information objects for each component.

Table 6.1 Statistics related to information objects connecting community members

	Component 1	Component 2	Component 3
No. of users	2,458	248	24
No. of links	19,115	1,306	29
No. of distinct URLs	51,847	6,414	831
Average of aggregation index	73.1%	86%	70.3%

Tables 6.2, 6.3, and 6.4 show the top ranked information objects in the first (giant), the second, and the third components, respectively[33]. As can be seen in these tables, each component revealed distinctive characteristics with a clear theme. While each table only shows the top 15 information objects in each community, it should be mentioned that not only the top ranked information objects shown in the tables but the majority of the information objects in each community shared the central theme of the community. Not surprisingly, the first community turned out to be about web design and/or web development. Considering that social bookmarking and *delicious.com* first gained attention and popularity in the domain of information architecture, it seems natural that the largest community in *delicious.com* is centered around the topics closely related to web design and/or development. The second

[33] As mentioned above, the fifth column of the table is the aggregation index of the information object. The second column shows the number of connections in which the information object is involved (as one of the 28 or more information objects creating each link). The third column shows the total number of users who bookmarked the information object in the original network, and the fourth column shows the number of users who are included in the community among those users shown in the third column. Note that the contribution index is calculated using the values in the second and the third columns, and the aggregation index is calculated using the third and the forth columns.

community is in a sense more interesting. The information objects are pieces of fiction, usually posted on a blog, by online amateur writers. Moreover, the majority of them fall into the category of *fan fiction*. Tushnet(1997) provides the following definition of fan fiction: "'Fan fiction,' broadly speaking, is any kind of written creativity that is based on an identifiable segment of popular culture, such as a television show, and is not produced as 'professional' writing." (p.655).It is fiction "where fans create stories using characters, settings, and events from their favorite books, movies, or television shows." (Burns & Webber, 2009, p.27). In fact, due to its distinctive characteristics such as community-driven progressive creation of works, fan fiction has been an area of interests in media studies. Note that all the top ranked URLs are on LiveJournal.com where, according to Hellekson and Busse (2006), online fan communities have flourished. The third community, although small, also has a clear theme: recipes for baking. Since the number of information objects connecting the users of this community is relatively small (831), it was possible to read through all the titles one by one. Interestingly, without a single exception, all of the 831 URLs are related to cooking and/or baking. Moreover, the vast majority of them are about baking pies, cakes, cookies, and so on.

Since the large number of information objects made manual examination of the topics in the first and second communities infeasible, a rudimentary text analysis of the titles of the information objects was conducted. Taking only the titles of the information objects in the given community, word frequency was calculated after removing special characters and a small set of stop words. From the list of words in descending order of frequencies, the most frequently-occurring noun or noun-

equivalent terms were selected to create the list of terms for each component. Tables 6.5 and 6.6 show the high frequency terms found in titles of the information objects belonging to the first and the second communities, respectively. In Table 6.5, technology-related terms assume a large portion of the frequent terms. It can be seen, again, that web design and development constitute the central theme: the term CSS (standing for Cascading Style Sheet) is ranked very high, following some generic terms like blog, web, design, or google. Terms like development and programming are also ranked high, along with terms referring to specific programming languages or development environments, such as ruby, rails, javascript, python, and ajax. One interesting finding is that 'japan' is in 12th place and 'japanese' in 25th place. In addition, 'itmedia' referring to a specific Japanese website is in 15th place. It seems that a considerable number of users in this community bookmarked information objects that reside in Japanese websites and are written in Japanese.

As for the second community, the title terms shown in Table 6.6 reveal a clearer pattern, although it may not be obvious at first glance. As discussed above, this community consists of users who are interested in fiction, especially fan fiction written by online amateur writers. The most frequent term (string) 'fic' appeared 1,208 times, which is almost 3.5 times higher than the second one on the list. In addition, other terms like 'ficlet', 'flashfic', 'fiction', 'fanfic', and 'fanfiction' are also ranked high. In total, there are around 1,500 occurrences of the string 'fic' either as a single term or as a substring in other terms. Considering that the total number of information objects connecting users in this community is only 6,414, it is a remarkably high frequency. Overall, the analysis of title words also confirms that user

interests in each community indeed centered around a coherent theme.

Table 6.2 Top ranked information objects in the first community (the giant component)

No	No of links	All users	Comm users	Aggregation index	URL	Title
1	1,319	147	117	79.6%	http://speckyboy.com/2008/03/28/top-12-css-frameworks-and-how-to-understand-them/	Top 12 CSS Frameworks and How to Understand Them | Speckyboy - Wordpress and Design
2	684	78	65	83.3%	http://www.ironmyers.com/layouts/750_pixel_Layouts/index.html	750 pixel Pure CSS Layouts - Iron Myers
3	944	110	89	80.9%	http://www.sitepoint.com/article/tomorrows-css-today	Tomorrow's CSS Today: 8 Techniques They Don't Want You To Know [CSS Tutorials]
4	1,445	173	136	78.6%	http://www.smashingmagazine.com/2008/04/17/web-form-design-modern-solutions-and-creative-ideas/	Web Form Design: Modern Solutions and Creative Ideas | Design Showcase | Smashing Magazine
5	451	55	48	87.3%	http://webtecker.com/2008/03/17/list-of-ajax-form-validators/	List of Ajax form Validators | WebTecker the latest tech, web resources and news.
6	1,606	196	151	77.0%	http://www.noupe.com/ajax/37-more-shocking-jquery-plugins.html	37 More Shocking jQuery Plugins
7	1,436	176	128	72.7%	http://www.noupe.com/javascript/37-great-ajax-css-tab-based-interfaces.html	37+ Great Ajax, CSS Tab-Based Interfaces
8	472	60	51	85.0%	http://webtecker.com/2008/04/14/programing-cheat-sheets/	Programing Cheat Sheets | WebTecker the latest Web Tech, Resources and News.
9	1243	159	120	75.5%	http://www.vivalogo.com/vl-resources/beautiful-javascript-flash-galleries.htm	33 Most Beautiful Javascript and Flash Galleries | Vivalogo Resources
10	442	57	48	84.2%	http://ntt.cc/2008/02/13/the-most-complete-ajax-framework-and-javascript-libraries-list.html	The Most Complete AJAX Framework and JavaScript Libraries List(124+) - Ntt.cc
11	822	106	82	77.4%	http://vandelaydesign.com/blog/design/photoshop-text-tutorials/	50 Essential Photoshop Text Tutorials | Vandelay Website Design
12	577	75	58	77.3%	http://www.bgoncalves.com/notes/2008/04/20/30-free-online-books/	Bruno Goncalves - 30+ Free Online Books
13	769	100	72	72.0%	http://www.webcredible.co.uk/user-friendly-resources/css/planning-stylesheet.shtml	Planning your stylesheet - the definitive guide
14	740	101	85	84.2%	http://www.labnol.org/internet/design/completely-test-website-errors-html-standards/2673/	How to Completely Test Your Website
15	849	116	93	80.2%	http://vandelaydesign.com/blog/design/resources-grid-based-design/	65 Resources for Grid-Based Design | Vandelay Website Design

Table 6.3 Top ranked information objects in the second community

No	No. of links	All users	Comm users	Aggregation index	URL	Title
1	306	36	36	100.0%	http://foxxcub.livejournal.com/455452.html	foxxcub: Fic: Jon and Spencer Make a Porno
2	231	30	29	96.7%	http://naotalba.livejournal.com/27521.html	naotalba: Bandslash: Telepathy Means Never Having Privacy When You're Jerking Off
3	285	40	37	92.5%	http://afterthefair.livejournal.com/175629.html	Well-Mannered Frivolity - Fic: Kick It Back, Jon/Spencer, R
4	210	30	29	96.7%	http://foxxcub.livejournal.com/460955.html	foxxcub: Fic: keep it in your back pocket
5	377	55	47	85.5%	http://sevenfists.livejournal.com/173275.html	sevenfists: fic: In the Sirocco, 1/2
6	334	49	42	85.7%	http://zarah5.livejournal.com/132620.html	zarah5: Fic: Agent Provocateur (P!ATD, Brendon/Ryan, Jon/Spencer; R) (1/3)
7	182	27	26	96.3%	http://community.livejournal.com/ficonastick/28753.html	ficonastick: PatD: Go With The Flow (Jon/Ryan)
8	266	40	36	90.0%	http://provetheworst.livejournal.com/440594.html	provetheworst: [fic] A Time and a Place [brendon/spencer; nc-17]
9	304	46	43	93.5%	http://enoughoflove.livejournal.com/252889.html	enoughoflove: Fic - Nothing Quite Like
10	180	28	28	100.0%	http://beingothrwrldly.livejournal.com/573044.html	the way i see it #282
11	256	40	36	90.0%	http://airgiodslv.livejournal.com/382901.html	airgiodslv: Oistros
12	430	67	59	88.1%	http://ignipes.livejournal.com/313344.html	ignipes: P!atD Fic: Dance Upon the Waves, 1/2 (Ryan/Brendon, PG-13)
13	381	60	51	85.0%	http://ignipes.livejournal.com/312652.html	ignipes: P!atD Fic: Manifesto (implied GSF, PG-13, 3420 words)
14	304	48	40	83.3%	http://just-katarin.livejournal.com/113137.html	Cool as the Fonz and deadly as Charles Bronson - the last good thing about this part of town | Fic | Bandom- Fall Out Boy part 1/2
15	163	27	26	96.3%	http://sathinks.livejournal.com/371052.html	Kiss Ninja by SA [Panic at the Disco, Brendon/Jon, 1/1, PG]

Table 6.4 Top ranked information objects in the third community

No	No. of links	All users	Comm users	Aggregation index	URL	Title
1	7	7	6	85.7%	http://www.culinaryconcoctionsbypeabody.com/2008/04/16/enjoy-the-now/	Culinary Concoctions by Peabody - Enjoy the now…
2	7	7	6	85.7%	http://alpineberry.blogspot.com/2008/04/candy-in-pie.html	alpineberry: Candy? In Pie?
3	5	6	5	83.3%	http://www.culinaryconcoctionsbypeabody.com/2008/04/09/7-years-strong/	Culinary Concoctions by Peabody - Lemon Berry Gratin
4	4	5	5	100.0%	http://www.culinaryconcoctionsbypeabody.com/2008/02/07/im-a-rat/	Sesame Almond Cookies
5	4	5	5	100.0%	http://alpineberry.blogspot.com/2008/02/i-heart-nutella.html	alpineberry: I Heart Nutella
6	4	5	4	80.0%	http://creampuffsinvenice.ca/2008/02/18/pretty-in-marble/	Pretty in Marble
7	4	5	4	80.0%	http://www.culinaryconcoctionsbypeabody.com/2008/04/01/gooey-tuesday-with-dorie/	Gooey Chocolate Cake
8	4	5	4	80.0%	http://smittenkitchen.com/2006/12/mounds-of-awesome/	Hazelnut Truffles
9	3	4	4	100.0%	http://alpineberry.blogspot.com/2007/02/nutella-cheesecake-brownies.html	alpineberry: Nutella Cheesecake Brownies
10	3	4	4	100.0%	http://novice-baker.blogspot.com/2008/03/cleaning-freezer-part-1-banana-cake.html	Fresh from the Oven: Cleaning the Freezer Part 1: Banana Cake with Caramel Espresso Frosting
11	3	4	4	100.0%	http://dessertfirst.typepad.com/dessert_first/2008/02/one-thing-about.html	Dessert First: Another New Year to Celebrate
12	3	4	4	100.0%	http://alpineberry.blogspot.com/2008/03/big-bag-of-brown-sugar.html	alpineberry: A Big Bag of Brown Sugar
13	3	4	4	100.0%	http://www.jasonandshawnda.com/foodiebride/?p=797	Confections of a Foodie Bride > Blog Archive > The secret lies with Charlotte
14	3	4	3	75.0%	http://www.culinaryconcoctionsbypeabody.com/2008/03/13/kiss-me-im-scottish-canadian/	Chocolate Stout Creme "Brew"lee
15	3	4	3	75.0%	http://cookandeat.com/2008/04/05/muffin-mixed-berries/	Cook & Eat > Blog Archive > Muffin Mixed Berries

Table 6.5 Frequent title words in the first community

Rank	Freq.	Term	Rank	Freq.	Term	Rank	Freq.	Term	Rank	Freq.	Term
1	2,777	blog	26	355	tips	51	270	flickr	76	216	mac
2	1,984	web	27	353	magazine	52	268	python	77	215	techcrunch
3	1,404	news	28	347	tools	53	268	bbc	78	211	gigazine
4	1,197	design	29	346	rails	54	265	data	79	210	linux
5	1,195	google	30	325	business	55	264	future	80	203	flash
6	778	home	31	317	time	56	261	yahoo!	81	203	education
7	741	archive	32	317	gizmodo	57	259	learning	82	202	language
8	582	css	33	316	wordpress	58	257	site	83	197	people
9	572	video	34	316	page	59	255	readwriteweb	84	196	videos
10	554	search	35	314	art	60	254	games	85	195	computer
11	546	world	36	313	music	61	253	resources	86	191	jpeg
12	544	japan	37	303	guide	62	252	ways	87	189	sites
13	526	twitter	38	301	vision	63	251	website	88	189	facebook
14	500	software	39	299	development	64	248	science	89	188	information
15	495	itmedia	40	294	source	65	244	microsoft	90	187	photos
16	470	internet	41	293	tutorials	66	240	network	91	187	links
17	440	media	42	287	tech	67	240	mobile	92	187	library
18	432	code	43	287	list	68	236	photo	93	187	game
19	420	youtube	44	282	ruby	69	232	wiki	94	185	book
20	414	image	45	281	pixels	70	228	jquery	95	179	day
21	413	life	46	281	javascript	71	227	engine	96	178	community
22	399	technology	47	280	firefox	72	225	programming	97	175	beta
23	385	cnet	48	278	part	73	225	health	98	173	php
24	384	windows	49	274	photoshop	74	223	things	99	172	apps
25	362	japanese	50	273	project	75	218	ajax	100	171	marketing

Table 6.6 Frequent title words in the second community

Rank	Freq	Term	Rank	Freq	Term	Rank	Freq	Term	Rank	Freq	Term
1	1,208	fic	26	74	mckay	51	42	atlantis	76	27	pants
2	349	nc-17	27	71	ficlet	52	38	jack	77	26	stereomer
3	302	sga	28	70	flashfic	53	38	ianto	78	26	need
4	166	dean	29	62	gen	54	38	adult	79	26	giddygeek
5	162	sam	30	59	gerard	55	37	something	80	26	days
6	154	r	31	58	life	56	37	fiction	81	25	ways
7	138	pg-13	32	57	frank	57	36	stargate	82	25	fanfic
8	127	brendon	33	56	thing	58	36	sardonicsmiley	83	24	sevenfists
9	122	bandom	34	56	story	59	36	heart	84	24	moon
10	115	panic	35	56	ryan	60	35	everything	85	24	mind
11	114	spencer	36	56	jared	61	34	boys	86	24	lamardeuse
12	107	john	37	55	j2	62	33	sex	87	23	match
13	105	rodney	38	54	jon	63	33	remix	88	22	stars
14	99	spn	39	54	fob	64	32	song	89	22	slashatthedisco
15	98	words	40	54	day	65	32	sky	90	22	skoosiepants
16	96	pg	41	52	slash	66	31	question	91	22	road
17	86	pete	42	48	title	67	31	nothing	92	22	primer
18	86	patrick	43	48	boy	68	31	art	93	22	meme
19	86	mcr	44	46	torchwood	69	30	disco	94	22	list
20	83	way	45	46	post	70	29	sheafrotherdon	95	22	impertinence
21	82	bandslash	46	45	night	71	29	patd	96	22	girl
22	79	jensen	47	44	times	72	29	airgiodslv	97	22	bob
23	74	time	48	44	picspam	73	28	rock	98	22	bang
24	74	sheppard	49	43	world	74	28	lavvyan	99	22	door
25	74	rps	50	43	recs	75	28	foxxcub	100	21	fanfiction

Chapter 7. Discussion

Social bookmarking produces a new information environment where users are actively involved, as a part of their own information management strategy, in the collective accumulation of knowledge. This study put the dual nature of social bookmarking (as personal information tool and as social software) in the center and assessed how the overlaps in personal information spaces of users within a social bookmarking site can be used to identify shared interest spaces among users. The problem of overlap and interest sharing was addressed in three consecutive phases, each with an increasingly specific question and the analysis built on the results of the previous phase.

This chapter begins with a discussion of the sampling strategy adopted in this study to examine the large information space of a social bookmarking site, *delicious.com*. Each research question, addressed in each phase, is then presented along with the major findings. Potential explanations for and implications of the findings are also discussed in each section.

7.1 Implications of the sampling strategy for the findings

Openness is one of the defining features of social bookmarking. In *delicious.com*, as in many other social bookmarking sites, all the activities on the site are, by default, open for everyone to see. However, while any bookmark on the site is potentially accessible[34]

[34] A site user can set his/her bookmarks to be private, so they would not be accessible to other users.

when a user accesses the site, what he/she can see is a particular part of a dynamically changing information space; one does not and can not see the information space in its entirety. Except for the rare case when a researcher is granted direct access to the internal database of the site being studied, a researcher is facing the same situation – only partial representations of the entire information space are visible. Data collection is usually done by crawling pages from the site, but in most cases it is not feasible to crawl the entire site because of many technical, legal, and practical reasons. This is particularly true for a site like *delicious.com*, where a large volume of activity is constantly going on and the size of the site is extremely large. Therefore, different studies create their own sample datasets of varying sizes. It is important to be aware of the potential impact of the sampling method employed in a particular study on the analysis and the study's result. Depending on where the sampling was started, for instance, different datasets may represent entirely different parts of the information space. One of the main methodological considerations in designing this study, therefore, was to define as clearly as possible how the datasets were constructed and which parts of the information space each was intended to represent.

The basic building block of the information space of social bookmarking is a bookmark posting, consisting of information on *who* bookmarked *what* information object *when*. A bookmark posting may contain tags and/or comments if the user opted to add them, but the above three elements -- the information object being bookmarked, the user who posted the bookmark, and the time of posting -- are always present. In this study, therefore, the information space of *delicious.com* is conceptualized as being comprised of three dimensions: information objects, users, and time. The sampling strategy used in the

first phase of this study aimed to capture both the breadth and the depth of the information space of *delicious.com*, while simultaneously limiting the sample to a size that could be analyzed with current methods. Two complementary methods were used to this end. First, using the RSS feature of *delicious.com*, a large number of recent bookmark postings were collected into a dataset called *Recent*. This dataset represents the *current* breadth of the information space with a broad range of information objects and users. Second, for capturing the depth of the space along the dimension of time, a sample of users and a sample of URLs were drawn from the *Recent* dataset and, for each sample set, the entire history of bookmarking activities associated with each element was collected, by crawling all the relevant pages from *delicious.com* and parsing out each bookmark posting. This second sampling process resulted in two datasets: *User History* and *URL History. In combination, the three datasets* serve the purpose of representing the characteristics of the information space as comprehensively as possible, with a limited sample size.

A different sampling strategy was used for the second phase, based on the analysis of the first phase data. A set of *active* users was defined by the intensity and frequency of their recent bookmarking activities. That is, the users to be included in this set were selected from the *Recent* dataset, based on the number of postings that they made within the given period and the number of different days on which they posted one or more bookmarks. For each user in this *active* user set (23,238 users in total), the entire bookmarking history of the user was collected by crawling all the pages on his/her account. Comparing this second phase dataset with the first phase *User History* dataset, which contains users randomly selected from the *Recent* dataset and their entire

bookmark collections, it turned out that the level of activities in this set, in terms of the volume of bookmark postings, is indeed significantly higher. On average, the total number of postings of a user in this set is almost three times higher than that of a random user, regardless of the time window (entire history, 12 months, 6 months, or 3 months) examined. This means that many of those users who were active in their recent postings had been active users over a longer period. While the obvious implication of this, in the context of this study, is that the dataset would yield a dense network of users, as intended, it also indicates that we can expect a certain level of stability in the patterns of user behaviors in *delicious.com.* As one of the earliest and most successful instances of social bookmarking, this site seems to have an active core in its user population, including early adopters who have continued using the site for several years.

The second phase also differed from the first phase in its scope. In the network analytic framework, the task of defining subjects (actors) to be included, called boundary specification, constitutes a critical part of determining the scope of an analysis. In this study, the boundary of the network to be analyzed was defined based on two criteria, the number of recent postings and the number of days at least one posting was made. More specifically, the set of users to be included in the network was selected from the *Recent* dataset by combining the top 10% on each criterion. This set of users represents the most *active* part of the *current* user population. This boundary specification is in part affected by the basic consideration mentioned above; that is, given the large scale and dynamic nature of the information space, it would be useful to define a specific target area to probe within the information space, so that the findings can be interpreted with reference to the relevant area. Another key factor supporting this boundary definition was

the findings of the first phase. All the distributions of bookmarking activities, being examined from different perspectives, revealed a great deal of heterogeneity. While the vast majority of information objects were bookmarked once or a few times, some information objects had extremely large number of postings. Similarly, while the majority of users were less active, with the number of bookmarks way below the average, some users were far more active. This finding suggested that a selection of users on a random basis might produce a network of many isolated nodes, mainly because of the lack of enough shared bookmarks to connect the users. More importantly, the structure of the network could be dominated by a small number of those users who have a massive number of bookmarks and happened to be chosen into the sample. In other words, depending on the very small portion of highly active users selected by chance, and the characteristics of their bookmark collections, the resulting network could vary to a large extent. By including active users as broadly as possible with the given size limitation, this study attempted to reduce the variation by chance and to acquire reliable results at least within the limit of the target area.

7.2 Discussion of first phase findings

The first phase was concerned with the question of accumulation and overlap in the information space of *delicious.com*. The main focus of the analysis was to assess the current level of shared interests among *delicious.com* users, by looking at aggregate data describing bookmarking activities from the resource-centric view and from the user-centric view.

The analysis in this phase was mainly comprised of statistical descriptions of

bookmarking activities. In *delicious.com*, bookmarking activities accumulate along the axis of information objects and the axis of users. When a bookmark is posted by a user for an information object (URL), it is simultaneously added to the corresponding URL page and the User page. Depending on which axis is taken first, therefore, a set of bookmark postings can be seen as a set of information objects, each of which has a subset of users who bookmarked the given information object (a resource-centric view), or as a set of users, each of whom has a collection of information objects (a user-centric view)[35]. Taking advantage of this interwoven structure inherent in *delicious.com*, the analysis in this phase was done both from the resource-centric view and from the user-centric view. A number of statistics and the distributions of their quantities were examined from each view, to bring a complementary understanding of how the information space of *delcious.com* is (being) shaped, and how bookmarking activities are spread and accumulated along the two axes. Thus, this phase assessed the overall *texture* of the information space and the applicability of the network approach to this information space at the data level, supporting the conceptual framework of the study. As the first step towards investigating shared interest spaces within *delicious.com*, a specific objective of this phase was to evaluate the extent to which bookmarking activities overlap among users over information objects. Such overlaps formed the basis for identifying and connecting users who share interests in the second phase.

Even in the first phase, the question of interest sharing was addressed indirectly by examining aggregate patterns. For instance, the popularity distribution of information

[35] Note that, if the entire information space was examined, the resource-centric view and the user-centric view would have shown the exactly same set of users and information objects, and the only difference would have been in their arrangement.

objects showed the overall scattering and gathering of users over the range of information objects, from which the level of interest sharing was inferred. Note that, while a complementary approach was taken by analyzing data from both resource-centric and user-centric perspectives, the first phase looked at information objects and users separately. Only in the second phase was social network analysis introduced to attend to the relations between them.

In the following subsections, the important findings of the first phase are summarized and some observations deemed noteworthy are also discussed.

7.2.1 Accumulation and overlap from the resource-centric view

From the resource-centric view, the main question of accumulation and overlap was primarily addressed by examining the distribution of popularity of the information objects. The number of times an information object was bookmarked is equivalent to the number of users who had an interest in that object, which represents the popularity of the information object. In both the *Recent* dataset and the *URL History* dataset, it was observed that the majority of information objects were bookmarked only a few times, while there were a relatively small number of information objects that were bookmarked a large, often extremely large, number of times. This kind of skewed distribution is often observed in information-related phenomena (Bates, 1998; Brynjolfsson et al., 2006; Anderson, 2006).

While both the *Recent* dataset and the *URL History* dataset have a large proportion of less popular items, the two popularity distributions showed a difference that is important in the context of this study. In the *Recent* dataset, almost 90% of all the

URLs in the dataset occurred only once, whereas the percentage of URLs that were posted once is significantly lower (36%) in the *URL History* dataset. Note that, in this study, the information objects that users have bookmarked in common serve as an indicator of shared interests, and only those URLs that were posted multiple times could create connections among the interested users. Because of that, the difference in the percentage of single postings in the two datasets was important[36].

In order to see whether the higher percentage of single-occurrence URLs in the *Recent* dataset was the consequence of different data collection methods, the full history of some of those URLs that occurred only once in the *Recent* dataset was tracked using the *URL History* dataset. In the *URL History* dataset, 7,216 URLs out of the 10,000 URLs were those which appeared only once in the *Recent* dataset. The examination of the history of those 7,216 URLs revealed that 5,239 (72.6%) URLs were indeed posted only once in the four month period (January, 2008 – April, 2008)[37]. It was concluded that, although some of the single postings in the *Recent* dataset can be ascribed to the time interval gaps in data collection, the higher proportion of single postings in the *Recent* dataset was not a mere artifact of the data collection method.

[36] It should be mentioned, however, that since the data collection of the *Recent* dataset involved time intervals between RSS fetches, the fact that a URL appeared only once in the *Recent* dataset does not necessarily mean that the URL was posted only once during the fourteen week period (from January 14, 2008 to April, 21, 2008). On the other hand, the popularity distribution of the information objects in the *URL History* dataset considers the complete history of bookmarking activities associated with each information object, that is, all bookmark postings added to each information object from its first appearance on *delicious.com* to the last posting before the data collection.

[37] At the time of data collection, the URL pages in *delicious.com* arranged the postings by month, and the information on the exact date for each posting was not provided. Therefore, it was not possible to trace records precisely from January 14, 2008 to April, 21, 2008. Instead, the four-month period, from January to April, was examined.

The extremely high percentage of single postings in the *Recent* dataset suggests that, while bookmark postings are constantly added to *delicious.com*, when a short period of time is considered, there would be little overlap in the information space. In other words, during a short period, the level of shared interests that can be observed from bookmarking choices of users is low. The lower percentage of single postings in the *URL History* dataset (36% as noted above) is encouraging, since it indicates that the level of accumulation increases over time, in general. While still the majority of URLs in the complete *URL History* dataset were posted only a few times, those URLs have two or more interested users, forming small circles of shared interests.

With the observed differences in the overall degree of overlap in the two datasets, *Recent* and *URL History*, a question may arise as to when bookmark postings start to accrue on information objects. One way to look into this question is to use the incremental nature of the way the *Recent* dataset was constructed. In order to see the effect of accumulation over time, two statistics were calculated at an interval of every 10,000 new postings added to the dataset: 1) the ratio of distinct URLs to the total number of bookmarks in the dataset and 2) the proportion of multi-posted URLs. The results showed a slow yet steady increase in the proportion of URLs that were posted multiple times. As the number of postings in the dataset grew, new postings started to duplicate the existing information objects, and the ratio of distinct URLs to the total bookmark postings decreased little by little. By the time 1,000,000 postings were collected into the *Recent* dataset, the ratio of distinct URLs to the total number of bookmarks was 8.2 to 10, and about 10% of the distinct URLs had more than one posting. Obviously, the *Recent* dataset as a whole is comprised of a large variety of

distinct information objects, indicating a broad range of user interests. The gradual increase in additional postings of existing information objects in the *Recent* dataset suggests that interests in certain information objects may spread among the users of the site over time. An examination the *URL History* dataset also supports this conclusion. For each URL in the *URL History* dataset, the URL's age in *delicious.com* was calculated by counting the number of months that had passed since its first posting. All URLs were then grouped by age, and the average number of postings per age group was calculated. The result showed that the average number of postings per URL tends to be higher for older age groups.

Finally, another question regarding accumulation from the resource-centric view concerned the different patterns of accumulation or growth in popularity across various information objects. Both in the *Recent* dataset and the *URL History* dataset, a highly skewed popularity distribution was observed, with a small number of information objects having a number of postings far above the average. In their early study of *delicious.com*, Golder and Huberman (2006) analyzed 212 URLs that appeared on the 'popular' page (a page on *delicious.com* where the system supposedly lists the most popular URLs at a given point in time), and reported that the majority of the URLs in their dataset reached the peak of their popularity pretty quickly, although other patterns were also observed in some of the popular URLs. A slightly different approach was taken in this study, to examine the relationship between popularity and age, by plotting the number of postings of each URL in the *URL History* dataset by its age. The result shows that, while the average number of postings per URL increases over time, there exists a great deal of variation in the popularity of information objects in any given period. Some information

objects gained a very high level of popularity in a short period of time, often within just a month. This finding is in accordance with Golder and Huberman's (2006) findings. As described in section 4.1.2, these information objects included blog entries or articles recently published or a new site just launched. This kind of burst of attention demonstrates the power of the viral spread of information on the Web in general and on *delicious.com* in particular. It may be due to a kind of preferential attachment, at least in part facilitated by *delicious.com* itself, since it allows users to see the most 'popular bookmarks' at any given moment. Not only can users copy any bookmark on *delicious.com*, but a user can 'subscribe' to bookmarks of other users so that each time the subscribed user adds a bookmark, the subscribing user is automatically informed. Some users, therefore, can be quite influential in spreading information on *delicious.com* due to this subscription mechanism. Another likely source of a sudden burst of interest in a certain item would be an introduction or comment about the item on a popular blog or website. It is hard to tell how much of the popularity of an item can be attributed to external factors and how much to *delicious.com*'s social navigation features. It seems clear, however, that various social processes or mechanisms, either within or outside the site, play a substantial role in the emergence of highly popular items, especially for those that were posted very heavily very quickly.

There are other information objects on which bookmark postings accumulate gradually over a long period of time. Those information objects include quite general sites related to people's everyday use of the Internet, such as A*mazon.com* or *The Internet Movie Database* (*IMDb)*. Sites that are known to be popular among tech savvy people, such as *Slashdot.org*, are also ranked high in the popularity distribution,

suggesting a characteristic of the user population in this particular site.

7.2.2 User behaviors and the level of shared interests

From the user centric view, we looked at the general level of activity. The statistics mostly pertain to individual aspects of bookmark collection, showing how frequently and heavily users in *delicious.com* are engaged in bookmarking activities in general.

In the *Recent* dataset, for each user, the number of postings made by a given user and the number of different days when he/she posted at least one bookmark were calculated [38]. The distribution of the number of postings again showed clear right-skewness, with a large portion of less active users and a relatively small number of highly active users. More specifically, about 81% out of the total of 288,727 users have five or less bookmarks, while some users posted more than 100 bookmarks during the 14-week period that the *Recent* dataset was collected. Among the total of 1,226,472 postings in the *Recent* dataset, about 37% were made by those 81% of less active users (with five or less postings) and the remaining 63% of the postings were made by the 19% of users who were moderate to highly active (with the number of postings ranging from 6 to a maximum of 935). In order to define the set of active users, the cut-point of 9 or more postings (the top 10%) was used. This top 10% of active users accounted for 50% of the total postings. In light of this finding, the large proportion of distinct information objects in recent postings can be interpreted as representing constant expansion of the information space by those users on the active front adding new information objects.

[38] As mentioned before, these two measures were used together as the criteria defining the set of *active* users for the second phase analysis.

Not surprisingly, the *User History* dataset also showed a heterogeneous distribution of users in terms of the size of their cumulative collections. Again there existed a small group of highly active users and another, much larger, group of users who used the site only a few times. However, unlike all other distributions examined before, this dataset has a thick middle layer, consisting of moderately active users. Among the 10,000 users in this dataset, nearly half (46%) of them have between 100 and 1,000 bookmarks in their collections. The result showed that while users vary in their level of activity, a considerable proportion of users use the service quite heavily over a long period time, making it an important part of their information management infrastructure.

The above statistics from the user-centric view depict bookmarking/usage patterns of individual users in *delicious.com*, but they do not show how much overlap exists among bookmark collections of different users. Whereas the distribution of the number of postings from the resource-centric view in and of itself shows the level of interest sharing (since the number of postings on a particular information object is in effect the number of users who share interests in that information object), the number of bookmark postings of a user pertains only to the individual. Since one of the main questions addressed in this phase was concerned with shared interests among users, an additional analysis was conducted to address the question from the user-centric view. In both the *Recent* dataset and the *User History* dataset, in addition to the total number of bookmarks a user has, the number of bookmarks he/she shares with one or more other users was calculated so that the proportion of shared bookmarks can be measured. Averaged over all users in the dataset, the proportion of shared bookmarks was about

27% in the *Recent* dataset and about 67% in the *User History* dataset. Once again, a large increase in the measure of shared interests was observed in longitudinal data. Moreover, within the *User History* dataset, the proportion of shared bookmarks is consistently higher for those users with larger collections. The ratio of shared bookmarks to the total number of bookmarks in a user's collection represents how unique his/her collection is. It is interesting to observe that less active or moderately active users tend to have more unique collections, while highly active users consistently have a larger proportion of shared bookmarks.

In summary, the two most notable patterns observed in the first phase results are 1) the diversity and heterogeneity of all the distributions examined, and 2) the increase in the level of overlap over time. Together these findings illuminate both personalized aspects and social aspects of social bookmarking. The highly skewed distribution of information objects with a long tail of less popular objects suggests that users have diverse interests and often make idiosyncratic choices. On the other hand, the existence of extremely popular information objects, especially ones that reached the highest level of popularity shortly after their first appearance on the site, indicates that a sort of social contagion or viral spread might be in play. While users vary to a large extent in the intensity and frequency of their use of *delicious.com*, many have built substantial collections of bookmarks over time. In most cases, a user's collection is comprised of a subset of unique bookmarks (belonging only to him/her) and a subset of shared bookmarks. An interesting finding was that users with larger collections tend to have a larger proportion of shared bookmarks, indicating that active users in this site might be more involved in its social aspects. Overall, while a broad range of diverse interests was

observed in recent bookmarking activity, the community also has expressed a considerable amount of shared interests.

7.3 Discussion of second phase findings

The main research question of the second phase was concerned with whether users of *delicious.com* can be connected to form a network based on their prior bookmarking choices, and whether such a network can be used to study the patterns of interest sharing among users that may exist in the information space. In the following, the basic properties of the induced network are discussed, as well as the results of the *m*-slice analysis.

7.3.1 Properties of the network of active users

In this phase, a network of users was induced based on their prior bookmarking choices. As described above, a new dataset was constructed so that the network would represent relations among a set of *active* users in the current user population of *delicious.com*. This involved 1) selecting a subset of users from the *Recent* dataset based on the intensity and frequency of their recent activity, and 2) acquiring the entire bookmarking history of each of those users. This sample of data was then used to draw relations among the included users.

With this new set of users and the information objects bookmarked by them, three different affiliation networks were constructed for three different time windows (12 months, 6 months, and 3 months). Each network was then transformed into an one-mode network, a social network of users. In each network, two users were connected if they had bookmarked at least one

common information object within the given period of time. While the most basic properties, including the number of nodes, the number of edges, and the average degree, were measured for each of the three networks, all the subsequent network analysis was conducted only with the 3-month data for the sake of computing efficiency.

The most notable finding related to these networks is that the level of connectivity, even in the network based on the shortest period of 3 months, was exceptionally high. Generally speaking, a one-mode network induced from a two-mode affiliation network tends to be dense, because all the actors affiliated with a group are, by definition, completely interconnected to form a clique in the one-mode network. In addition, depending on the extent to which actors have multiple affiliations, there can be many connections bridging those cliques, further raising the overall connectivity of the network. However, the level of connectivity observed in the above networks is high, even for an induced one-mode network. A comparison with other affiliation-based networks will make this point clear. Table 7.1 presents the basic network statistics commonly used for investigating the internal structure of a network:

n, the total number of vertices,

m, the total number of edges,

z, the average degree,

l, the average shortest path length (distance) between any pair of nodes, and

C_1 and C_2, two clustering coefficients.

The first row shows calculated values for the network of *delicious.com* users induced from their bookmarking activities within the three month period (February 2008 – April 2008). The next

three rows show the same statistics for well-known empirical networks studied elsewhere[39].

Table 7.1 Basic network statistics of induced one-mode networks

Network	n	m	z	l	C_1	C_2
Delicious.com users	23,172	17,918,940	1546.60	2.01	0.32	0.43
Film actors	449,913	25,516,482	113.43	3.48	0.20	0.78
Company directors	7,673	55,392	14.44	4.60	0.59	0.88
Coauthors in physics	52,909	245,300	9.27	6.19	0.45	0.56

As can be seen in Table 7.1, the average degree per node, which is known as being comparable among networks of different sizes, is much higher in this network. Although a shared membership creates actual connections, there are two factors contributing to the volume of connections in an affiliation network and a network derived from it: the number of affiliations one could have and the number of members a group might have. In this network, the upper bound for each mode is much higher compared to other networks. For instance, whereas the number of papers a scholar could possibly write would be clearly limited even for a very prolific writer, the number of bookmarks one can add to his/her collection would be only limited by their choice. Likewise, there is virtually no constraint imposed on the number of postings (thereby interested users) gathered on an information object in social bookmarking, while all other types of affiliation have a more or less limited space to carry members. Therefore, the

[39] The original study for film actors was reported in Watts and Strogatz (1998), company directors in Davis et al. (2001), and physics coauthors in Newman (2001a, 2001b).

high level of connectivity in this network can be, at least in part, attributed to the large number of postings the active users in our dataset have, as well as the existence of popular items. In other words, the chance that a pair of users in this dataset is connected by sharing at least one bookmark was high thanks to the two contributing factors. However, it should be noted that this is an after-the-fact analysis of the reason for the observed volume of connections.

A study conducted on different social bookmarking sites using a similar approach reported a different finding as to interest sharing among users. Santos-Neto et al. (2009) studied two social bookmarking sites, *CiteULike* and *Connotea*, by reportedly collecting all the activities on each site from late 2004 to early 2009. With the datasets including 1,325,565 items posted by 40,327 users on *CiteULike* and 509,311 items posted by 34,742 users on *Connotea*, they found that 99.9% of user pairs in *CiteULike* had no items in common and, similarly, 99.8% of users in *Connotea* had no shared items. If a network of users had been constructed based on shared bookmarks using either of the datasets, the resulting network would have been very sparse with few connections. The authors supposed that users of the sites maintain their accounts primarily for personal information management needs, and such behavior could explain the lack of commonality in people's bookmarking choices. Interestingly, the nature of social bookmarking as a personal information management tool was one of the key motivations for this study, along with its nature as social software. As discussed above, the results of the first phase of this study, with various statistics depicting user behaviors, showed that the personal aspect was indeed apparent in *delicious.com*, while a substantial level of overlap was also observed. The main differences between *delicious.com* and the two sites that

Santos-Neto et al. (2009) studied would, then, be in their realization of, or lack thereof, social characteristics. One possible explanation could be found in the size of the user population and the volume of their activities. As for the user population, recall the *Recent* dataset constructed in the first phase of this study. The number of users who posted one or more bookmarks in *delicious.com* during the fourteen week period of data collection was 288,727, far exceeding (7 to 8 times) the total number of users in Santos-Neto et al.'s datasets which reportedly included all the activity from late 2004 to early 2009 in *CiteULike* and *Connotea*, respectively. Moreover, the intensity of user activity in *delicious.com*, in terms of the number of postings, is higher in orders of magnitude. On average, a user of *CiteULike* made about 33 postings, and a user of *Connotea* made only about 15 postings. In contrast, in the group of active *delicious.com* users studied in the second phase of this study, a user made, on average, about 1100 postings throughout their membership with *delicious.com* and about 200 postings during the 3 month period (Feb. 2008 – April 2008). Even for the random users included in the *User History* dataset in the first phase, the average number of postings per user was about 352. Both in the user population and the volume of activity, it appears that *delicious.com* achieved a critical mass that can bring about social behavior, while the other two sites did not. It should be noted, however, that *CiteULike* and *Connotea* have a narrower scope than *delicious.com*. Since the sites provide a reference/citation management tool, both the range of information objects and the potential user bases are limited compared to *delicious.com*, which is a general-purpose bookmark management tool. In any case, the different findings between this study and the Santos-Neto et al. (2009) study, as to interest sharing among users, indicate that *delicious.com* is much closer to the 'social'

end of the spectrum, compared to *CiteULike* and *Connotea* which are on the other end of the spectrum, 'personal'.

The level of connectivity is the primary indicator of the overall cohesion of a network. Other structural properties of the network of *delicious.com* users also indicate that users in this network are connected in a cohesive manner. First, the giant component of this network is extremely large, containing 99% of the entire network, while the second largest component contains only 2 nodes. With random graphs, it has been shown that as the density of a network increases, vertices and subsets of vertices tend to be continuously joined together and eventually form a giant component, which covers an extensive portion of the network (Newman, 2001). In fact, the existence of a giant component is observed in most real world networks. More often than not, there also exist multiple components of varying sizes outside the giant component. In a large scale network, researchers often attend to the medium to small size components since those components might represent separate groups. In this network, however, almost all users were connected into the giant component, and there were no other components consisting of a relatively small number of users, except for a few duos and isolates. In other words, almost all users in this network belong to a single community without separating into small groups, at least at the global level. Second, not only can users reach almost everyone else in the network through one or more paths (by being in the giant component), the shortest path length between any pair of users in the giant component is typically very small, with an average shortest path length of 2.01 as shown in Table 7.1. Third, the average local clustering coefficient, as well as the global clustering coefficient, was substantially higher in this network than a randomized

counterpart of the network (i.e., an Erdös Rényi (ER) network (Erdös & Rényi, 1959) with the same number of nodes and edges). Note that a high clustering coefficient, together with a short path length, is a key parameter that Watts and Strogatz (1998) incorporated into their model of the small-world effect. Their model explains how a network with local clusters is at the same time globally connected, by virtue of a small number of 'short cuts' connecting otherwise separate clusters. In a study of the link typology of the Web (Kleinberg, 1999), the small world properties of the network were in part attributed to its degree distribution, where a relatively small number of well-connected 'hubs' are interlinked to create a backbone of the network, providing global connectivity. In our network of *delicious.com* users, it was also observed there was a negative correlation between the node degree and the local clustering coefficient, suggesting that nodes with a higher degree tend to have connections stretching across local clusters. All in all, these results indicate that the active users' network consists of a single community, which constitutes a 'small world' where everyone can reach everyone else within a few steps.

The small world effect was first described in the seminal paper by Milgram (1967), which reported a counter-intuitive observation that two people in distant locations can be connected through a short chain of acquaintances. A substantial body of literature since then has confirmed the pervasiveness of the small world effect in many real world networks, and has discussed the possible impact of the structure on the functioning of the systems of interest. In general, the advantage of the network structure characterized by a small world is that it provides both local cohesion and global accessibility (White & Houseman, 2003). Being situated in a locally dense cluster which also has a few

bridges spanning boundaries of multiple clusters, a node can have access to a variety of non-overlapping resources outside the cluster as well as to the resources of primary relevance within its own cluster. From a global perspective, short distances allow fast transmission of information on the network, and redundant connections in local clusters promote diffusion since a new piece of information can spread quickly over all the nodes in a cluster through multiple paths (Yamaguchi, 1994; Schilling, 2004). In the context of the current study, the dense structure found in the network of *delicious.com* users, along with the small world properties, suggest that the basic structural conditions for building an efficient information sharing mechanism are already in place. The fast spread of popular information objects observed in the *Recent* dataset in the first phase supports this supposition.

It should be mentioned, however, that the high level of connectivity appears to be a quite dominant factor affecting other properties and the overall structure of the network[40]. Moreover, one may argue that the high level of connectivity could be a consequence of the low benchmark established for connecting users in this set. Given the characteristics of this dataset, consisting of active users with the average of about 200 bookmark postings, the condition of 'one or more shared bookmarks' might not be a sufficient indicator of shared interests. That is, having one out of 200 in common with another user may or may not reflect shared interests between the two users. As shown

[40] For instance, when an ER random network was constructed using the same number of nodes and edges as this network, the characteristic path length of the random network was as short as this network (about 2) even though its local clustering was substantially lower. In other words, the characteristic short path length in our network could be simply the function of the large number of edges, unlike those more intriguing cases where two nodes belonging to distant local clusters in a sparse network could still be connected within a short chain of intermediaries. In addition, the ER random network also had a giant component of a comparable size.

in the popularity distribution of the information objects, there exist extremely popular information objects. Regardless of the content of the rest of their bookmark collection, an addition of one such popular item would connect the users with a large number of other users.

7.3.2 *M-slice* analysis: persistent structural patterns

The *m*-slice technique adopted in the second part of the second phase allowed a further investigation of the internal structure of the network, while addressing the above mentioned issue of the low benchmark established for connecting users. In an *m*-slice, users are connected if they have at least *m* information objects in common, instead of a single item. By incrementally increasing the threshold value *m* of the minimum shared bookmarks, it was possible to make the benchmark progressively stricter.

The *m*-slice process was repeated from *m*=1 to *m*=600. At each step, the *m*-slice represents a sub-network of the original network, where each and every edge has a weight equal to or greater than the given *m* value. In the context of this study, it means that each and every connected pair of users in a given *m*-slice has *m* or more bookmarks in common. For each iteration, the number of edges removed and the number of remaining nodes were recorded, and the same set of the basic network properties as in the original network were measured for the resulting network (*m*-slice). A number of notable patterns emerged during the process. First, while the number of edges decreased exponentially as *m* increased, the number of nodes reduced slowly, especially in the early iterations. More specifically, 74% of the edges were removed at *m*=2, but about 93% of the nodes were still connected. At *m* = 4, while almost 94% of all the edges in

the original network were removed, about 70% of nodes were remaining. By the time *m* reached 30, almost all (99.9%) of the edges were taken out, but the network still had almost 12% of the nodes of the original network. Second, even after the vast majority of edges were removed, the distance between the remaining nodes, either the diameter of the network or the average shortest path length, did not increase substantially. These two findings, in combination, suggest that a large portion of the removed edges had provided redundant paths among users in this network. In other words, the ties representing a few shared information objects played rather marginal roles in sustaining the structural integrity of the network. Third, as *m* increased, the average of the local clustering coefficients grew higher whereas the global clustering coefficient steadily dropped. Such a discrepancy between the two clustering coefficients has been observed in other networks, as can be seen in Table 7.1, and is generally attributed to the fact that the average of the local clustering coefficients tends to give more weight to lower degree nodes. The reason why the discrepancy increased with higher *m* value in this network can be found, in part, in the changes observed in the degree distribution. While the degree distribution of the original network was far from a power-law form due to the large number of moderately well-connected nodes, the degree distribution of higher *m*-slices approached a power-law distribution (see Figure 5.4 and Figure 5.10) with a substantially larger number of nodes having a low degree. In addition, the negative correlation between the node degree and the local clustering coefficient observed in this network suggests that nodes with lower degrees tend to have higher local clustering than those with higher degrees. The two observations in combination indicate that the increased number of nodes with a lower degree, which have relatively higher local

clustering coefficients, raised the average measure. In any case, the measure clearly shows that nodes in this network tend to form local clusters.

An important implication of the above findings is that the small-world effect is indeed a genuine characteristic of the network of *delcious.com* users. The two characteristic properties – short path length and high clustering – turned out to be persistently present in all *m*-slices. Note that in *m*-slices the effect of a few extremely popular information objects on the volume of the connections was reduced. Since each connection is established by a set of information objects, the resulting network is less driven by the popularity of individual information objects. For instance, suppose there was an information object bookmarked by 1,000 users. By virtue of the single information object, those 1,000 users would be completely interconnected (everyone has a direct link to everyone else) in the original network. In 2-slice (i.e., m=2), however, the users would lose (direct) connections to other users in the group unless they had at least one additional information object in common. Therefore, as m (the minimum required number of shared bookmarks to create connections) increases, it is more likely that the structure of the network represents a non-random pattern of overlaps in user interests.

The last part of the *m*-slice analysis was concerned with identifying relatively cohesive regions within the network's giant component. As discussed above, the giant component of the original network encompasses almost 99% of the entire network. Therefore, the initial division of components in the network did not depict any subgroup structure. In order to address the third question of this study – whether and how the network of *delicious.com* users breaks down into communities – the analysis of components within each *m*-slice was conducted to examine the subgroup structure that

may exist in the network. The result showed that as *m* increased, the number of components of size two or more increased quickly and the size of the giant component decreased steadily. However, the second largest component remained small even after 20 iterations. This finding means that the components taken apart from the giant component were mostly pairs or clusters of only a few nodes. The first break-down of the network into communities of substantial size occurred at *m*=28, where the size of the second largest component was 248 and the size of the third largest component was 14. Along with the largest component (consisting of 2,458 nodes) of the 28-slice, these two components were regarded as separate communities. An exploratory analysis of the content (the shared interests) of the communities was conducted in the third phase.

Another internal structure of this network that was revealed through the iterative component analysis was its core-periphery structure. A core-periphery structure is characterized by a densely connected core and peripheral nodes that are attached to the core but not connected with one another (Wasserman & Faust, 1994). According to Everett and Borgatti (2005), "a graph has a core-periphery structure to the extent that it lacks subgroups. Another way of putting it is that all nodes can be regarded as belonging (to a greater or lesser extent) to a single group, either as core members or peripheral members." (p.68). The way in which our network was decomposed as *m* increased was consistent with the above definition. While at each step a considerable number of users were continuously taken out from the giant component, those users themselves were, by and large, not connected to each other. The majority of users detached from the giant component became isolated or formed small groups of a few users. Up until *m* reached 28, the network of *delicious.com* users did not have any communities of substantial size except the giant component. This finding indicates that the core was not only densely connected, but the

strength of ties within the core was high enough not to be divided into parts, while peripheral nodes were connected to the core with relatively weak ties and were easily split off from the core as the *m* value increased.

A core-periphery structure is commonly found in a social network, especially within a well-established system (Borgatti & Everett, 1999). It is a stable structure in that the dense connections that the members of the core have developed provide an effective mechanism for consensus and compliance. The members of the core tend to share information, adopt similar views on issues, and assimilate behaviors of the others. On the other hand, the periphery holds open the possibility for new members, diverse resources, or innovations to be introduced. The peripheral nodes are connected to the core typically with relatively weak ties and they are not as visible or engaged in the community as the members of the core. Note that a large network may contain multiple core-periphery structures, each of which is a well-developed community. In that case, peripheral nodes in one community may include bridges to another community, or those who span boundaries of multiple communities.

In observance of the prevalence of this structure in social networks, there have been discussions on the advantages that the distinct characteristics of this structure bring to the network as a whole as well as the different leverages that a core or peripheral position may allow individual actors to have (Barsky, 1999; Cummings & Cross, 2003). Of interest to this study, it has been suggested that a core-periphery structure has particular strengths for an information and/or knowledge network (Borgatti, 2005). In general, information tends to spread fast to the entire network once it gets to the core. The nodes in the core have greater access and/or exposure to relevant resources and,

although there would be redundancies, multiple channels within the core also ensure reliable access to high-quality information. On the other hand, the peripheral nodes reach both the information circulating in the core and new information from outside. Their distance from a dominant core allows them to be more open to innovative ideas or front-end developments.

The above advantages can be applied to the network of *delicious.com* users. Information objects endorsed by the users in the core of the network can quickly spread and, thus, further increase the cohesion of the information space, while the boundaries of the information space can be expanded by peripheral users who may introduce areas relatively unknown to the core. It is promising that the structural features that are known to carry social advantages for its constituents and for the system as a whole were observed in this network.

7.4 Discussion of third phase findings

The last question of this study was concerned with whether it is possible to locate communities of shared interests within the network of *delicious.com* users. The question was addressed in two parts: the identification of the community structure of the network based on structural patterns (in the second phase), and the examination of the cohesiveness of the identified communities in terms of relational content, or shared interests (in the third phase). As described above, using the nested structure of the *m*-slices of the network, it was found that the network breaks down into communities at $m = 28$. While the structural patterns -- the overall strength of connections within the subgroup in this case -- led to the division of regions, it needs to be further investigated

whether those regions of the network indeed constitute coherent communities of shared interests. Freeman et al. (1989) pointed out that, "Most network researchers seek to uncover clusters of actors. They stress structural patterns and seemingly avoid any discussion of relational content at all" (p.5). The third phase of this study, in recognition of the needs for the analysis of relational content, was a preliminary exploration of the communities identified through *m*-slice component analysis, in terms of their *relational content*. More specifically, taking the three largest components found in the 28-slice[41] as separate communities, this phase looked at the content and coherence of shared interests within communities.

The duality of the affiliation network provided the basis for this analysis. Recall that links in the network of users were drawn from their common relations to the information objects. Every link in a community was traced back to the information objects involved in creating the link, and the union set of information objects for the community was constructed. For each information object in the set, its contribution to the connectivity and coherence of the community was measured with an index, named the contribution index. In addition, the extent to which the users who bookmarked the given information objects were brought together into the community in question was measured by another index, named the aggregation index. The set of information objects for a community was then sorted by the contribution index, and the content and cohesiveness of shared interests defining the community was examined by looking at the URLs and titles of the information objects and by averaging over the aggregation index

[41] The 28-slice was chosen because it was the first breakpoint where the second largest component, broken off from the giant component of the previous *m*-slice, reached a substantial size.

within the set.

The results indicated that each of the three largest communities had a distinct theme defining the interests of the community members. Not only the top ranked information objects (by their contribution index) were clearly on a topic, but the majority of the information objects in each community shared the central theme of the community. The first community consists of 2,458 users. The number of links among them was 19,115, and after tracing every link the union set of information objects included 51,847 distinct URLs. The theme of the first community was web design and/or web development, reflecting the general characteristics of the user population of *delicious.com*. Both the top ranked URLs shown in Table 5.8 and the frequent title terms shown in Table 5.11 were all tightly related to the theme. The second community turned out to be a community of *fan fiction*. From 1,306 links among the 248 users in this community, 6,414 distinct URLs were traced. The coherence of shared interests in this community was remarkably high: all the top linked information objects were found on LiveJournal.com which is a well-known host of online fan communities, and the string 'fic' referring to 'fiction' occurred very frequently in titles of the information objects in the union set of this community. The third community consisted of only 24 users and 29 links among them. Interestingly, the number of distinct information objects was relatively high, with 831. Compared to the other two communities, there was little overlap on individual information objects, meaning that different pairs of users had, by and large, different sets of shared information objects. However, the information objects in the union set also showed a clear theme. All of the 831 URLs were related to cooking and/or baking, with the vast majority of them being about baking pies, cakes,

cookies, and so on.

The aggregation index averaged over all the information objects in each community was 0.73, 0.86, and 0.70, respectively. Note that the aggregation index of an individual information object in a community was defined as the proportion of the users who were included in the community among all the users who had bookmarked the given information object in the original network. The aggregation index of 0.73 in the first community means that, on average, 73% of all the users (in the original network) who had bookmarked the information objects in its union set were included in the community. Considering that the number of users in the first community (2,458) was less than 11% of the total number of users (23,172) in the original network, the aggregation index of 0.73 was notably high. In the case of the second community, consisting of only 248 users, the average value of the aggregation index was even higher. This result suggests that each community indeed encompassed the users who shared interests in the given set of information objects. In addition, it also suggests that highly popular items did not assume a large portion of the union set of each community. In other words, relatively less popular items collectively defined a coherent set of shared interests in each community.

Overall, the results of the third phase indicated that the *m*-slice component analysis appeared to be successful in detecting communities in the network of *delicious.com* users. While the giant community of the original network did not break down easily, after removing weak ties iteratively, strongly knit communities emerged. Each of the three communities identified in the 28-slice of the original network showed distinct shared interests and a high level of coherence.

Chapter 8. Conclusion

This study was motivated by the unique characteristics of social bookmarking, being at the intersection of personal and social information spaces. While users of a social bookmarking site build their own bookmark collections, in doing so, they collectively and cumulatively weave the information space of the site as a whole. What are the implications of the novel possibility of looking at the personal information spaces of the wide variety of users? What patterns or structures emerge when personal information spaces of individual users intersect and overlap in an open space? Would this data be useful for improving our understanding of the social dimension of information access? What are the appropriate tools for investigating this large scale data where personal and social features are intertwined? With these broad motivational questions in mind, this study was conceived as the first step toward a better understanding of the phenomenon of social bookmarking and its theoretical and practical implications.

The overall goal of this study was to investigate the structure of the information space of a social bookmarking site in terms of the shared interests of its users. *Delicious.com* was chosen as the case in this study because it was one of the first social bookmarking sites and it has a large population of users. The main purpose of the study was two-fold: first, to see whether and how we can identify shared interest space(s) within the general information space; and second, to evaluate the applicability of the theories and methods developed in social network analysis to this end. This study was

carried out in three phases asking separate yet interrelated questions concerning the overall level of interest overlap, the structural patterns in the network of users connected by shared interests, and the communities of interest within the network. Each phase dealt with increasingly narrower and more specifically defined areas in the information space, adopting different methods.

The first phase aimed to characterize the information space of *delicious.com* as a whole, in terms of accumulation and overlap of bookmarking activities. The basic statistics of bookmarking activities along three axes – users, information objects, and time – were analyzed with three related datasets (*Recent*, *User History*, and *URL History*). The results showed that bookmarking activities are spread over a wide variety of information objects, with little overlap during a short period of time. However, the majority of information objects accrue multiple postings over time, raising the overall level of overlap (i.e., interest sharing). In all three datasets, the distribution of bookmark postings, showing either the popularity of information objects or the activities of users, turned out to be highly skewed. The heterogeneous distributions demonstrate the coexistence of personal and social characteristics. The long tail of less popular information objects in the popularity distribution indicates that users have diverse personal interests. On the other hand, the extremely popular information objects, especially ones that attracted attention quickly, suggest a sort of social contagion or viral spread. Individual collections of users (their personal information spaces) consist of a subset of unique selections and a subset of shared bookmarks, while their sizes and composition vary to a large extent. Overall, it was concluded that, while individual users of the site have a broad range of diverse interests, there is a certain level of overlap

and commonality, providing a ground for creating a network of users with shared interests.

In the second phase, the network of active users was created and explored. From the *Recent* dataset, a smaller set of users representing the most *active* group of users in the *current* user population of the site was selected, and an affiliation network consisting of the users in the new set and the information objects in their bookmark collections (3 month data) was constructed. The social network of users was then created by transforming the affiliation network. This network of *delicious.com* users presented common features of traditional social networks. Network properties commonly found in many other networks -- a highly heterogeneous degree distribution, the existence of the giant component, and the small-world effect characterized by high local clustering and short average path length -- were observed in this network. The most prominent factor defining this network was a high volume of connections, indicating that the active users of this site are well-connected based on their common possession of bookmarks. The level of connectivity and cohesion in this network suggests that the basic structural conditions for building an efficient information sharing mechanism are already in place and, even if individual users approach the site primarily for managing their personal collections of bookmarks, they can easily be embedded in the dense structure of shared interests with the small number of bookmarks they have in common with other users.

A further examination of the internal structure of the network was conducted with the *m*-slice analysis. In an *m*-slice, users are connected if they have at least *m* information objects in common, instead of a single item. By incrementally raising the threshold value, *m,* of the minimum shared bookmarks, sub-networks with increasingly

stronger connections emerged from the original network. Perhaps the most interesting finding of this study is the core-periphery structure uncovered in the *m*-slice analysis. The network had a strongly connected dense core, which did not break into parts even after the vast majority (more than 99%) of the connections in the network was taken out. The first major split of the core was observed at $m = 28$, where each and every remaining tie consisted of 28 or more shared information objects. The core periphery structure observed in this network was known to be a stable and effective structure for an information network. Note that each node in the network represents an individual user and his/her personal information space at the same time; the tie between a pair of users represents the intersection of their personal information spaces. Therefore, the structure of the network can be interpreted as depicting interrelated areas of interests within the site, as well as implicit relations among users. On the other hand, we need to keep in mind that, in the current interface of *delicious.com*, when a user adds a bookmark, she/he is presented with an option to 'view' other users who have bookmarked the same information object and to further explore those users' other bookmarks. Thus, the shared information objects between two users may indeed serve as a potential channel of interaction or influence. In any case, it is promising that the structure of this network resembles one that is commonly found in well-established social systems.

The third phase was a small exploratory analysis of the communities identified by the *m*-slice technique. The three communities detected at the 28-slice of the network were examined by looking at the information objects constituting the links within each community. The results indicated that each of the three largest communities had a distinct theme defining the interests of the community members, at a high level of

coherence. We can conclude that the information objects that were involved in creating links among members of each community were related to the respective topic area and, thus, brought together the users who shared interests in the given set of information objects.

The results of this study have implications for further research and also have practical applications. First of all, the results showed that users of social bookmarking can be connected to form a network which enables further investigation of social mechanisms of information sharing in this new information environment. In general, one of the main points of representing a system as a network and studying its structure is to understand how the structural features affect behaviors and/or outcomes of the system constituents. With the network representing shared interests of users, it would be possible to look at the behaviors of users and information objects in relation to the observed structure. The communities of interest that were identified in this study lay the groundwork for a comparative study of different communities. Tag usage or tagging behavior within and across communities of interest, for instance, can be compared to suggest more productive ways of exploiting tag data.

At the beginning of this dissertation, we discussed a conceptual picture of the information space of a social bookmarking site consisting of interconnected 'local' interest spaces. It was posited that, if coherent regions of shared interests can be identified in the network representing the information space of a social bookmarking site, such 'local' interest spaces, or communities, provide users with a dynamic local 'view' of the overall information space, allowing him/her to take advantage of others' efforts. The structure of the network analyzed in this study was in accordance with the

conceptual picture. Schek et al (2000) pointed out that a problem in the current search environment is that users with different sets of interests are provided with the same 'view' of a global information space. They argued for a dynamically constructed information space that is tailored to the personal context of a particular user. The results of this study suggest the possibility of creating a flexible and adaptable information environment for users, building upon social bookmarking. Björneborn (2004) introduced the concept of 'exploratory capability' that had been originally discussed as an evaluation criterion for information retrieval (Doyle, 1963) into his study of the small world network of an academic website. According to his definition, the exploratory capability of an information system includes the possibility for users to navigate and access a broad range of information objects with a structure fostering serendipity and diffusion. The small world effect and the core-periphery structure of the network support the notion of exploratory capability.

Although the findings this study shed light on the question of shared interests represented by bookmarking behaviors, this study has its limitations. This study took a single site, *delicious.com*, as a case, so the generalizability of the result is limited to the site. In fact, as discussed in the previous chapter, while the level of shared interests in this particular site was high enough to build a dense network of users, a study conducted with different social bookmarking sites (CiteULike and Connotea) reported disparate findings regarding interest sharing among users. Although some possible explanations on what might have caused the differences were discussed, the conditions necessary for a social bookmarking site to realize its potential as a social information tool still remain to be studied.

The sampling strategy adopted in this study and the boundary specifications used to select a sample of the network of users further require that care must be taken when generalizing the results of this study. The network of *delicious.com* users constructed in this study was intended to represent the active part of the information space, with its boundary definition based on activity criteria. Clearly, this network may not be representative of the entire information space. The exceptional volume and strength of connections observed in the present network of active users probably has to do, at least in part, with the boundary specification. Therefore, a network constructed with a random sample of users, for example, may reveal a different structure. In order to get a more complete picture of the entire information space, the structure of the network studied here may be compared and contrasted with other network(s) using different boundary specification(s), in a follow-up comparative study.

Having defined the set of users, the scope of the bookmarking records to be used for establishing connections among them was also reduced, mainly due to technical and computational limitations. While the number of users in the second phase dataset was only 23,238, the total number of bookmark postings made by them was 25,559,506 when the entire history of each was considered. There are network analysis tools that can handle a large scale dataset, and most of those tools, including the ones used in this study (*Pajek* and the *igraph* package in R), support the transformation of a two-mode network (consisting of users and information objects, in this case) into an one-mode network. However, a transformation of this dataset could not be done using the existing tools due to the excessive number of entities in the second mode (information objects). Even when the number of information objects was reduced substantially by applying a

shorter time-window (12 months, 6 months, and 3 months), this was still the case. Therefore, for each time window, a one-mode social network of users had to be constructed through a series of steps, involving separate database tables and scripts, outside those tools. For the sake of computational efficiency, only the network built with the 3 month window was further analyzed. As a consequence, the structure of the network only reflects user interests within this short period of time. While it may be argued that the network reflects the current interests of users, data with a longer time window might have offered more reliable indicators of shared interests. This kind of limitation can be overcome with increased computing capacity in future studies.

In conclusion, while much remains to be studied to further our understanding of the social bookmarking phenomenon, the main purpose of this study -- understanding the information space in terms of shared interests by adopting network approaches -- was achieved. The overall level of interest sharing in *delicious.com* appeared to be sufficient to construct a network of its users based on their common bookmarks. The network of *active* users of the site turned out to be a dense network exhibiting the small world effect, suggesting that interest spaces in this site are locally coherent and globally connected. In addition, the dense core of the network suggests that, to a large extent, there is a set of central themes that binds the users of this site together. The communities identified with the structural definition of minimum tie strength turned out to have distinctive and coherent shared interests. The findings of the three phases of this study, in combination, suggest the possibility of identifying communities of shared interests within a social bookmarking site, by adopting network analysis methods. Social bookmarking data can be used to create a network of users where various dimensions of

user behaviors both at the personal level and the community level can be observed. The structure of the network allows a further investigation of social mechanisms of interest sharing in this new information environment.

While most discussions and empirical studies on social bookmarking have concentrated on one aspect of this phenomenon -- the value or use of the tags assigned by users -- this study contributed to the body of research by addressing a broader problem: how the information space of a social bookmarking site can be understood with regard to the shared interests of its users. The approach taken in this study to identifying communities of interests within the large information space will also be useful for pursuing the research problems related to the tags that have been the center of attention in the research community.

References

Aaronson, S. (1975). The footnotes of science. *Mosaic, 6*(2), 22-27.

Abrams, D., Baecker, R., & Chignell, M. (1998). *Information archiving with bookmarks: Personal web space construction and organization.* Paper presented at the Human factors in computing systems: CHI 98 conference proceedings.

Adomavicius, G., Sankaranarayanan, R., Sen, S., & Tuzhilin, A. (2005). Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems, 23*(1), 103-145.

Aggarwal, C. C., Wolf, J. L., Wu, K.-L., & Yu, P. S. (1999, August 15 - 18). *Horting hatches an egg: a new graph-theoretic approach to collaborative filtering.* Paper presented at the the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, California, United States.

Albert, R., & Barabasi, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*, 47-97.

Allen, M.P. (1982). The identification of interlock groups in large corporate networks: Convergent validation using divergent techniques, *Social Networks* 4, 349-366.

Anderson, C. (2006). *The Long Tail: Why the Future of Business Is Selling Less or More.* New York: Hyperion.

Anderson, C. J., Wasserman, S., & Crouch, B. (1999). A p* primer: logit models for social networks. *Social Networks, 21*(1), 37-66.

Balabanovic, M., & Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM, 40*(3), 66 - 72.

Barabási, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications, 311*(3-4), 590-614.

Barreau, D. K. (1995). Context as a factor in personal information management systems. *Journal of the American Society for Information Science, 46*(5), 327-339.

Barreau, D., & Nardi, B. A. (1995). Finding and reminding: file organization from the desktop. *SIGCHI Bull., 27*(3), 39-43.

Barsalou, L. W. (1982). Context-independent and context-dependent information in concepts. *Memory & Cognition, 10*, 82-93.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition, 11*, 211-227.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 629-654.

Barsalou, L. W. (1990). On the indistinguishability of exemplar memory and abstraction in category representation. In T. K. Srull & R. S. Wyer (Eds.), *Advances in social cognition, Volume III: Content and process specificity in the effects of prior experiences* (pp. 61-88). Hillsdale, NJ: Lawrence Erlbaum Associates.

Barsalou, L. W. (1991). Deriving categories to achieve goals. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 1-64). San Diego, CA: Academic Press.

Barsalou, L. W., & Sewell, D. (1984). *Constructing representations of categories from different points of view.* Atlanta, GA: Emory University.

Barsky, N. P. (1999). A core/periphery structure in a corporate budgeting process. *Connections*, 22(2), 22–29.

Basu, C., Hirsh, H., & Cohen, W. (1998). *Recommendation as Classification: Using Social and Content-Based Information in Recommendation.* Paper presented at the the 15th National Conference on Artificial Intelligence (AAAI '98).

Bates, M. J. (1986). Subject Access in Online Catalogs: A Design Model. *Journal of the American Society for Information Science, 37*(6), 357-376.

Bates, M. J. (1998). Indexing and Access for Digital Libraries and the Internet: Human, Database, and Domain Factors. *Journal of the American Society for Information Science, 49*(13), 1185-1205.

Bates, M. J. (2002). Speculations on browsing, directed searching, and linking in relation to the bradford distribution. *Proc. CoLIS*, 4**,** 137–150.

Bates, Marcia J. (1977) "System meets user: Problems in matching subject search terms." *Information Processing and Management,* 13(6), 367-375

Bayer, A. E., Smart, J. C., & McLaughlin, G. W. (1990). Mapping Intellectual Structure of a Scientific Subfield through Author Cocitations. *Journal of the American Society for Information Science, 41*(6), 444-452.

Bearman, P., & Everett, K. (1993). The Structure of Social Protest. *Social Networks, 15*(2), 171-200.

Begelman, G., Keller, P., & Smadja, F. (2006, May 22--26). *Automated Tag Clustering: Improving search and exploration in the tag space.* Paper presented at the 15th International World Wide Web Conference (WWW2006), Edinburgh, UK.

Berkowitz, S. D. (1982). *An introduction to structural analysis: The network approach to social research*. Toronto: Butterworths.

Berlin, B. (1978). Ethnobiological classification. In E. H. Rosch & B. Lloyd (Eds.), *Cognition and categorization* (pp. 9-26). Hillsdale, NJ: Erlbaum.

Berlin, B. (1992). *Ethnobiological classification: principles of categorisation of plants and animals in traditional societies.* Princeton, New Jersey: Princeton University Press.

Berlin, B., Breedlove, D. E., & Raven, P. H. (1974). *Principles of Tzeltal Plant Classification*. New York: Academic.

Bernard, H. R., Killworth, P., & Sailer, L. (1981). Summary of Research on Informant Accuracy in Network Data and on the Reverse Small World Problem. *Connections, 4*(2), 11-25.

Bjorneborn, L. (2004). *Small-World Link Structures across an Academic Web Space - a Library and Information Science Approach.* Royal School of Library and Information Science, Copenhagen.

Bonacich, P. (1972). Factoring and weighting approaches to clique identification. *Journal of Mathematical Sociology, 2*, 113-120.

Borgatti, S. P., & Cross, R. (2003). A Relational View of Information Seeking and Learning in Social Networks. *Management Science, 49*(4), 432-445.

Borgatti, S. P., & Everett, M. G. (1997). Network analysis of 2-mode data. *Social Networks, 19*(3), 243-269.

Borgatti, S. P., & Everett, M. G. (1999). Models of corerperiphery structures. *Social Networks, 21*, 375-395.

Borgatti, S. P., & Foster, P. C. (2003). The Network Paradigm in Organizational Research: A Review and Typology. *Journal of Management, 29*(6), 991-1013.

Borgman, C. L. (1986). Why Are Online Catalogs Hard to Use? Lessons Learned From Information-Retrieval Studies. *Journal of the American Society for Information Science, 37*(6), 387-400.

Borgman, C. L., & Furner, J. (2002). Schoarly Communication and Bibliometrics. *Annual Review of Information Science and Technology, 36*(1), 2-72.

Bowker, G. C., & Star, S. L. (1999). *Sorting things out : classification and its consequences*. Cambridge, Mass.: MIT Press.

Brass, D. J. (1995). A social network perspective on human resources management. *Research in Personnel and Human Resources Management, 13*, 39-79.

Brass, D. J., Galaskiewicz, J., Greve, H. R., & Tsai, W. (2004). Taking stock of networks and organizations: A multilevel perspective. *Academy of Management Journal, 47*(6), 795-817.

Breiger, R. (2003). Emergent themes in social network analysis: Results, challenges, opportunities. In R. Breiger, K. M. Carley & P. Pattison (Eds.), *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers* (pp. 19-34). Washinton, D. C.: The National Academies Press.

Breiger, R. L. (1974). The duality of persons and groups. *Social Forces, 53*(1), 181-190.

Breiger, R., Carley, K. M., & Pattison, P. (Eds.). (2003). *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*. Washinton, D. C.: The National Academies Press.

Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks, 30*(1-7), 107-117.

Brooks, C. H., & Montanez, N. (2006). Improved annotation of the blogosphere via autotagging and hierarchical clustering. *Proceedings of WWW 2006*, 625-632.

Brown, R. (1965). *Social Psychology*. New York: Free Press.

Bruner, J. S., Goodnow. J., Austin, G. A. (1956). *A study of thinking*. New York; John Wiley & Sons.

Brynjolfsson E, Hu Y, Smith MD. (2006). From niches to riches: Anatomy of the long tail. *MIT Sloan Management Review, 47*(4), 67-.

Burns, E., & Webber, C. (2009). When Harry met Bella. *School Library Journal, 55*(8), 26-29.

Burt, R. S. (1983). Studying Status/Role-Sets Using Mass Surveys. In R. S. Burt & M. J. Minor (Eds.), *Applied Network Analysis*. Beverly Hills, CA: Sage Publication.

Calado, P., Cristo, M., Goncalves, M. A., Moura, E. S. d., Ribeiro-Neto, B., & Ziviani, N. (2006). Link-based similarity measures for the classification of Web documents. *Journal of the American Society for Information Science and Technology, 57*(2), 208-221.

Campbell, D. G. (2006). A phenomenological framework for the relationship between the Semantic Web and user-centered tagging systems. Proceedings of the 17th SIG Classification Research Workshop, Austin, Texas.

Carlyle, A. (1999). User categorisation of works: Toward improved organisation of online catalogue displays. *Journal of Documentation, 55*(2), 184-208.

Carlyle, A. (2001). Developing organized information displays for voluminous works: a study of user clustering behavior. *Information Processing & Management, 37*(5), 677-699.

Carrington, P., Scott, J., & Wasserman, S. (Eds.). (2005). *Models and Methods in Social Network Analysis*. New York: Cambridge University Press.

Case, D. O. (1991). Conceptual organization and retrieval of text by historians: The role of memory and metaphor. *Journal of the American Society for Information Science, 42*(9), 657-668.

Case, D. O., & Higgins, G. M. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science, 51*(7), 635-645.

Chakrabarti, S., Dom, B. E., & Indyk, P. (1998). Enhanced hypertext classification using hyper-links. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 307-318.

Chakrabarti, S., Dom, B. E., Kumar, S. R., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D., & Kleinberg, J. (1999a). Mining the Web's link structure. *Computer*, 32(8), 60-67.

Chalmers, M. (1999). Comparing information access approaches. *Journal of the American Society for Information Science, 50*(12), 1108-1118.

Chi, E. H. and Mytkowicz, T. (2007). Understanding Navigability of Social Tagging Systems. In Proceedings of CHI'07.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Choy, S.-O., & Lui, A. K. (2006). Web Information Retrieval in Collaborative Tagging Systems. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 352-355): IEEE Computer Society.

Christiaens, S. (2006). Metadata Mechanisms: From Ontology to Folksonomy … and Back. *Lecture Notes in Computer Science*, 4277, 199-207.

Chrysikou, E. G. (2006). When Shoes Become Hammers: Goal-Derived Categorization Training Enhances Problem-Solving Performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(4), 935-942.

Chubin, D. E., & Moitra, S. (1975). Content analysis of references: Adjunct of alternative to citation counting. *Social Studies of Science, 5*, 423-44

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E, 70*, 066111.

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., & Sartin, M. (1999, August

19). *Combining content-based and collaborative filters in an online newspaper.* Paper presented at the ACM SIGIR Workshop on Recommender Systems.

Cochrane, P. A., & Markey, K. (1983). Catalog Use Studies - Since the Introduction of Online Interactive Catalogs: Impact on Design for Subject Access. *Library & Information Studies Research, 5*, 337-363.

Cohen, B., & Murphy, G. L. (1984). Models of Concepts. *Cognitive Science, 8*, 27-58.

Cole, C., & Leide, J. E. (2006). A Cognitive Framework for Human Information Behavior: The Place of Metaphor in Human Information Organizing Behavior. In A. Spink & C. Cole (Eds.), *New Directions in Human Information Behavior* (pp. 171-202). Dordrecht: Springer.

Cole, J. R. (2000). A Short History of the Use of Citations as a Measure of the Impact of Scientific and Scholarly Work. In B. Cronin & H. B. Atkins (Eds.), *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield* (pp. 281-300). Medford, NJ: Information Today Inc.

Cole, J. R., & Cole, S. (1971). Measuring the quality of sociological research: Problems in the use of the Science Citation Index. *The American Sociologist, 6*, 23-29.

Cole, J. R., & Cole, S. (1973). *Social Stratification in Science*. Chicago, IL: University of Chicago Press.

Condliff, M.K., Lewis, D., Madigan, D., Posse, Bayesian, C. (1999). Mixed-effects Models for Recommender Systems. *Proceedings of 1999 ACM SIGIR Workshop on Recommender Systems*.

Cooper, L. Z. (2004). The Socialization of Information Behavior: A Case Study of Cognitive Categories for Library Information. *Library Quarterly, 74*, 299-336.

Cooper, W.S. (1969). Is interindexer consistency a hobgoblin? *American Documentation, 20*, 268-278

Coser, L. (1977). *Masters of Sociological Thought: Ideas in Historical and Social Context* (2nd ed.). New York: Harcourt, Brace and Jovanovich.

Cronin, B. (1981). The need for a theory of citing. *Journal of Documentation, 37*(1), 16-24.

Cronin, B. (1984). *The citation process: The role and significance of citation in scientific communication.* London: Taylor Graham.

Cronin, B., & Atkins, H. B. (Eds.). (2000). *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*. Medford, NJ: Information Today Inc.

Cummings, J. and Cross, R. (2003). Structural properties of work groups and their consequences for performance. *Social. Networks*, 25(3), 197-210.

Custers, E. J. F. M., Regehr, G., & Norman, G. R. (1996). Mental representations of medical diagnostic knowledge: A review. *Academic Medicine, 71*(10), 555-561.

Cutter, C. A. (1876). *Rules for a printed dictionary catalog*. Washington: U. S. G. P. O.

Davenport, E., & Cronin, B. (2000). The citation network as a prototype for representing trust in virtual environments. In B. Cronin & H. B. Atkins (Eds.), *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield* (pp. 517-534). Medford, NJ: Information Today Inc.

Davis, G. F., & Greve, H. R. (1997). Corporate elite networks and governance changes in the 1980s. *American Journal of Sociology, 103*, 1-37.

Davis, G. F., Yoo, M., and Baker, W. E. (2001). *The Small World of the Corporate Elite*, preprint, University of Michigan Business School, Ann Arbor, MI.

de Mey, M. (1982). *The Cognitive Paradigm: Cognitive Science, A Newly Explored Approach to the Study of Cognition Applied in an Analysis of Science and Scientific Knowledge*. Dordrecht, Holland: D. Reidel.

Dieberger, A., Dourish, P., Hook, K., Resnick, P., & Wexelblat, A. (2000). Social navigation: techniques for building more usable systems. *interactions, 7*(6), 36-45.

Doyle, L.B. (1963). Is Relevance an adequate criterion in retrieval system evaluation? Proceedings of the American Documentation Institute; 26th annual meeting (pp. 199–200). Chicago, IL.

Doreian, P., & Woodard, K. L. (1994). Defining and locating cores and boundaries of social networks. *Social Networks, 16*(4), 267-293.

Edge, D. O. (1977). Why I am not a co-citationist. *Society for Social Studies of Science Newsletter, 2*, 13-19.

Edge, D. O. (1979). Quantitative measures of communication in science: A critical review. *History of Science, 17*, 102-134.

Emirbayer, M. (1997). Manifesto for a Relational Sociology. *American Journal of Sociology, 103*(2), 281-317.

Emirbayer, M., & Goodwin, J. (1994). Network analysis, Culture, and the Problem of Agency. *American Journal of Sociology, 99*(6), 1411-1454.

Ennis, J. G. (1992). The Social Organization of Sociological Knowledge: Modeling the Intersection of Specialties. *American Sociological Review, 57*(2), 259-265.

Erdös, P. and Rényi. A. (1959). On random graphs. *PublicationesMathematicae*, 6, 290–297.

Everett, K., & Borgatti, S. P. (2005). Extending Centrality. In P. Carrington, J. Scott & S.

Wasserman (Eds.), *Models and Methods in Social Network Analysis* (pp. 57-76). New York: Cambridge University Press.

Everett, K., & Borgatti, S. P. (2005). Extending Centrality. In P. Carrington, J. Scott & S. Wasserman (Eds.), *Models and Methods in Social Network Analysis* (pp. 57-76). New York: Cambridge University Press.

Faust, K. (1997). Centrality in affiliation networks. *Social Networks, 19*(2), 157-191.

Faust, K., Willert, K., Rowlee, D., & Skvoretz., J. (2002). Scaling and Statistical Models for Affiliation Networks: Patterns of Participation Among Soviet Politicians during the Breszhnev Era. *Social Networks, 24*, 231-259.

Felcher, E. M., Malaviya, P., & McGill, A. L. (2001). Categorization in, within, and across Category Judgments. *Psychology & Marketing, 18*(8), 865-887.

Fidel, Raya (1985). "Individual variability in online searching behavior." In Carol A. Parkhurst (ed.), *ASIS '85: Proceedings of the American Society of Information science 48th Annual Meeting* (October 20-24) *Las Vegas, Nevada*. White Plains, NY: Knowledge Industry Publications, 69-72.

Field, S., Frank, K. A., Schiller, K., Riegle-Crumb, C., & Muller, C. (2006). Identifying positions from affiliation networks: Preserving the duality of people and events. *Social Networks, 28*, 97-123.

Fillmore, C. (1982). Towards a descriptive framework for spatial deixis. In R. J. Jarvella & W. Klein (Eds.), *Speech, place and action: studies in deixis and related topics* (pp. 31-59). Chichester: Wiley.

Fortunato, S., Latora, V., & Marchiori, M. (2004). Method to find community structures based on information centrality. *Physical Review E, 70*, 056104.

Frank, O. (2005). Network Sampling and Model Fitting. In P. Carrington, J. Scott & S. Wasserman (Eds.), *Models and Methods in Social Network Analysis* (pp. 31-56). New York: Cambridge University Press.

Franks, B. (1995). Sense Generation: A "Quasi-Classical" Approach to Concepts and Concept Combination. *Cognitive Science, 19*, 441-505.

Freeman, L. C. (1979). Centrality in social networks: conceptual clarification. *Social Networks, 1*, 35-41.

Freeman, L. C., & White, D. R. (1993). Using Galois Lattices to Represent Network Data. *Sociological Methodology, 23*, 127-146.

Freeman, L. C., White, D. R., & Romney, A. K. (Eds.). (1989). *Research methods in social network analysis*. Fairfax, Va.: George Mason University Press.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM, 30*(11), 964 - 971.

Garfield, E. (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science, 122*(3159), 108-111.

Garfield, E. (1974). The citation index as a subject index (Editorial). *Essays of an Information Scientist, 2*(62-64).

Garfield, E. (1988). Announcing the SCI Compact Disc Edition: CD-ROM Gigabyte Storage Technology, Novel Software, and Bibliographic Coupling Make Desktop Research and Discovery a Reality. *Essays of an Information Scientist, 11*, 160-170.

Garton, L., Haythornthwaite, C., & Wellman, B. (1997). Studying Online Social Networks. *Journal of Computer-Mediated Communication, 3*(1).

Garton, L., Haythornthwaite, C., & Wellman, B. (1999). Studying on-line social networks. In S. Jones (Ed.), *Doing Internet research : critical issues and methods for examining the Net* (pp. 75-105). Thousand Oaks, Calif.: Sage Publications.

Gilbert, G. N. (1977). The transformation of research findings into scientific knowledge. *Social Studies of Science, 7*, 113-122.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, 99*(12), 7821-7826.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. B. (1992). Using Collaborative Filtering to Weave an Information Tapestry. *Communications of the ACM, 35*(12), 61-70.

Golder, S. A. and B. A. Huberman (2005). The Structure of Collaborative Tagging Systems. Information Dynamics Lab, HP Labs.

Golder, S. A. and B. A. Huberman (2006). "The Structure of Collaborative Tagging Systems." Journal of Information Science 32(2): pp. 198-208.

Gottlieb, L., & Dilevko, J. (2001). User preferences in the classification of electronic bookmarks: Implications for a shared system. *Journal of the American Society for Information Science and Technology, 52*(7), 517-535.

Gottlieb, L., & Dilevko, J. (2003). Investigating how individuals conceptually and physically structure file folders for electronic bookmarks: The example of the financial services industry. *Journal of the American Society for Information Science and Technology, 54*(2), 124-139.

Granovetter, M. (1973). The Strength of Weak Ties. *American Journal of Sociology, 78*(6), 1360-1380.

Granovetter, M. (1976). Network Sampling: Some First Steps. *American Journal of Sociology,*

81(6), 1287-1303.

Granovetter, M. (1982). The strength of weak ties: A network theory revisited. In P. V. Marsden & N. Lin (Eds.), *Social Structure and Network Analysis* (pp. 105-130). Beverly Hills: Sage Publishers.

Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological Theory, 1*, 201-233.

Granovetter, M. S. (1974). *Getting a job*. Cambridge, MA: Harvard University Press.

Grudin, J. (1994). Groupware and social dynamics: eight challenges for developers. *Communications of the ACM, 37*(1), 92-105.

Guy, M. and E. Tonkin (2006). "Folksonomies: Tidying up Tags?" D-Lib Magazine,12(1).

Halpin, H., Robu, V., & Shepherd, H. (2007). The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web* (pp. 211-220). Banff, Alberta, Canada: ACM Press.Kipp, M. E. I. and D. G. Campbell (2006). Patterns and Inconsistencies in Collaborative Tagging Systems: An Examination of Tagging Practices. Proceedings Annual General Meeting of the American Society for Information Science and Technology, Austin, Texas (US).

Harter, S. P. (1992). Psychological Relevance and Information Science. *Journal of the American Society for Information Science, 43*(9), 602-615.

Heit, E. (1997). Knowledge and concept learning. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 7-41). London: Psychology Press.

Hellekson, K. and Busse, K., editors (2006). Fan Fiction and Fan Fiction Communities in the Age of the Internet. McFarland and Company.

Heller, K. (1989). The return to community. *American Journal of Community Psychology*, 17(1), 1-15.

Hill, W. C., Hollan, J. D., Wroblewski, D., & McCandless, T. (1992). Edit wear and read wear. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 3-9). Monterey, California, United States: ACM.

Hotho, A., R. Jäschke, C. Schmitz and G. Stumme (2006). Information Retrieval in Folksonomies: Search and Ranking. Proceedings of the 3rd European Semantic Web Conference, Budva, MontenegroLNCS, Springer.

Huang, Z., Chen, H., & Zeng, D. (2004). Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Transactions on Information Systems, 22*(1), 116-142.

Huang, Z., Chung, W., Ong, T.-H., & Chen, H. (2002, July 14 - 18). *A Graph-based*

Recommender System for Digital Library. Paper presented at the the 2nd ACM/IEEE-CS joint conference on Digital libraries, Portland, Oregon, USA.

Huisman, M., & Duijn, M. A. J. v. (2005). Software for Social Network Analysis. In P. Carrington, J. Scott & S. Wasserman (Eds.), *Models and Methods in Social Network Analysis* (pp. 270-316). Cambridge ; New York: Cambridge University Press.

Hummon, N. P., & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks, 11*(1), 39-63.

Jacob, E. K. (2004). Classification and categorization: a difference that makes a difference. *Library Trends, 52*(3), 515-540.

Jacoby, J., and Slamecka, V. (1962). *Indexer Consistency under Minimal Conditions*. Bethesda, MD: Documentation, Inc.

Ji, A., Yeon, C., Kim, H., & Jo, G. (2007). Collaborative Tagging in Recommender Systems. *Lecture Notes In Computer Science, 4830*, 377.

Jiao, Y., & Cao, G. (2007). *A Collaborative Tagging System for Personalized Recommendation in B2C Electronic Commerce.* Paper presented at the International Conference on Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007.

Johnson, S. (2001). *Emergence: The connected life of ants, brains, cities, and software.* New York: Scribner.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93, 136-153.

Kalman, M. E., Monge, P., Fulk, J., & Henino, R. (2002). Motivations to resolve communication dilemmas in database-mediated collaboration. *Communication Research, 29*(2), 125-154.

Kautz, H., Selman, B., & Shah, M. (1997). Referral Web: combining social networks and collaborative filtering. *Communications of the ACM, 40*(3), 63-65.

Keller, R. M., Wolfe, S. R., Chen, J. R., Rabinowitz, J. L., & Mathe, N. (1997). A bookmarking service for organizing and sharing URLs. . *Computer Networks and ISDN Systems, 29*, 1103-1114.

Kelly, D. (2004). *Understanding implicit feedback and document preference: a naturalistic study.* Unpublished dissertation, Rutgers, The State University of New Jersey.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*, 10-25.

Kim. J., Oard, D. W., & Romanik, K. (2000). Using implicit feedback for user modeling in internet and intranet searching. University of Maryland CLIS Technical Report 00-01.

Kleinberg, J. M. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM, 46*(5), 604-632.

Knoke, D., & Kuklinski, J. H. (1982). *Network analysis*. Beverly Hills, Calif.: Sage Publications.

Konstan, J., Miller, B. N., Malt, D., Herlocker, J., Gordon, L. R., & Riedl, J. (1997). GroupLens: applying collaborative filtering to Usenet news. *Communications of the ACM, 40*(3), 77-87.

Krikelas, J. (1972). Catalog Use Studies and Their Implications. *Advances in Librarianship, 3*, 195-220.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Kumar, S. R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Trawling emerging cyber-communities automatically. *Proceedings of the 8th International World Wide Web Conference.*

Kumbasar, E., Romney, A. K., & Batchelder, W. H. (1994). Systematic Biases in Soical Perception. *American Journal of Sociology, 100*, 477-505.

Kwasnik, B. H. (1989). The influence of context on classificatory behavior. Ph.D. Dissertation, Rutgers University.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind.* Chicago: University of Chicago Press.

Lambiotte, R.; Ausloos, M.(2006). Collaborative tagging as a tripartite network. Lecture Notes in Computer Science, 3993 (2006) 1114-1117

Laumann, E. O., Marsden, P. V., & Prensky, D. (1989). The Boundary Specification Problem in Network Analysis. In L. C. Freeman, D. R. White & A. K. Romney (Eds.), *Research methods in social network analysis* (pp. 61-87). Fairfax, Va.: George Mason University Press.

Leydesdorff, L. (1998). Theories of citation? *Scientometrics, 43*(1), 5-25.

Leydesdorff, L., & Amsterdamska, O. (1990). Dimensions of Citation Analysis. *Science, Technology, & Human Values, 15*(3), 305-335.

Lin, N. (1982). Social resources and instrumental action. In P. V. Marsden & N. Lin (Eds.), *Social Structure and Network Analysis* (pp. 131-145). Beverly Hills: Sage Publishers.

Ma'ayan, A., Jenkins, S. L., Neves, S., Hasseldine, A., Grace, E., Dubin-Thaler, B., et al. (2005). Formation of regulatory patterns during signal propagation in a mammalian cellular network. *Science, 309*, 1078-1083.

Macgregor, G. and McCulloch, E. (2006) Collaborative tagging as a knowledge organisation and

resource discovery tool. Library Review, 55 (5). pp. 291-300.

MacRoberts, M. H., & MacRoberts, B. R. (1986). Quantitative measures of communication in science: A study of the formal level. *Social Studies of Science, 16*, 151-172.

MacRoberts, M. H., & MacRoberts, B. R. (1987). Another test of the normative theory of citing. *Journal of the American Society for Information Science, 38*(4), 305-306.

Malone, T. W. (1983). How do people organize their desks?: Implications for the design of office information systems. *ACM Transactions on Information Systems, 1*(1), 99-112.

Markey, K. (1984). Interindexer consistency tests: A literature review and report of a test of consistency in indexing visual materials. *Library & Information Science Research, 6*, 155-177.

Marlow, C., M. Naaman, d. boyd and M. Davis (2006). Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead. WWW 2006, Edinburgh, UKACM Press.

Marsden, P. V. (1990). Network Data and Measurement. *Annual Review of Sociology, 16*, 435-463.

Marsden, P. V. (2005). Recent Development in Network Measurement. In P. Carrington, J. Scott & S. Wasserman (Eds.), *Models and Methods in Social Network Analysis* (pp. 8-30). New York: Cambridge University Press.

Marsden, P., & Campbell, K. E. (1984). Measuring tie strength. *Social Forces*, *63*, 482-501.

Marwell, G., Oliver, P. E., & Prahl, R. (1988). Social Networks and Collective Action: A Theory of the Critical Mass. III. *The American Journal of Sociology, 94*(3), 502-534.

Mathes, A. (2004) Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Available at http://www.adammathes.com/academic/computer-mediatedcommunication/folksonomies.html

McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science, 41*(6), 433-443.

McMillan, D. W. and D. M. Chavis (1986). Sense of community: A definition and theory. *Journal of Community Psychology*, 14(1), 6-23.

McPherson, J. M. (1982). Hypernetwork sampling: Duality and differentiation among voluntary organizations. *Social Networks, 3*(4), 225-249.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*, 254-278.

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology, 32*(1), 49-96.

Mika, P. (2005). *Ontologies are us: A unified model of social networks and semantics.* Paper presented at the The SemanticWeb - ISWC 2005, 4th International SemanticWeb Conference (November 6-10), Galway, Ireland.

Milgram, S. (1967). The small world problem. *Psychology Today*, 1(1), 61-67.

Millen, D. R. & Feinberg, J. (2006). Using social tagging to improve social navigation. In *Workshop on the Social Navigation and Community based Adaptation Technologies*, 2006.

Mirza, B. J.,Keller, B. J., and Ramakrishnan, N. (2003). Studying Recommendation Algorithms by Graph Analysis. *Journal of Intelligent Information Systems*, 20(2), 131–160.

Mirza, B.J. (2001). Jumping Connections: A Graph-Theoretic Model for Recommender Systems. Master's thesis, Virginia Tech. Available at http://scholar.lib.vt.edu/theses/available/etd-02282001-175040/.

Mische, A., & Pattison, P. (2000). Composing a civic arena: Publics, projects, and social settings. *Poetics, 27*(2-3), 163-194.

Moravscsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science, 5*, 86-92.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289-316.

Nadel, S.F., 1957. The Theory of Social Structure. Cohen and West, London.

Newman, M. E. J. (2000). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America, 98*(2), 404-409.

Newman, M. E. J. (2001a). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E, 64, art. no.* 016131.

Newman, M. E. J. (2001b). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E, 64, art. no.* 016132.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review, 45*, 167--256.

Newman, M. E. J. (2004a). Detecting community structure in networks. *The European Physical Journal B, 38*(2), 321-330.

Newman, M. E. J. (2004b). Fast algorithm for detecting community structure in networks. *Physical Review E, 69*, 066133.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*, 026113.

Nooy, W. d., Mrvar, A., & Batagelj, V. (2005). *Exploratory Social Network Analysis with Pajek*. Cambridge ; New York: Cambridge University Press.

Olson, H. A. (2002). The power to name : locating the limits of subject representation in libraries. Dordrecht, Netherlands: Kluwer Academic.

Paolillo, J. C., & Penumarthy, S. (2007). The Social Structure of Tagging Internet Video on del.icio.us. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences* (pp. 85): IEEE Computer Society.

Pattison, P., & Wasserman, S. (1999). Logit models and logistic regressions for social

Perugini, S., Goncalves, M. A., & Fox, E. A. (2004). Recommender Systems Research: A Connection-Centric Survey. *Journal of Intelligent Information Systems, 23*(2), 107 - 143.

Popescul, A., Ungar, L. H., Pennock, D. M., & Lawrence, S. (2001). Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*.

Pritchard, A. (1969). Statistical bibliography or bibliometrics? *Journal of Documentation, 25*, 348-349.

Quintarelli, E. (2005). Folksonomies: power to the people. ISKO Italy-UniMIB meeting, Milan, ItlayRetrieved February 18, 2007 from http://www.iskoi.org/doc/folksonomies.htm.

Ratneshwar, S., Barsalou, L. W., Pechmann, C., & Moore, M. (2001). Goal-Derived Categories: The Role of Personal and Situational Goals in Category Representations. *Journal of Consumer Psychology, 10*(3), 147-157.

Ratneshwar, S., Pechmann, C., & Shocker, A. D. (1996). Goal-Derived Categories and the Antecedents of Across-Category Consideration. *Journal of Consumer Research, 23*(3), 240-250.

Resnick, P., & Varian, H. R. (1997). Recommender Systems. *Communications of the ACM, 40*(3), 56-58.

Robins, G., Pattison, P., & Wasserman, S. (1999). Logit models and logistic regressions for social networks: III. Valued relations. *Psychometrika, 64*(3), 371 - 394.

Rosch, E. H. (1975). Cognitive reference points. *Cognitive Psychology, 7*, 532-547.

Rosch, E. H. (1978). Principles of categorization. In E. H. Rosch & B. Lloyd (Eds.), *Cognition and categorization* (pp. 28-49). Hillsdale, NJ: Erlbaum.

Rosch, E. H., & Lloyd, B. (Eds.). (1978). *Cognition and categorization*. Hillsdale, NJ: Erlbaum.

Rosch, E. H., & Mervis, C. B. (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Psychology, 7*, 573-605.

Rosch, E. H., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382-352.

Rosch, E. H., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance, 2*, 491-502.

Roth, E. M., & Shoben, E. J. (1981). The effect of context on the structure of categories Psychology. *Cognitive Psychology, 15*, 346-378.

Rucker, J., & Polanco, M. J. (1997). Siteseer: personalized navigation for the Web. *Communications of the ACM, 40*(3), 73-76.

Salton, G. (1971). Automatic indexing using bibliographic citations. *Journal of Documentation, 27*, 98-110.

Santos-Neto, E., Condon, D., Adrade, N., Iamnitchi, A., and Ripeanu, M. (2009). Individual and Social Behavior in Tagging Systems. In *the 20th ACM Conference on Hypertext and Hypermedia*, July, 2009.

Saracevic, T., & Kantor. P. (1988) A study of information seeking and retrieving. II. Users, questions and effectiveness. *Journal of the American Society for Information Science*, 39 (3), 177-196.

Saracevic, T., & Kantor. P. (1988) A study of information seeking and retrieving. III. Searchers, searches and overlap. *Journal of the American Society for Information Science*, 39 (3), 197-226

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000). *Analysis of recommendation algorithms for e-commerce.* Paper presented at the the 2nd ACM conference on Electronic commerce, Minneapolis, Minnesota, United States.

Sarwar, B., Konstan, J., Borchers, A., Herlocker, J., Miller, B., & Riedl, J. (1998) Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System. *Proceedings of the ACM Conference on computer Supported Cooperative Work (CSCW) 1998.*

Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2002). *Collaborative Filtering: Methods and metrics for cold-start recommendations.* Paper presented at the the 25th annual international ACM SIGIR conference on Research and development in information retrieval, Tampere, Finland.

Schek, H. J., Schuldt, H., and Weber, R.(2002). Hyperdatabases - Infrastructure for the Information Space. In *Proc. of VDB Conf.*, pp 1-15, Brisbane, Australia, 2002.

Schilling, M. A. (2005). A "small-world" network model of cognitive insight. *Creativity Research Journal*, 2(3), 131-154.

Schneider, J. W., & Borlund, P. (2004). Introduction to bibliometrics for construction and

maintenance of thesauri. *Journal of Documentation, 60*(5), 524-549.

Schwartz, M.F. and Wood, D. C. M. (1993). Discovering Shared Interests Using Graph Analysis. *Communications of the ACM*, 36(8), 78–89.

Scott, J. (2000). *Social network analysis : a handbook*. London ; Thousands Oaks, Calif.: SAGE Publications.

Seidman, S. B. (1983). Network structure and minimum degree, *Social Networks*, 5, 269-287.

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., et al. (2006). Tagging, communities, vocabulary, evolution. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work* (pp. 181-190). Banff, Alberta, Canada: ACM Press.

Shadish, W., Tolliver, D., Gray, M., & Gupta, S. (1995). Author judgments about works they cite: three studies from psychology journals. *Social studies of Science, 25*, 477-497.

Shardanand, U., & Maes, P. (1995). *Social Information Filtering: Algorithms for Automating "Word of Mouth"*. Paper presented at the ACM Conference on Human Factors in Computing Systems.

Shaw, W. M. (1991a). Subject and Citation Indexing. Part I: The Clustering Structure of Composite Representations in the Cystic Fibrosis Document Collection. *Journal of the American Society for Information Science, 42*(9), 669-675.

Shaw, W. M. (1991b). Subject and Citation Indexing. Part II: The Optimal, Cluster-Based Retrieval Performance of Composite Representations. *Journal of the American Society for Information Science, 42*(9), 676-684.

Shaw, W. M., Jr. (1990a). Subject indexing and citation indexing. Part I: Clustering structure in the cystic fibrosis document collection. *Information Processing and Management*, 26, 693-703.

Shaw, W. M., Jr. (1990b). Subject indexing and citation indexing. Part II: An evaluation and comparison. *Information Processing and Management*, 26, 705-718.

Shirky, C. (2005). Ontology is Overrated: Categories, Links, and Tags. From the shirkey.com blog. http://shirky.com/writings/ontology_overrated.html.

Simmel, G. (1955). *Conflict and the web of group affiliations*. Glencoe, IL: Free Press.

Sinclair, J. & Cardew-Hill, M. (2007). The folksonomy tag cloud: when is it useful? Journal of Information Science.

Small, H. G. (1973). Co-citation In Scientific Literature - New Measure Of Relationship Between 2 Documents. *Journal of the American Society For Information Science, 24*(4), 265-269.

Small, H. G. (1978). Cited Documents as Concept Symbols. *Social Studies of Science, 8*(3), 327-340.

Small, H. G. (1982). Citation context analysis. In B. Dervin & M. J. Voigt (Eds.), *Progress in communication sciences* (Vol. 3, pp. 287-310). Norwood, NJ: Ablex.

Small, H. G., & Griffith, B. C. (1974). The structure of scientific literature, i: Identifying and graphing specialties. *Science Studies, 4*, 17-40.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, Massachusetts: Harvard University Press.

Snyder, H., Cronin, B., & Davenport, E. (1995). What's the use of citation? Citation analysis as a literature topic in selected disciplines of the social sciences. *Journal of Information Science, 21*(2), 75-85.

Tennis, J. T. (2006). Social Tagging and the Next Steps for Indexing. Proceedings 17th SIG/CR Classification Research Workshop, Austin, Texas.

Terveen, L., & Hill, W. (2001). Beyond Recommender Systems: Helping People Help Each Other. In J. Carroll (Ed.), *HCI in The New Millennium*: Addison-Wesley.

Terveen, L., Hill, W., Amento, B., McDonald, D., & Creter, J. (1997). PHOAKS: A System for Sharing Recommendations. *Communications of the ACM, 40*(3), 59-62.

Tushnet, R. (1997). Legal fictions: Copyright, fan fiction, and a new common law. *Loyola of Los Angeles Entertainment Law Journal, 17*(3), 651-686.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.

Vader Wal, T. (2005). Folksonomy Definition and Wikipedia. http://www.vanderwal.net, Nov. 2, 2005.

Wasserman, S., & Faust, K. (1994). *Social network analysis : methods and applications*. Cambridge ; New York: Cambridge University Press.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p. *Psychometrika, 61*(3), 401 - 425.

Wasserman, S., Scott, J., & Carrington, P. (2005). Introduction. In P. Carrington, J. Scott & S. Wasserman (Eds.), *Models and Methods in Social Network Analysis* (pp. 1-7). New York: Cambridge University Press.

Watts, D. J. (1999). Networks, dynamics, and the small-world phenomenon. *American Journal of Sociology, 105*, 493-527.

Watts, D. J. (2004). The "new" science of networks. *Annual Review of Sociology, 30*, 243-270.

Watts, D. J. and Strogatz, S. H. (1998). *Collective dynamics of "small-world" networks*, Nature,

393, 440–442.

Watts, D. J., & Strogatz, S. H. (1998). Collective Dynamics of 'Small-World' Networks. *Nature, 393*, 440-442.

Wellman, B. (1983). Network analysis: Some basic principles. In R. Collins (Ed.), *Social Theory 1983* (pp. 155-200). San Francisco: Jossey-Bass.

Wellman, B. (1988). Structural analysis: from methods and metaphor to theory and substance. In B. Wellman & S. D. Berkowitz (Eds.), *Social Structures: A Network Approach* (pp. 19-61). Cambridge: Cambridge University Press.

Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia, M., & Haythornthwaite, C. (1996). Computer Networks as Social Networks: Virtual Community, Computer Supported Cooperative Work and Telework. *Annual Review of Sociology, 22*, 213-238.

Wexelblat, A., & Maes, P. (1999). Footprints: history-rich tools for information foraging. In *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit* (pp. 270-277). Pittsburgh, Pennsylvania, United States: ACM Press.

White, H. D. (1990). Author co-citation analysis: Overview and defense. In C. L. Borgman (Ed.), *Scholarly communication and bibliometrics* (pp. 84-106). Newbury Park, CA: Sage.

White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure on intellectual structure. *Journal of the American Society for Information Science, 32*, 163-172.

White, H. D., & McCain, K. W. (1989). Bibliometrics. *Annual Review of Information Science and Technology, 24*, 119-186.

White, H. D., & McCain, K. W. (1997). Visualization of literatures. *Annual Review of Information Science and Technology, 34*, 99-168.

Whittaker, S., & Sidner, C. (1996). Email overload: exploring personal information management of email. In *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground* (pp. 276-283). Vancouver, British Columbia, Canada: ACM Press.

Wilson, C. S. (1999). Informetrics. *Annual Review of Information Science and Technology, 34*, 107-247.

Wilson, T. P. (1982). Relational networks: An extension of sociometric concepts. *Social Networks, 4*, 105-116.

Wisniewski, E. J., & Bassok, M. (1999). What Makes a Man Similar to a Tie? Stimulus Compatibility with Comparison and Integration. *Cognitive Psychology, 39*, 208-238.

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science, 18*, 221-281.

Wittgenstein, L. (1974). *Philosophical investigations*. Oxford: Blackwell.

Wouters, P. (1998). The signs of science. *Scientometrics, 41*(1-2), 225-241.

Wrobel, S. (1994). *Concept Formation and Knowledge Revision*. Norwell, MA: Kluwer Academic Publishers.

Wu, H., Zubair, M., & Maly, K. (2006). Harvesting social knowledge from folksonomies. In *Proceedings of the seventeenth conference on Hypertext and hypermedia* (pp. 111-114). Odense, Denmark: ACM

Yamaguchi, K. (1994). The flow of information through social networks: Diagonal-free measures of inefficiency and the structural determinants of inefficiency. *Social Networks*, 16: 57-86.

Yuan, Y., Fulk, J., Shumate, M., Monge, P., Bryant, J. A., & Matsaganis, M. (2005). Individual participation in organizational information commons: The impact of team level social influence and technology-specific competence. *Human Communication Research, 31*(2), 212-240.

Zunde, P, & M.E. Dexter. (1969). Indexing consistency and quality. *American Documentation, 20*(3), 259-267.