Finding Communities in Typed Citation Networks

by

Frederick W. Kroon

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2008

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

As the Web has become more and more important to our daily lives, algorithms that can effectively utilize the link structure have become more and more important. One such task has been to find communities in social network data. Recently, however, there has been increased interest in augmenting links with additional semantic information. We examine link classification from the point of view of scientometrics, with an eye towards applying what has been learned about scientific citation to Web linking. Some community detection algorithms are reviewed, and one that has been developed for topical community finding on the Web is adapted to typed scientific citations.

Acknowledgements

First, I would like to thank my supervisor, Chrysanne DiMarco, for all her support throughout my stay at Waterloo. Special thanks also goes out to the members of my supervisory committee, Bob Mercer, Paul Thagard, Robin Cohen, and Peter Carrington. I have benefited greatly from our discussions and their suggestions.

Thanks also to all of those that worked on the IN^{3}SCAPE project, especially Radoslav Radoulov, who put together the manual citation annotation system that led to the insights on citation classification systems. Also, thanks to Pascal Poupart for introducing me to many statistical methods in Natural Language Processing.

I would also like to thank the many current and former members of the Artificial Intelligence Lab, both student and faculty. Many of my ideas have come out of informal discussions with people in the lab, and listening to others talk about their work is a constant source of inspiration. Thanks especially to all of the following for their input:

Ed Barsalou, Allan Caine, Matt Enss, Greg Hines, Reid Kerr, Kevin Regan, Tyrel Russell, Aaron Scriver, Adi Seth, Matthew Skala, Ray Singh, Martin Talbot, John Whissell, Dana Wilkinson, Rob Warren, Jie Zhang

Finally I would like to thank my girlfriend, Julie, who has put up with me droning on about “communities”, “scientometrics”, and “social networks” for far too long.

Dedication

I dedicate this thesis to my father, who has been tirelessly supportive and patient for many, many years.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the recent increase in the amount of textual material available on-line, interest in navigating the available literature has also increased. Early attempts to search through on-line documents met with limited success. All such approaches focused purely on the contents of the documents themselves, without taking into account how the documents were inter-related. As such, there was no good way to judge the quality of the information contained in any given document.

The most important breakthrough in information retrieval on the Web was the incorporation of the link structure of the Web into search-ranking algorithms. The most successful of these approaches was Google's PageRank algorithm [66], which incorporated the simple idea that the more Web pages that link to a given target page, the more important the target page. Another such system is IBM's Clever Project [14].

The Web was clearly not the first collection of documents to be interlinked. Other, earlier collections included legal documents, patents, and scientific journal articles. All three make extensive use of citations to situate the document in the context of earlier work. These citations serve to link each document to other, related documents, in much the same way as hyperlinks on the Web.

Algorithms such as PageRank were inspired by previous work on scientific citation. For many years, researchers have realized the importance of inter-document citation for the navigation of the world of scientific and legal documentation [32]. Furthermore, citation counting had been used for assessing the importance or quality of scientific research output [19]. It was this intuition that guided the development of the PageRank algorithm. As the number of Web pages that link to a given page increases, the perceived importance of that page also increases.

1.1 The Future of the World Wide Web

1.1.1 The Semantic Web

One dream for the Web that has yet to be fully realized is that of the Semantic Web. The Semantic Web is an extension of the Web as it currently stands, but with the inclusion of additional information aimed specifically at automated processing. Tim Berners-Lee, creator of the original World Wide Web, expresses this goal as follows[1] [8]:

> In the second part of the dream, collaborations extend to computers. Machines become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A "Semantic Web," which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy, and our daily lives will be handled by machines talking to machines, leaving humans to provide the inspiration and intuition. The intelligent "agents" people have touted for ages will finally materialize.

So far, proposals for the Semantic Web have focused primarily on the semantics of the information contained *within* documents. However, as for search, the semantics of the links *between* documents will be as important as the documents themselves. In fact, in the earliest stages of the construction of the current Web, Berners-Lee had a similar intuition: "One thing I wanted to put in the original design was the 'typing' of links".[2] One obvious way to capture such semantic relationships is to simply label each link with a type, drawn from an ontology of possible relationships.

Numerous systems for attaching semantics to Web information are currently being developed. As mentioned earlier, most of these focus primarily on the semantics of documents, rather than the semantics of the relationships between documents. For example, the Dublin Core Ontology, defined by the Dublin Core Metadata Initiative[3], at this point defines a 'relation' between two resources, but does not go any further, leaving open the particular relationships that might be defined.

Though the current vocabularies for link semantics are extremely sparse, they are clearly important to the development of the future Web. As time goes on, algorithms that take advantage of semantic links will become increasingly important.

[1] I leave the first part of his dream for the next section.

[2] Tim Berners-Lee, quoted by Andy Carvin, "Tim Berners-Lee: Weaving a Semantic Web", digital divide network, http://www.digitaldivide.net/articles/view.php?ArticleID=20 (accessed Feb. 4, 2007).

[3] Described on their website as "...an organization dedicated to promoting the widespread adoption of interoperable metadata standards and developing specialized metadata vocabularies for describing resources that enable more intelligent information discovery systems." http://dublincore.org/about/ (accessed Jan. 15, 2007).

1.1.2 The Social Web

In addition to research on the Semantic Web, there has recently been increased focus on on-line communities. However, unlike the semantic web, which remains largely in the future, the social Web is firmly rooted in the present. With the advent of the so-called 'Web 2.0' has come the enormous popularity of on-line social networks, collaborative bookmarking, recommender systems, etc. This collaborative aspect of the Web makes up the other half of Tim Berners-Lee's dream for the Web.

> In the first part [of the dream], the Web becomes a much more powerful means for collaboration between people. I have always imagined the information space as something to which everyone has immediate and intuitive access, and not just to browse, but to create. . . . Furthermore, the dream of people-to-people communication through shared knowledge must be possible for groups of all sizes, interacting electronically with as much ease as they do now in person. [8]

As with the semantic web, there have been proposals made for typed links in the realm of the social web. For example, SIOC[4] (Semantically-Interlinked On-line Communities) and FOAF[5] (Friend of a Friend) are both ontologies interested in capturing information about relationships between individuals within on-line communities. SIOC primarily focuses on documents and their relations. For example, a message posting in an on-line forum might be linked by the 'reply_of' relation to an earlier post to which it was replying. FOAF focuses more on the various aspects of the people that make up on-line communities.

The move from networks of documents to networks of people has generated interest in not only navigating the world of linked documents, but also linked *people*. This has brought the much older field of social network analysis right to the forefront of modern information technology. The availability of large-scale social networks[6] has provided a fertile ground for the use of large-scale social networks analysis techniques.

However, though there are many social networks available on-line currently, explicit representations of such networks are not usually available for a particular on-line community of interest. Even if there were such a network, it might be incomplete. For example, links might not exist between nodes, even though there is a relationship between the people represented by those nodes. When this is the case, networks must be inferred from available information, whether it is e-mail patterns [21], discussion group postings [2], blog linking-patterns [38], scientific paper collaborations [59], or potentially many other information sources.

[4] http://sioc-project.org/

[5] http://www.foaf-project.org/

[6] For example, the networks contained within the larger social networking websites, such as LiveJournal (http://www.livejournal.com/), Facebook (http://www.facebook.com/), LinkedIn (http://www.linkedin.com/), MySpace (http://www.myspace.com/), and others.

One specific problem that has been receiving attention recently is the detection and characterization of communities (both on-line and off-line), given a graph of the social relationships between individuals. Such graphs include friendship networks, graphs of e-mail correspondence [77], and scientific collaboration networks [60].

As mentioned earlier, since such graphs are not always readily available, it may be of interest to mine community structure from information that does not strictly represent social relations. For example, Web links and scientific citations do not strictly represent social relations between the authors of the documents in question. In both cases, the author of the linking document is aware of the document linked, but not necessarily vice versa. In addition, the network formed by such links is over the set of documents, rather than authors.

However, there is something sufficiently social about hypertext and citation networks that makes them very nearly 'true' social networks. Citation links (or hyperlinks) also indicate a relation between authors, though not the direct one that would be present in, say, a network of scientific collaborations. Certainly the citing author is aware of the cited authors, and knows something about them.

So, if the citation network is not itself a straightforward example of a social network, it can certainly be interpreted as such, or a more straightforward social network might be inferred. Community identification can still be done on such structures, as, for example, on the linking patterns between blogs [15], the Web more generally [36], and networks of citations.

1.1.3 Web 3.0

Recently, the term 'Web 3.0' has been used with increasing frequency. Current usage of the term 'Web 3.0' is vague and inconsistent. Sometimes it is used to refer to continuations of trends that are already present in 'Web 2.0'. In other instances, it is used as simply as a synonym for the semantic web. Perhaps it is best not to think of Web 3.0 as either of these, but a synthesis of both: the social character of Web 2.0 combined with some or all of the machine processable features of the semantic web. This view of the future web is very close to Berners-Lee's dream of a 'read-write' web.

If this vision is anything close to reality, the future Web will require algorithms that not only take advantage of the link structure and the document contents, but the content of the links themselves. Data mining on the Web will not only require advances in current information retrieval techniques, but the development of new techniques to exploit the semantic content of links. Furthermore, the social aspect of the future Web will require algorithms to generate, navigate, and reason about social structures.

1.2 From the Web to Scientific Literature

1.2.1 Scientific Citation

The fact that the Semantic Web still lies in the future, as well as the lack of standards with respect to the typing of links on the current Web, creates problems for any attempt to develop algorithms that will be able to take advantage of semantic links in the future. However, since the study of scientific citation was so fruitfully used as a model of Web hyperlinking for the development of current document search and retrieval methods, it is reasonable to suggest that scientific citation may be a good model for document linking in general.

While the notion of typed links between documents is in its infancy on the Web, this is not the case with the scientific literature. There have been a number of proposals for classification schemes for scientific citations (see [34] for a review). As such, the study of classified citations may provide insight into links on the future Semantic Web.

1.2.2 Scientific Communities

If networks of typed citations are to serve as a model for the future web, it needs to be shown that the scientific literature also shows community structure. Like the web, the scientific literature is written by a body of people, most of whom will never meet face-to-face. Journal articles, like email, blog entries, or forum postings, are a sort of correspondence, albeit more official, between members of a community.

Researchers differ in their assumptions about what constitutes a scientific community. Reasonable possibilities include groups of collaborators, topical areas, and professional organizations. One notion that has received considerable attention is that of "the invisible college" [20]. Invisible colleges are groupings of scientists, joined by *informal* communications, in addition to traditional publishing. With the arrival of increasingly effortless methods for near-instant communications between scholars, such invisible colleges are arguably becoming increasingly important.

To be useful in the development of tools to navigate on-line scientific communities, however, these intuitive notions of community need to be formalized. For example, Girvan and Newman [37] have formalized the notion of collaborative groups by examining graphs of co-authorship of published papers. In this graph, authors are represented by vertices, and two vertices are joined by an edge if and only if the two corresponding authors have co-authored a paper. Collaborative communities, then, are simply clusters on the graph, such that there are more edges within groups than between groups.

Likewise, Gibson, Kleinberg, and Raghavan [36] have attempted to find communities linked by topic using Kleinberg's HITS model [48]. Without going into detail here, the HITS model is based on the bibliometric notions of *bibliographic* *coupling* [45] and *co-citation* [72]. These measures capture the intuition that if two papers either cite the same target, or are cited by the same source, they are likely to be topically related. Thus papers linked by bibliographic coupling and co-citation can be said to be in the same topical community.

1.2.3 Citations and Community Structure

When writing a scientific article, an author typically does not target the article at a particular individual. Rather, the article is aimed more broadly at an audience or community of scholars who might be interested in the research. When a citation is made, it indicates that the citer is among the actual audience of the cited work (though not, perhaps, the *intended* audience). A first approximation for a definition of a scientific community is a number of authors, each of whom is in the intended audiences of the others.

So, citations are clearly a form of communication to the scientific community of which the author is a part. In addition to identifying the citer as an interested party, citations fulfil other functions. For example, the citation may be used to direct a reader to further reading, point out a contrasting point of view, or assign credit to earlier work. If citations were labelled with information indicating the citer's intent, and information about the cited content, we would have a number of different networks, each with its own particular relation. These networks would give a clearer understanding of the flow of information among researchers, as well as the attitudes of some researchers to the research of others.

The notion of community I have described so far might be understood as an amalgam of a number of different communities. For example, a community of researchers might share a common problem upon which they are working, a common methodology, a common theoretical model, or possibly even a common geographical location or mother tongue. Moreover, communities have complex structures. For example, communities may overlap with one another, there may be ideological rifts within a community, and so forth.

1.3 Organization of the Thesis

Chapter 2 introduces citation networks and reviews the literature on citation classification. Some important characteristics of citation classification schemes are introduced, and a novel citation classification scheme is presented.

Chapter 3 reviews the literature on community identification. Community identification over both collaboration networks and networks of linked documents are examined.

Chapter 4 presents three possible models for converting typed citation data into a form more suitable for community identification algorithms that were developed for the web. Chapter 5 presents the results of applying one such algorithm, Kolda and Bader's TOPHITS algorithm [49] to the resulting social network.

Chapter 6 makes a case for the usefulness of typed citations to community detection as well as other information retrieval problems. In addition, future avenues of research are presented.

Chapter 2

Citation Analysis

2.1 Representations for Citation Analysis

Before examining prior work in the field of community identification, it is important to understand the two graphs that can be easily derived from citation indexes: the *citation graph* and the *collaboration graph*.

2.1.1 The Citation Graph

An *untyped* citation graph is simply a graph $G = (V, E)$ defined over a set of vertices V representing the papers themselves. An edge $e = (u, v)$ indicates that paper u cites paper v. Thus the citation graph is *directed*. While not common, it is possible to have reciprocal citations. That is, it is possible for two papers to cite each other (for example, if they are in the same volume). As such, the graph is not acyclic, but there are generally only a small number of cycles relative to the total number of edges. Additionally, as one can only cite a paper that has been published (whether formally or informally), each vertex has an associated time t at which it was published.

It is important to note that the citation graph is not, strictly speaking, a social network in and of itself. The citation graph gives relations among *documents*, not among people (at least not directly). Assuming that there is only one author per paper, and all authors can be unambiguously identified, the citation graph(s) could be mapped to a graph over authors. However, even then there is no particular relationship implied among the authors. Even the relationship "has read a paper by" might not be present, as it is quite possible that some cited papers are never actually read by the citing author[1].

[1] This seems especially likely when a citation is used simply to acknowledge a pioneer, or some oft-cited source, simply to register its existence. In the scheme of Moravcsik and Murugesan, this is known as a *perfunctory* citation. See Section 2.2.1 for more information.

Also, assuming that a paper has a single author is clearly wrong. Many papers have multiple authors, perhaps even hundreds of authors in the case of some experimental physics papers. As the research reported in a paper, and indeed the writing of the paper, might be split into smaller pieces among authors, there is no direct way to know which author made which citation. The second assumption, that authors can be uniquely identified, is also not the case. This will be seen in Section 4.2.2.

2.1.2 The Collaboration Graph

The other immediately available graph that can be inferred from the bibliographic information is the collaboration graph. Unlike the citation graph, the collaboration graph is defined over people, rather than documents. The graph contains an edge $e = \{u, v\}$ iff two authors have collaborated on a paper. As the relation of "collaborates with" is symmetric, this graph is clearly undirected. The graph might be weighted, however, as pairs of authors may collaborate on multiple documents.

Unlike the citation graph, the collaboration graph is clearly a social network in the strictest sense. Researchers who collaborate are clearly communicating with one another. If this is the case, though, why not simply find communities using the collaboration graph? Unfortunately, this graph is incomplete, as many pairs of authors communicate (perhaps even extensively), perhaps even collaborate, and never actually co-author a paper. A perfect example of this can be observed simply by reading the Acknowledgements section of a paper. Here, other researchers are given credit for contributing in some way, yet they are not co-authors, nor are they necessarily cited. Alternatively, an author might benefit considerably from communication with a colleague, but rather than co-authoring the paper, the author simply cites relevant information from one of that colleague's published papers.

The collaboration graph, however, clearly contains important information for determining community structure. At the very least it could be used as an evaluation mechanism or 'sanity check' on the social network derived from the citation graph. Clearly if the inferred social network fails to include many known collaborations, it is of inferior quality.

2.2 Citation Classification

Typed citation graphs are standard citation graphs in which each of the edges has been labelled according to some classification system. There have been numerous such schemes proposed for the classification of citations.

2.2.1 A Brief History of Citation Classification

Over the past 40 years or so, there have been a number of researchers interested in having a more complete understanding of citation, particularly in scientific writing. Probably the earliest attempt to classify citations was Eugene Garfield [31], one of the great pioneers of the field of scientometrics. Garfield defines fifteen distinct categories into which a given citation might fall.

These categories are primarily concerned with the motivation of the citing author in providing the citation. For example, a citation might be made to give credit to another researcher, to provide background material for the current work, or perhaps to dispute a prior claim. While Garfield gives no deep theoretical motivation, nor experimental evidence, to justify his classification scheme, it was well enough received to form the basis of many other schemes.

Similar schemes have been produced by a number of other authors [18, 24, 27, 31, 30, 33, 40, 52, 53, 54, 65, 68, 73]. While it is beyond the scope of the thesis to go into detail for each of these schemes, it is worthwhile exploring what they all have in common. For a closer examination of the various schemes, see Garzone [34].

Like Garfield's early scheme, each of these schemes define a set of discrete categories, and attempt to place each citation into one of these categories. While the specifics of the categories vary due to the various research interests of the scheme's designer, there are a number of categories that appear in all or most of the schemes. For example, one such classification is that of citations that *provide background.* This might be further broken down into *specific* versus *general* background (e.g., Hodges [40], Finney [27], Garzone and Mercer [33, 34], Radoulov [71]), but some claim that this distinction is artificial [68]. Another category that always appears is *methodological* citations, in which a method developed in the cited paper is used in the citing paper. On the other hand, some categories are less often included. For example, *used to suggest future research* is relatively rarely encountered in the various schemes that have been proposed.

Generally, there is a good deal of agreement between the various schemes. Where there is not, usually a category has simply been omitted, as it is unimportant to the scheme designer's research goals. That being said, nearly all schemes define a long list of categories, with little attention payed to the particular kinds of information that each category represents.

One classification scheme that deviates from the typical enumeration of categories is that of Moravcsik and Murugesan [57]. In their analysis, they identified four binary dimensions upon which a particular citation may vary. The four dimensions are as follows:

Conceptual versus operational

Is the information cited a theory (conceptual) or is it a method or tool (operational) that is used in the citing work?

Organic versus perfunctory

Is the information cited critical to the understanding of the citing work (organic), or is the reference merely mentioning that the cited work exists (perfunctory)?

Evolutionary versus juxtapositional

Is the citing work building on the cited work (evolutionary), or does it present a competing view (juxtapositional)?

Confirmative versus negational

Is the cited work being confirmed (confirmative) or disputed (negational)?

Of these dimensions, the first concerns the content of the cited article, while the remainder address the relationship between the cited and citing work. Of the three latter categories, it is particularly interesting that two of the dimensions address the question of whether the two papers are in agreement or disagreement, and one addresses the question of how critical the cited work is to the citing work. This is likely due to the fact that the scheme was developed primarily to assess the viability of bibliometric methods for measuring intellectual achievement in the sciences [56]. When viewed from this perspective, the citer's opinion of a cited work and the extent to which new work is based on the cited work are critically important. Clearly if a citation is negative it should not contribute to the evaluation of one's academic prowess.

Schemes for Automated Citation Classification

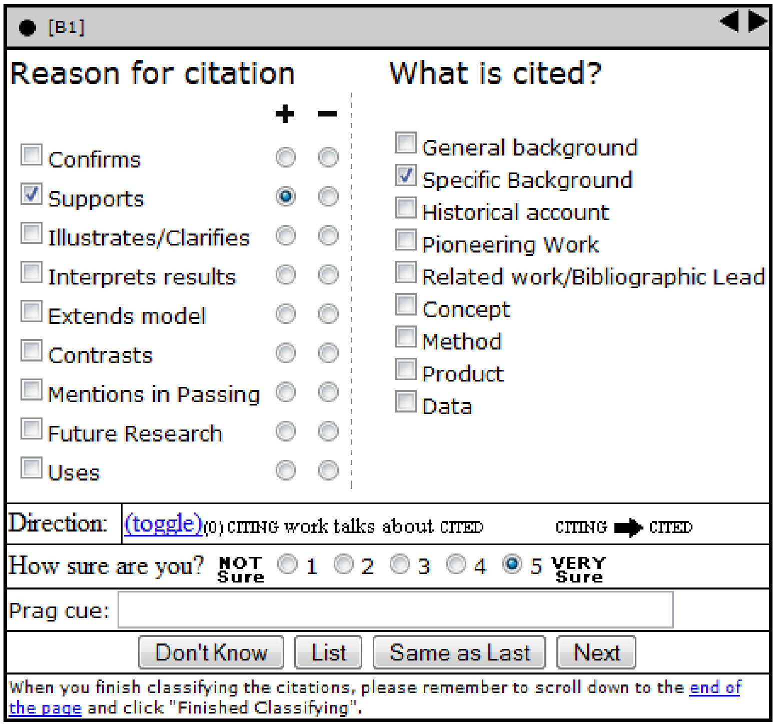
As already mentioned, each scheme is tailored to the particular application to which it is to be applied. One application that has been of increasing interest as of late is automated citation classification [33, 58, 75]. Since the great majority of citations have not been manually classified or annotated by their authors (or anyone else), it looks as though this is an ideal problem for research in natural language processing and information retrieval.

One scheme in particular that was designed specifically for the purposes of automated citation classification is that of Garzone and Mercer [33, 34]. Like all of the schemes mentioned so far, their scheme consists primarily of a list of possible classes (in this case, 35). Unlike many of the previous schemes, however, they attempt to organize the various classes into a hierarchy. This allows the scheme to contain both very broad categorizations, as well as very fine-grained distinctions.

There are benefits to such a system from the perspective of automated citation classification, since a system that is unsure about the fine-grained classification can fall back to the broader category. However, in addition, this same feature affords manual classifiers the same opportunity. For these reasons, this scheme was used as a starting point for the scheme of Radoulov [71].

Like Garzone, Radoulov was interested in producing an automated citation classifier. However, as very few citations have been classified, there was lack of data from which to train a model. This led to the need to produce a reasonably sized manually classified set of citations, in a relatively short period of time. Unfortunately, Garzone's scheme, with 10 broad categories and 35 sub-categories, was far too unwieldy for a manual classifier, especially since a citation could fall into many of the categories simultaneously.

Borrowing from the ideas of Moravcsik and Murugesan, Radoulov factored out some of the features of each of the categories to provide a more manageable user interface.



● [B1]

Reason for citation

	+	−
□ Confirms	○	○
☑ Supports	●	○
□ Illustrates/Clarifies	○	○
□ Interprets results	○	○
□ Extends model	○	○
□ Contrasts	○	○
□ Mentions in Passing	○	○
□ Future Research	○	○
□ Uses	○	○

What is cited?

□ General background
☑ Specific Background
□ Historical account
□ Pioneering Work
□ Related work/Bibliographic Lead
□ Concept
□ Method
□ Product
□ Data

Direction: (toggle)(0) CITING work talks about CITED   CITING ➡ CITED

How sure are you? NOT Sure ○ 1 ○ 2 ○ 3 ○ 4 ● 5 VERY Sure

Prag cue:

When you finish classifying the citations, please remember to scroll down to the end of the page and click "Finished Classifying".

Figure 2.1: Radoulov's [71] manual citation annotation tool. Note that the tool allows multiple selections, generating an immense number of possible classifications. In addition to the typical categories, the tool includes the ability to make a positive versus negative distinction for possible pragmatic function and incorporates a surety scheme for use in training an automatic classifier.

By factoring out various components of the various classifications, Radoulov's system reduces the number of decisions an annotator must make, and allows the creation of some very fine-grained classifications. Most importantly, the system breaks out information about the citer's purpose for making the citation and information about which particular part of the cited paper is of interest. We will come back to this distinction in the next section.

Though it is an advance over prior classification systems, the scheme, as presented in Figure 2.1 conflates information, and may lead to confusion on the part of the annotator. For example, there is no way to state that the citer agrees with the citee's theory, but disagrees with their experimental methodology. Example 2.2.1 shows a similar case, in which the citers both agree and disagree with the cited work. These problems will be addressed by a novel classification scheme, presented in the next section.

Example 2.2.1
"Although we agree that without a probability sample, the results should only apply to individuals in studies and should not be generalized to a population, we disagree with the contention made by Mage et al. that no useful information can come from studies using samples that do not fulfill their criteria."[7]

2.3 A Novel Classification Scheme

While all of the citation classification schemes presented here capture important information about citations, most confuse different kinds of information within the classification scheme. A citation indicates the existence of a relationship between two documents. What is not made explicit in the schemes developed so far is that a citation can capture just about *any* relationship between a pair of documents.

Untyped citations indicate the existence of a relationship between two documents, but fail to provide more information about that relationship. Normally this relationship is between the citing and cited documents, but it is also possible that the citing document is expressing a relationship between two distinct cited documents, as is illustrated in Example 2.3.1.

Example 2.3.1
"The acceptance of the concept of hormesis, a specific type of nonmonotonic dose response, has accelerated in recent years (Academie Nationale de Medecine 2005; Cendergreen et al. 2005; Kaiser 2003; Puatanachokchai et al. 2005; Randic and Estrada 2005; Renner 2003). Nonetheless, it has not been without its detractors. One article critical of the concept was published last year in *Environmental Health Perspectives* (Thayer et al. 2005)."[15]

That being said, there are certainly kinds of information that are much more likely to be expressed by a citation. By examining the various schemes and in what ways they agree and disagree, four distinct kinds of information can be identified that are relevant to the classification of a citation:

1. Author attitudes toward cited work
    - What is the citer's opinion of the cited work? (negative, positive, or neutral)
2. Function of citation

- What function does the citation serve with respect to the reader?
    - e.g., citation points reader to related material

3. Content of cited work

    - What information within the cited work is of interest to the citing author?
        - e.g., cited work contains a method used in citing work
        - e.g., cited work is a historical account

4. Relationship of citing to cited work

    - What is the relationship between the citing and cited work?
        - e.g., citing work is supported by cited work
        - e.g., citing work disputes claims made in cited work
        - e.g., citing work makes use of methods or data from cited work

In addition, a citation might express multiple relationships between two documents. As such it is not sufficient to simply specify each of the pieces of information independently, as was done by Radoulov. Each piece must be coupled with other pieces to show the complete relationship. Thus, no scheme that simply uses a set of dimensions, like that of Moravcsik and Murugesan, will suffice to show all of the possible relationships. Example 2.3.2 shows such a case, which would be classified as both confirmative and negational in the Moravcsik and Murugesan scheme.

Example 2.3.2
"We agree with all of the possible explanations provided by Fries and Krishnan, although we disagree with the potential magnitude of the biases discussed."[26]

Thus, each citation C can be expressed as a set of *relationships* between two documents.

$$C = \{r_1, r_2, \ldots, r_n\} \tag{2.1}$$

Each relationship is in turn a tuple of *features.* The most immediately obvious feature of citation is the *valence* of the citation. Most or all previous work on citation classification has included valence as a critical component. Valence indicates the citing author's sentiment toward the cited work. For example, if the author is disputing the work, the valence is negative. If, on the other hand, the citing work relies upon the cited work, the valence is positive. Rather than simply a binary value, the valence corresponds to a continuous dimension that ranges from -1 to 1, where -1 denotes the strongest negative opinion, and 1 denotes the strongest positive opinion. Thus, if a citation has a value of -1, the citer has a strongly negative opinion of the cited work.

The remaining features are broken down into three broad categories. The first category contains features describing the *relationship* between the citing and cited work ("how" they are related). The second contains those features describing the *use* of the cited material ("why" they are related). Finally, the third contains features describing the *content* of the cited work ("what" content is related).

Relationship between citing and cited document :

- Cited work supports/does not support/contradicts citing work.
- Cited work illustrates or clarifies citing work
- Citing work corrects cited work
- Cited work contrasts with the citing work
- Citing work uses a method, tool, or data from cited work

Use of cited work in current document :

- Cited work is used to assist with the interpretation of results
- Cited work is used to develop or extend a model
- Cited work is used in the formulation of future research
- Cited work is mentioned in passing

Content of the cited work :

- General or specific background information
- Historical account
- Pioneering work
- Bibliographic lead
- Concept (e.g., theory, equation, model)
- Method or procedure
- Physical product
- Data

Thus, each feature is a 4-tuple $< valence, relation, use, content >$.

When each edge in the citation graph is classified in this way, the resulting graph looks something like the example in Figure 2.2.

It is important to note here that this scheme is not intended to be exhaustive. There are almost certainly relationships between documents that have not been captured here, as well as possible uses or contents. Alternatively, perhaps the scheme is not specific enough for some needs. It is simple enough to expand the scheme by breaking down the features listed above into smaller classes. For example, the type 'concept' has been broken down into 'theory', 'equation', and 'model'.

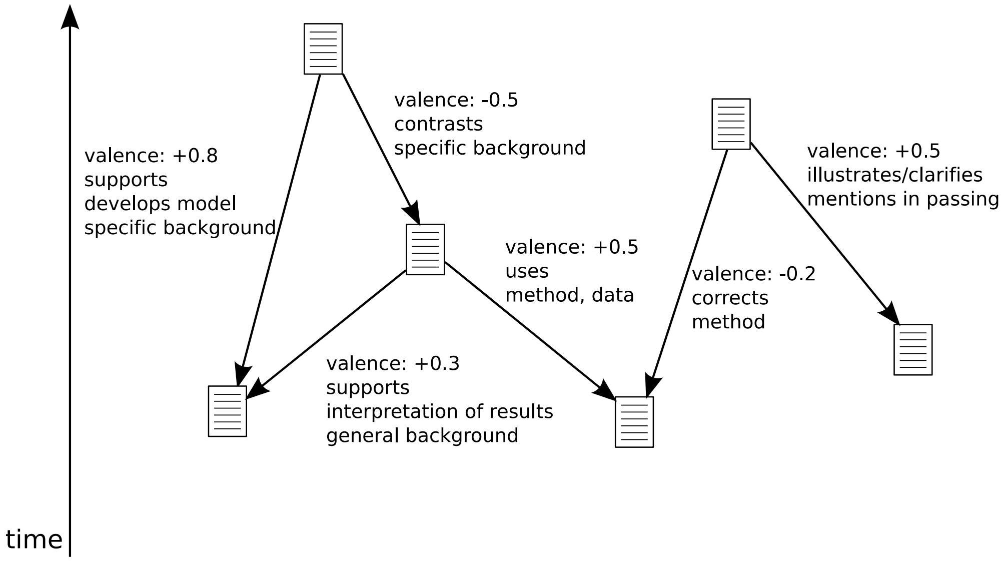


Figure 2.2: An example of a typed citation graph. Each link is labelled with its respective features. All citations are assumed to express only one relationship between papers. Note that the graph is directed, weighted, and signed. The vertices are sorted vertically according to their time of publication.

2.4 Web Link Classification

Though the idea of adding types or classifications to links has been around since the early days of the web, there has been little work on providing any formal classification scheme for Web links. However, there have been a few studies of the motivations for linking.

Chu [17], for example, examined links to the Web sites of 54 schools of Library and Information Science. From this, he identified 24 reasons for hyperlinking. The identified reasons are quite diverse, and very particular to academic websites. They do not lend themselves to the kind of analysis that has been given here for scientific citation.

Kim [46] performed interviews with faculty and graduate students at Indiana University, regarding their motivations for linking. He found 19 motivations, grouped into three broad classes: scholarly, social and technological. Of these, 16 were approximately equivalent to types given for scientific citations by previous authors.

Thelwall [76] examined inter-university links between 111 universities in the United Kingdom. He finds four broad categories: general navigational, ownership, social, and gratuitous. General navigational links are provided mostly as a navigational aid to the visitor. Thelwall states that they can be seen as roughly analogous to scientific citations that provide general background material. Ownership links are those links that provide information on the authorship, co-authorship, or ownership of the linking page. For example, a link attached to a university crest and leading to the main page of the university would be an ownership link. Finally, gratuitous links are those that seem to have been created with no obvious communicative motivation.

While these kinds of studies are necessary to produce an ontology of kinds of linking on the web, they have all come from the library and information science community, and thus have focused on academic linking. All of them have looked exclusively at links that either point to a university site, from a university site, or both. From the point of view of understanding Web linking as part of scientific communication these studies are worthwhile. However, the majority of linking on the Web is not academic, so to understand the motivations behind Web linking in general, more studies are necessary.

Chapter 3

Community Detection

While there have been many attempts in the literature to find communities in social networks, there is no firm agreement among authors as to what exactly constitutes a community. Each algorithm presented for community detection makes an assumption about what feature of the graph they are examining should be termed a “community”. There are two main camps, which I will call “modular” and “structural”.

The modular view of community structure within networks is often attributed to Newman [62]. Under this view, a graph showing community structure contains groups of nodes within which the edges are relatively dense, and between which the edges are relatively sparse. Each of these groups can be identified as a community. This view of community structure is most suitable for traditional sociograms, in which each vertex represents an individual, and each edge represents a social relationship between two individuals. Typically these graphs are undirected. The collaboration graph described in Section 2.1.2 is just such a graph.

The structural view is most often associated with finding communities in sets of interlinked documents, rather than amongst people. Unlike sociograms, these graphs are typically directed. The citation graph from Section 2.1.1 is such a graph, as is a graph of linked Web pages. In this view, there is some characteristic structure to a subgraph that represents a community. So, identifying communities is a matter of finding particular substructures within the graph.

3.1 Community Detection in Untyped Networks

3.1.1 Clustering

There are numerous clustering algorithms that have been applied to graphs: link-based clustering [78], spectral clustering [64], probabilistic clustering [74], latent space clustering [39, 41], among others. The methods reviewed below have specifically been used to ascertain community structure.

Bibliographic Coupling and Co-Citation

The simplest approach to identifying communities is to convert the citation graph into a structure better-suited for traditional clustering techniques. This requires the computation of a similarity measure from the available link structure. Once the similarity measure is computed for all pairs of documents, any one of many clustering algorithms can be applied (see Jain, Murty, and Flynn [43] for a review).

There are numerous candidates for an appropriate similarity metric, each with its own benefits and drawbacks [47]. Two that have been popular for many years in the scientometrics literature are *bibliographic coupling* [45] and *co-citation* [72].

Two documents are bibliographically coupled if they reference one or more of the same documents. The more documents that they both cite, the closer the documents are said to be. Co-citation is based on the number of documents that cite both of the two target documents. The more documents that cite both, the more similar the documents are said to be. Figure 3.1 illustrates the two measures.

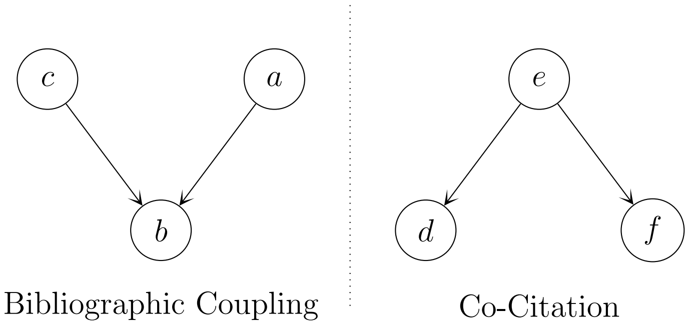


Figure 3.1: An example of bibliographic coupling and co-citation. In the Figure, documents represented by vertices a and c are bibliographically coupled, as they both cite b. Documents represented by vertices d and f are related by co-citation as they are both cited by e.

One major distinction between the two measures is that bibliographic coupling is static, whereas co-citation is dynamic. As new papers are being written all the time, two papers that were not previously related might become so, if both are cited in the same paper. However, as papers typically do not change their bibliographies after publication[1], two papers bibliographically coupled will stay that way over time.

One problem with using a similarity based approach is that by reducing all of the information represented by labelled edges in the classified citation graph to a single similarity value, all information on what binds the communities together is lost.

[1] Although with easy access to online publication through e-print archives such as arXiv (http://arxiv.org/) and BioMed Central (http://www.biomedcentral.com/) this might not always be the case.

In order to keep this information, more than one similarity metric would need to be computed (one for each classification, for example). Also, both the valence and weight of each citation would need to be considered, to identify possibly mutually antagonistic sub-communities.

Flow-Based Clustering

Flake et al. [28] give an algorithm for finding communities on the Web using techniques based on *flow*. The *max-flow min-cut theorem* [25] states that the maximum flow between a source vertex s and a sink vertex t is equal to the minimum size of a cut separating s and t. As there are polynomial time algorithms for max-flow, computing a graph partition given two 'seed' vertices can be done efficiently. This, however, assumes that there are reasonable choices for the seed vertices, which may not be the case.

An, Janssen, and Milios [3] applied the max-flow min-cut algorithm to a citation graph built from the ResearchIndex[2] digital library. They manually separated the corpus into topics, and identified authoritative articles for each topic. They treated all directed links as undirected. The algorithm was then applied using the hand-picked authorities as seed vertices. They found that the 'communities' discovered by this approach were less than satisfactory, and concluded that more sophisticated algorithms would be required.

Betweenness Clustering

Girvan and Newman [37, 63] provide a community identification algorithm based on the notion of 'betweenness' of edges. Rather than start with an unlinked graph, then add edges between individuals who are close together on the graph, as in many hierarchical agglomerative methods, they start with a full graph, then selectively delete edges that fall between (rather than within) communities. Locating these edges depends on an index called *edge betweenness*.

The betweenness centrality of a vertex is defined as the number of shortest paths upon which the vertex lies [29]. Edge betweenness is simply an adaptation of this measure to edges. More formally, let σ_{ij} be the number of shortest paths from i to j, and $\sigma_{ij}(e)$ be the number of shortest paths from i to j that include edge e. Then the edge betweenness e is defined as:

$$B_e = \sum_{i,j \in V} \frac{\sigma_{ij}(e)}{\sigma_{ij}} \tag{3.1}$$

The intuition is that edges that lie between clusters will be on many shortest paths, as paths that originate in one cluster and terminate in another will route

[2] Now CiteSeer (http://citeseer.ist.psu.edu/).

through them. Thus, the higher the betweenness of a link, the more likely it belongs between larger social groupings. Figure 3.2 illustrates the intuition behind their algorithm.

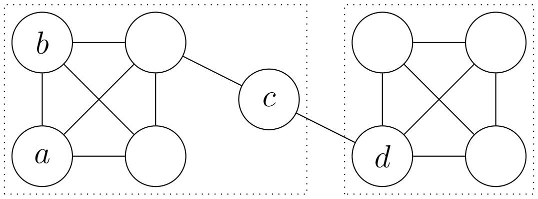


Figure 3.2: An example of edge betweenness. In the above graph, the edge (a, b) has a betweenness value of 1, as it is on a single shortest path (from a to b). The edge (b, c), however has an edge betweenness of 20, as all paths that pass between cliques must include it. If (c, d) is removed, the graph breaks into two components (shown in the dotted boxes).

3.1.2 HITS

The HITS (Hyperlink-Induced Topic Search) algorithm [48] began as a technique to identify authoritative sources on the Web, by making use of the link topology. In HITS, websites are divided into *hubs*, which act as pointers to information, and *authorities*, which contain that information. Thus, hubs will link predominantly to authorities, but less to each other, and authorities may not link at all.

The intuition here is that authoritative sources will be linked to by many pages. Furthermore, pages that link to many authorities have a different sort of authority—the knowledge of which sources should be considered authoritative for a given subject. As such, each page is given both a hub score and an authority score. This process can be performed by computing the principle eigenvectors of two graphs, M_{hub} and M_{auth}.

HITS was initially proposed to provide a ranking of documents for some given query. The algorithm works in two stages: In the first stage, documents are prefiltered to be relevant to a particular topic. In the second stage, the hub score $x^{<p>}$ and authority score $y^{<p>}$ are computed for every document in the filtered set.

The two scores are iteratively calculated as follows:

$$a(p) \leftarrow \sum_{q \to p} h(q) \tag{3.2}$$

$$h(p) \leftarrow \sum_{p \to q} a(q) \tag{3.3}$$

Let $M_{HUB} = AA^T$ and $M_{AUTH} = A^T A$, where A is the adjacency matrix of the graph. The hub and authority weights computed by the HITS algorithm converge to the principle eigenvectors of AA^T and $A^T A$ respectively.

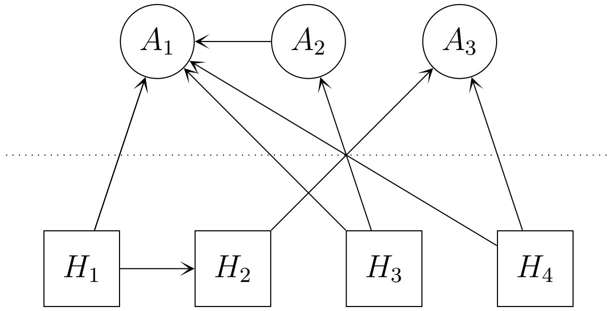


Figure 3.3: An example of a hubs and authorities structure. Though this graph is not actually bipartite, it is nearly so. The dotted line represents the partition that would partition the graph if it were actually bipartite (i.e., if the link from H_1 to H_2 and from A_2 to A_1 were deleted).

While HITS was designed to rank the importance of documents relevant to a particular query, it has been used to perform community detection on the Web as well [36]. Broadly speaking, a community is identified by the existence of a hubs and authorities structure on the graph of all pages. Since one does not know the topics that bind the communities to be discovered in advance, it is not possible to pre-filter the documents by some topical query.

In the original HITS algorithm, the principle eigenvectors of M_{HUB} and M_{AUTH} correspond to the hub and authority scores of the pages in the largest hub and authority structure in the original graph. To find communities, some of the non-principle eigenvectors are computed as well. If the eigenvectors are placed in descending order by their associated eigenvalues, each pair of eigenvectors represents a community with decreasing density of connections between the hubs and authorities.

ARC [13], a part of the IBM Clever Project, is an extension of the HITS algorithm to take information about topic into account. As such, it follows the same broad outline as HITS, beginning with a traditional search query to limit the set of documents by topic, and then computing the hubs and authority values for the documents in the reduced set.

ARC differs from HITS, however, in how the hubs and authorities scores are calculated. Rather than treat all links as identical, as is done in HITS, ARC increases the importance of links that are surrounded by text relevant to a given topic of interest. This is accomplished by re-weighting some entries in the adjacency matrix of the graph.

The original adjacency matrix is given by:

$$W_{ij} = \begin{cases} 1 \text{ if } i \rightarrow j \in G \\ 0 \text{ otherwise} \end{cases} \tag{3.4}$$

The weighted adjacency matrix of ARC is given by:

$$W_{ij} = \begin{cases} 1 + n(t) \text{ if } i \rightarrow j \in G \\ 0 \text{ otherwise} \end{cases} \tag{3.5}$$

where $n(t)$ is the number of terms shared between the topic description and the text surrounding the link anchor. The size of the window was established by a series of simple experiments to be 50 bytes on either side of the anchor.

Text surrounding a link could be seen as information about the connection between the linking and linked documents. Indeed, immediately surrounding text has been targeted as important for automated link classification [55, 58].

3.1.3 PageRank

As the original algorithm behind the wildly successful Google search engine, the PageRank algorithm [10, 66] is quite probably the most-used ranking algorithm for Web data of all time. Therefore, no examination of algorithms related to Web information retrieval would be complete without it. Unlike HITS, however, PageRank has never been particularly successfully applied to the identification of communities on the Web (or elsewhere).

That being said, it is an excellent method for providing information about the authority of individuals (whether they be documents or people), in a linked environment. Thus, it might be useful as part of a community detection algorithm, or as part of a system that gives information about individuals in an already-identified community.

The PageRank algorithm is based on a ‘random surfer’ model. Intuitively, the random surfer begins at some page in the collection of documents to be searched, and begins clicking on links, completely at random. In addition, the surfer will jump to a page at random with probability E. The PageRank $R'(u)$ of a given document u is the probability that the random surfer will arrive at that document:

$$R'(u) = c \sum_{v \in B_u} \frac{R'(v)}{N_v} + cE(u) \tag{3.6}$$

where B_u is the set of documents that point to u, N_v is the number of links leaving node v, and E is a vector over documents that provides the probability that the surfer will jump to a document at random.

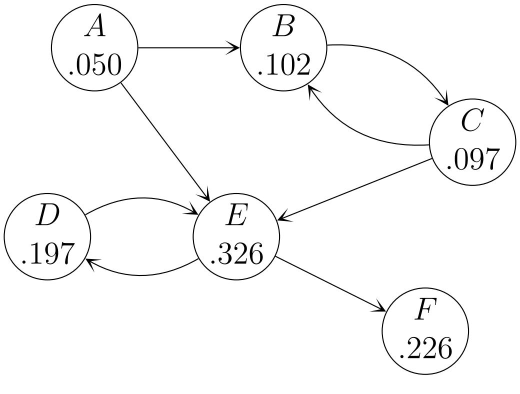


Figure 3.4: PageRank values for a small artificial network

3.2 Community Detection in Typed Networks

In all the research on community detection presented so far, there has been no attempt at finding communities on networks of typed links. While there has been little work on such algorithms, there has been some progress on each of the important aspects of typed citation networks.

3.2.1 Multiple Relationships

Cai et al. [11, 12] have gone beyond the case of single relations in community identification. In their work, they attempt to mine information about community from graphs that have multiple relations. They do this by attempting to determine the relative importance of each relation, given a group of individuals. For example, they suggest one relationship between authors might be “has presented at conference x” where x is one of a number of possible conferences.

The system would then be provided with the names of a number of people, and work out which conference (and thus which shared interest) best unites that group. However, this is best for finding unifying information about a particular group of individuals, rather than determining the community structure of the entire population of authors.

3.2.2 Weighted Networks

There has also been work on community identification on weighted graphs. Newman [61] has approached the problem by treating weighted graphs as multi-graphs and then clustering though the use of maximum-flow techniques. The graph is first transformed into a multi-graph, creating duplicate edges in proportion to the weight of the original edge. Once this is accomplished, clustering the graph proceeds very much as it does in the approach taken by Flake, et al. [28] (described above).

3.2.3 Signed Networks

One last network type that has received relatively little attention is the signed network. A signed graph is simply a standard graph in which a sign (positive or negative) is associated with each edge. Signed networks present problems for clustering since approaches such as spectral clustering and flow-based clustering are based on the assumption that the weight of edges in the graph are positive. This means that for citation graphs that include a sign[3], many of the best clustering techniques are unavailable.

Clustering a signed graph requires that groups of positively linked vertices are clustered together, but negatively linked vertices are not. This is in addition to the typical clustering criteria that many positive edges should be within a cluster, but few should cross between clusters.

Doreian and Mrvar [23] have proposed a method to partition such signed graphs. Their method uses local optimization techniques to minimize the objective function $P(C)$ given as:

$$P(C) = \alpha \sum_{n} + (1 - \alpha) \sum_{p} \tag{3.7}$$

where

$$0 \leq \alpha \leq 1$$

C is a partition of the graph into k clusters (decided in advance), $\sum_n$ is the number of negative edges within clusters, $\sum_p$ is the number of positive edges crossing between clusters, and α is a weighting constant.

While their technique uses iterated local search, it is clear that any local optimization technique could be used instead, with the same objective function.

3.2.4 Toward Clustering Citation Graphs

While there has been some work on networks that go beyond the simplest types, no attempt has yet been made to automatically identify communities over a network as rich as a typed citation network. As was seen in Section 2, the richest citation graphs can be interpreted as having many of the problems addressed here. As they have multiple relationships, they can be seen as a multi-graph. Citations vary in their importance, so the graphs are weighted. In addition, as the citer's opinion of a cited article may be positive or negative, they are signed.

Straightforward combinations of the techniques mentioned above do not suffice when coping with such complicated graphs. To illustrate, Newman's approach to weighted graphs first converts them to multi-graphs. However, rich citation

[3] For example, the confirmative/negational distinction of Moravcsik and Murugesan [57] could be interpreted as a sign. See Section 2.2.1 for more details.

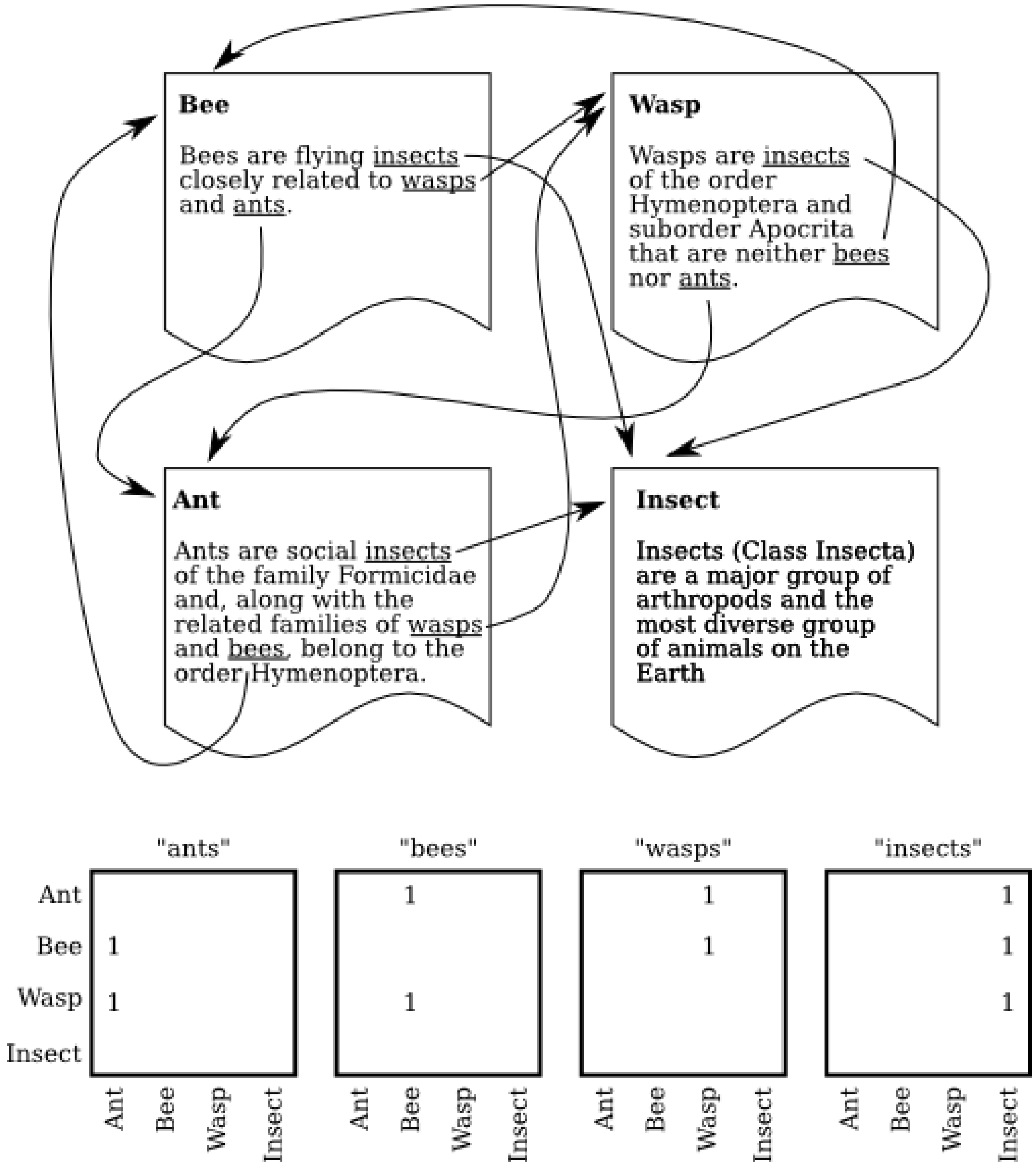
networks are already best understood as multi-graphs. Even if Newman's procedure were to be applied repeatedly to each edge (producing a multi-graph with many more duplicate edges), it is difficult to see how the multiple edges in a rich citation graph could be meaningfully reduced to a single flow.

3.2.5 TOPHITS

In an extension to the HITS algorithm by Kolda and Bader [49], some text analysis is added to identify not only the hubs and authority scores, but the *topic* of each of the found communities. Instead of simply working with simple Web links, an attempt is made to take into account the topic of each link. Since the topic is not actually known (since Web links are not classified), the TOPHITS (Topical HITS) algorithm uses words from within the link text to act as a stand-in for an actual topic. Thus, in a sense, the links are typed, where the "type" (or topic) is a list of word counts.

Instead of a two dimensional matrix representing the link structure, the TOPHITS algorithm starts with a 3-way tensor (a multi-dimensional generalization of a matrix). The first two dimensions of the tensor are the same as in the adjacency matrix of the original graph. The third dimension corresponds to the words in the associated text of the link. Thus M_{ijk} is only non-zero if there is a link from document i to document j containing the word associated with k. Figure 3.5 illustrates the structure of this tensor.

Instead of a singular-value decomposition (SVD) on the adjacency matrix, they perform a PARAFAC (parallel factors) decomposition on the adjacency tensor. This results in three vectors for each community (a hubs-and-authorities structure). The first two give the hubs and authorities scores to each document as in the HITS algorithm. The third gives scores to each word found in the link text. Thus, ideally, the most important terms for each topical community are identified.



Bee

Bees are flying insects closely related to wasps and ants.

Wasp

Wasps are insects of the order Hymenoptera and suborder Apocrita that are neither bees nor ants.

Ant

Ants are social insects of the family Formicidae and, along with the related families of wasps and bees, belong to the order Hymenoptera.

Insect

Insects (Class Insecta) are a major group of arthropods and the most diverse group of animals on the Earth

"ants"

	Ant	Bee	Wasp	Insect
Ant				
Bee	1			
Wasp	1			
Insect				

"bees"

	Ant	Bee	Wasp	Insect
Ant		1		
Bee				
Wasp		1		
Insect				

"wasps"

	Ant	Bee	Wasp	Insect
Ant			1	
Bee			1	
Wasp				
Insect				

"insects"

	Ant	Bee	Wasp	Insect
Ant				1
Bee				1
Wasp				1
Insect				

Figure 3.5: Example TOPHITS tensor. Text taken from Wikipedia articles for bee [81], ant [80], wasp [84], and insect [82].

Chapter 4

Applying TOPHITS to Typed Citations

Though algorithms such as HITS have been successful in identifying communities of documents on the Web, the goal of this thesis is the identification of communities from document collections with *typed* links. As the TOPHITS algorithm is an extension of HITS to links that are associated with some semantic information, it seems a worthwhile experiment to apply TOPHITS to a network of classified citations.

However, as pointed out by Kleinberg [48], scientific citation graphs do not have the hubs and authorities structure required for HITS (and by extension, TOPHITS) to work. Therefore, instead of attempting to find communities among documents directly, it is better to infer a social network among the authors that displays the requisite hubs and authorities structure, and then identify communities on the induced social network.

There are numerous ways that this could be accomplished. The simplest of these is simply aggregating all citations by author. This situation is summed up by the following rule:

$$cites_c(d_1, d_2) \rightarrow r_c(author(d_1), author(d_2)) \tag{4.1}$$

where r_c is a relationship based on the classification of the original citation, d_i are documents. Such an approach eliminates all temporal properties of the citation graph, and thereby allows the possibility of the hubs and authorities structure. In order to maintain the directed nature of the graph, the relation r will need to be something like an "is aware of" relationship. Additionally the graph could be weighted by counting the number of citations to a particular author,

$$w_{ij} = \frac{\text{\# of citations from } a_i \text{ to } a_j}{\text{\# of citations made by } a_i} \tag{4.2}$$

where w_{ijc} is the weight of the edge of class c from author i to author j. This is the likelihood of an author i citing author j.

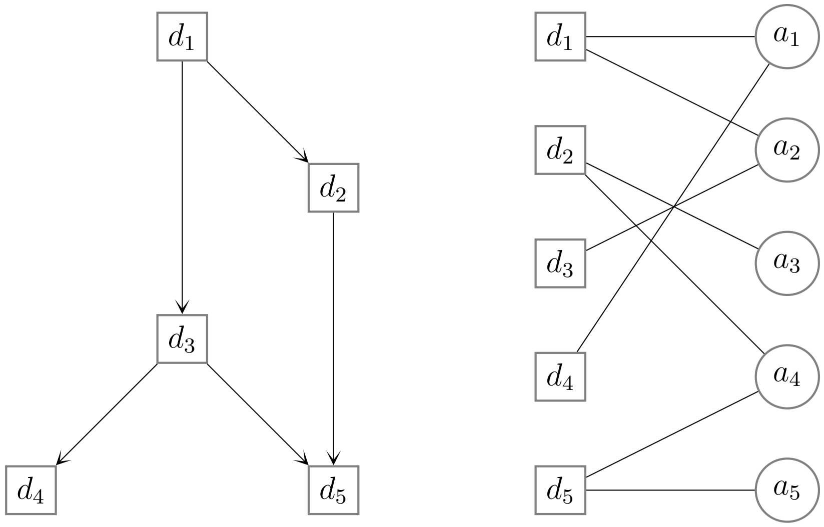


Figure 4.1: Hypothetical citation graph and affiliation network. The citation graph on the left, has an edge from d_i to d_j if document i cites document j. The affiliation network has an edge from d_i to a_j if j is an author of i. Note that the affiliation network is another method of representing the information in the collaboration graph.

A second possibility for a social network is derived from the citation network, but augmented with vertices representing each of the authors. In this augmented network of citations, there are now two kinds of nodes: document nodes and author nodes. Each document is linked to the nodes for the authors which contributed to it and to any document which is cited by it.

This is essentially a combination of two graphs. The first is the citation graph, and the second is an *affiliation network*, in which individuals are connected to groups (or in this case documents) with which they are affiliated. Thus, aspects of the citation and collaboration graphs are included in this network.

To derive the new network, begin with the citation network, add a vertex a_i for each author in the document collection:

- if an author a was credited in a document d:
  - add (d, a) to the edges
  - add (a, d) to the edges

Of the two graphs described above, the former was chosen as more appropriate for the application of the TOPHITS algorithm. Except for the difference in the graph, and the citation typing (explained fully in the next section), the TOPHITS algorithm was computed according to the method described by Kolda and Bader [49]. The top five communities were detected by performing a PARAFAC decomposition on the tensor. Despite the large size of the adjency tensor (270,665 × 270,665

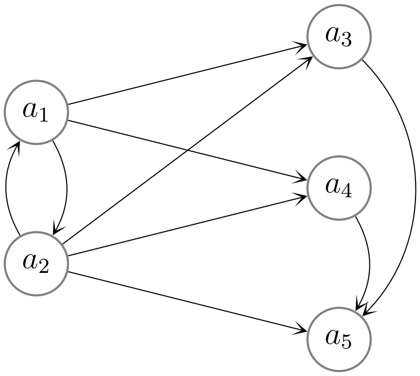


Figure 4.2: Author citation graph. There is an edge from a_i to a_j if a document written by author i cites a document written by author j. Derived from the graphs in Figure 4.1.

× 5), the calculations ran relatively quickly. The PARAFAC decomposition took almost exactly 6 minutes to run. This is largely due to the use of the sparse tensor toolbox [4, 5, 6] for MATLAB, which takes advantage of the sparse nature of the adjacency tensor.

4.1 The Simplified Classification Scheme

Since a corpus with fully classified citations does not exist, and the automatic classification of citations is still an active area of research [71], it is necessary to retreat to a scheme which can be computed automatically, with high accuracy. The document section within which the citation falls has been shown to be a useful feature in determining the more fine-grained classification of the citation. Since the document section is easily obtained, it will provide a useful, though simplified, stand-in for the full classification scheme.

Under the simplified scheme, each citation C can be expressed as a vector of n dimensions, where n is the number of sections in the document. Since many scientific documents are in IMRaD form, each citation vector will have 4 dimensions. So, the (multi)graph describing the network of citations can be represented as a 3-way tensor C. The first two dimensions of the tensor represent the set of documents, and the third represents the document sections. So, for each $c_{ijk} \in C$:

$$c_{ijk} = \begin{cases} 1 \text{ if there is a citation from } i \text{ to } j \text{ in section } k \\ 0 \text{ otherwise} \end{cases} \tag{4.3}$$

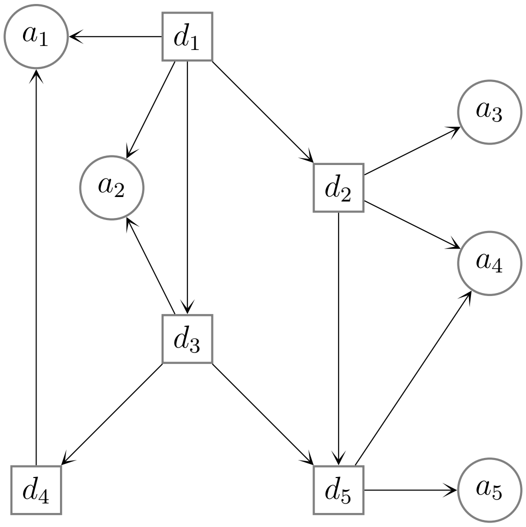


Figure 4.3: Example of an author-augmented citation graph. In this graph, the citation network is augmented with information taken from the affiliation graph. Derived from the graphs in Figure 4.1.

4.2 Implementation

4.2.1 Corpus

The initial corpus consisted of 48,630 full-text documents from the PubMed Central database. Full-text was required to extract the section in which the citations occurred.

The corpus contained a very large number of distinct section titles. Many of these could be meaningfully mapped to one of the IMRaD section types. An attempt was made to map each section title to one of five sections. The sections used were:

1. introduction
2. methods
3. results
4. discussion
5. other

Any section that contained one of the words from this list was assumed to map to that section. If a section title combined two or more of these section names, it was counted as being both. For example, a section titled "results and discussion" would be counted as both results and as discussion. To ensure a larger final corpus, several common sections were mapped to the IMRaD framework:

1. background → introduction
2. objectives → introduction
3. implementation → methods
4. case report → methods
5. algorithms → methods
6. utility → discussion
7. applications → discussion
8. future work → discussion
9. conclusion → discussion
10. contact → other
11. availability → other
12. author's contributions → other
13. appendix → other
14. abbreviations → other
15. acknowledgements → other
16. example → other

4.2.2 Author Identification

The documents in the PubMed collection contain markup identifying the authors of each paper, as well as their institutional affiliation. However, authors are not assigned a unique ID that persists from document to document. As such, identifying a particular individual as being the author of more than a single paper is difficult.

For two authors to have been considered the same individual, the following two conditions *must* have held:

1. They shared the same surname
2. They shared the same first initial

Because it is possible for two distinct authors to have the same name, at least one of the following additional conditions must have been met:

1. They shared the same institutional affiliation
2. They shared the same email address

The text describing a given institution had a substantial amount of variation. For example, one document may list the author's department and institution, whereas another might give only the name of the institution.

Linking records by performing this kind of string matching is a research topic unto itself, and beyond the scope of this thesis. While more sophisticated tools are available (see, for example [16]), for the purposes of this thesis, the following simple matching algorithm was used:

1. Numbers, punctuation, postal codes, etc. were stripped from the text.
2. The text was converted to lower case
3. After normalization, the resulting strings were compared using a variation of the Winkler distance metric [85] described below. Two institutions were considered to be the same if the Winkler distance was greater than 0.75.

Winkler distance is an approximate string matching algorithm, based on Jaro distance [44]. Jaro distance is defined as:

$$d_j = \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) \tag{4.4}$$

where s_1 and s_2 are the two strings to be compared, m is the number of matching characters, and t is the number of transpositions.

Winkler distance is different in that matches that occur early in the strings are given greater weight. However, for matching affiliations, the most important matches are more likely to occur at the end of the string, since relatively unimportant words such as "university" and "department" occur early, and relatively important words such as country names occur at the end. Thus the strings were first inverted, then unmodified Winkler was used.

Not all documents gave the institutional affiliation or email address for every author. In this case, the co-authorships of each author were checked against other authors with a matching name. In the case that the authors shared at least one matching co-author, they were assumed to be the same.

A number of improvements to this algorithm are possible. First, using Winkler distance is very conservative, given the text of the affiliations. Since affiliations vary significantly in length from document to document (even for the same author), the Winkler distance is often low, even for matching affiliations. Moreover, matching certain words is far more important than matching others. For example, the word "university" occurs in the vast majority of affiliations, and thus should be ignored. On the other hand, particular names of the institutions are significantly more important.

Often, the institutional affiliation includes common information. For example, in the PubMed Central corpus, the majority of affiliations include a department name, the name of a university, hospital, or government agency, a state or province, and a country. Sometimes street addresses or postal codes are included as well.

In addition to problems with the matching of institutional affiliations, the use of a more sophisticated measure of co-authorship would significantly reduce the number of identifications of distinct authors as identical. In practice, the simplistic method described here works reasonably well for less common names. However, this leads to problems with many common surnames, especially if only the author's first initials are available. Particularly difficult are Asian names, especially Chinese and Korean names. In China, for example, it is estimated that 85% of the population share only 129 surnames. Compounding this problem is the fact that multiple Chinese surnames are romanized identically by the Pinyin romanization system [70]. Korean surnames are even more problematic, with approximately 54% of the population of Korea sharing three surnames (Kim, Lee, and Park) according to 2000 figures from the Korean National Statistical Office [83].

4.2.3 Generation of Citation Graph

Unlike authors and their institutional affiliations, many articles have a unique ID that persists from document to document. Two of these are used in the PubMed Central document collection: PubMed ID, and DOI (Digital Object Identifier). Every article in the PubMed Central collection has a PubMed ID, but some cited articles do not.

There were a total of 871,110 citations in the document collection. Of these, the majority (721,036) were identified by PubMed ID or DOI. The largest number of citations to any particular article was 746.

Relatively few of the cited papers are in the original PubMed Central document collection. Of 486,873 documents cited, 6,579 were in the original collection, and thus full-text was available.

4.2.4 The Reduced Corpus

In order to reduce the size of the adjacency matrix, and thus keep down the processing time and memory usage, a decision was made to focus solely on those articles that cited another article within the collection. There were a total of 10,128 articles that either cited or were cited by another document in the full corpus. This was then augmented by adding references to any articles that were cited by a document within the set of inter-citing documents, and had a valid unique identifier (PubMed ID or DOI).

In total, there were 270,665 article vertices in the reduced graph, and 408,777 citation edges. The total number of unique authors was 956,859, with a median of 4 authors per article. A few documents contained a very large number of authors, and were later excluded from the graph.

Chapter 5

Experimental Results

Since the algorithm was run on a graph over authors, rather than documents, it is somewhat more difficult to validate that the found communities are natural. One possibility would be to calculate the average distance between authors with high hub and authority scores on the collaboration graph. However, since such networks typically display the small-world property[1], paths between authors would likely be short regardless of the quality of the clusters. Furthermore, such an analysis would tell us nothing of whether using the section data provided any advantage over simply leaving the citations untyped.

Assuming that one of the link types was scored more highly with respect to a given community, one would expect that the relationships between highly-scored authors within that community would be characterized by that link type. Since the most salient difference between links typed by section is the distinction between methodological versus conceptual citations, one would expect that a community that scored relatively highly on the "method" section would be linked more by methodology than broad research topic.

Since we do not have independent information about the research topics or methodologies of the various authors, we need to extract this information from the corpus. There are two possible ways that this might be accomplished. The first is to examine the full text of the documents written by the highest scoring authors, and perform some manner of topic identification (for example, keyword extraction [42, 50], latent semantic analysis [22], latent Dirichlet allocation [9], etc.).

The second possibility is to make use of the fact that all documents indexed by PubMed have been hand-annotated with terms from the MeSH (Medical Subject Headings) vocabulary. Using MeSH avoids the difficulties inherent in natural language processing, but it is possible that there is some term or concept, not present in MeSH, yet important to the discovered community.

In order to provide some description of the communities, we decided to identify MeSH terms that characterized the documents authored by the highest scoring

[1] For a discussion of the small-world property, see Watts and Strogatz [79].

members of the identified communities. For each score (i.e., hub score and authority score), the top 100 authors were identified, and the MeSH terms describing their documents were extracted. MeSH terms were scored based on the number of documents in which they occurred, as well as the score of the authors in question. The score s_t for a given term t is given by:

$$s_t = \sum_{a \in A} \#\text{ docs in docs}(a)\text{ containing } t \tag{5.1}$$

where A is the set of high-scoring authors, docs(a) is the set of documents that have a as an author.

Tables 5.1 and 5.2 show the hub, authority, and term scores for the top two identified communities.

Hubs		Authorities		Sections	
Animals	3.97	Sequence Alignment	3.79	Methods	0.88
Drosophila melanogaster	2.38	Software	3.03	Results	0.33
Genome	2.08	Algorithms	2.78	Discussion	0.25
Genes, Insect	1.88	Amino Acid Sequence	2.02	Introduction	0.21
Drosophila	1.49	Molecular Sequence Data	1.77	Other	0.06
Computational Biology	1.19	Proteins	1.77		
Sequence Analysis, DNA	1.19	Internet	1.26		
Humans	0.99	Sequence Analysis, Protein	1.26		
Databases, Genetic	0.89	Databases, Protein	1.01		
Drosophila Proteins	0.89	Humans	1.01		

Table 5.1: Hub, authority, and section scores for the principle PubMed community

Hubs		Authorities		Sections	
Oligonucleotide Array Sequence Analysis	8.57	Oligonucleotide Array Sequence Analysis	15.95	Results	0.59
Humans	7.87	Humans	15.67	Introduction	0.50
Gene Expression Profiling	5.40	Gene Expression Profiling	10.82	Discussion	0.48
Female	2.47	Saccharomyces cerevisiae	8.83	Methods	0.41
Saccharomyces cerevisiae	2.35	Saccharomyces cerevisiae Proteins	5.13	Other	0.04
Gene Expression Regulation, Neoplastic	1.88	Gene Expression Regulation, Neoplastic	3.99		
RNA, Messenger	1.88	Transcription Factors	3.99		
Animals	1.76	Animals	3.99		
Transcription, Genetic	1.64	Female	3.70		
Cluster Analysis	1.53	Genome, Fungal	3.70		

Table 5.2: Hub, authority, and section scores for the second PubMed community

Looking first to Table 5.1 it can be seen that citations in the methods section are relatively important. That is, the methods section can be said to characterize the community in some way. The second community (Table 5.2), however, cannot be said to have any section dominating the community.

Returning to Table 5.1, looking at the authorities column, one notices that the top-ranked MeSH terms are predominantly associated with gene sequencing technologies, primarily computational techniques. This is exactly what would be expected of a community that was primarily held together by their sharing of techniques. Even more interesting is that the hubs do *not* share this property, at least not to the same degree.

One possible characterization of the first community is one in which a number of disparate biomedical researchers, working on somewhat different topics (with some focus on *Drosophila* (a.k.a. the common fruit fly, a common model organism in genetic research), but all (or most) of whom rely on genome sequencing technologies. Each citer is a hub, and all or most of the citers cite another set of documents, all of which provide a common technique. If this is the case, this is a clear case of a "methodological community", that is, one that is bound by a common method, but not necessarily a common research topic.

Returning to the second extracted community, no particular citation section stands out. Examining the top scoring MeSH terms for this community shows more unity in terms of topic, as well as method. In this case, many of the terms are related to fungal genetics, especially *Saccharomyces cerevisiae* (a.k.a. Brewer's Yeast), a commonly used model organism for genetics, for which the entire genome has been sequenced.

Tables of top-scoring MeSH terms for the remaining three communities can be found in Appendix B.

Chapter 6

Conclusions and Future Work

While the results presented in Chapter 5 are not conclusive, they do provide some evidence that typed links are helpful in characterizing communities of scholars. That being said, there are numerous avenues for future research on networks of typed links. As type systems become more complex, many algorithms used for untyped links will no longer suffice, and will have to be modified. Basic research needs to be done with regard to the motivations behind linking on the Web, in order to build ontologies of link types. Finally, significantly more sophisticated techniques can be used to generate social networks from citation networks than the ones that have been used here. Producing better networks over authors can only improve the quality of identified communities.

6.1 Community Identification and Typed Links

While spectral techniques such as TOPHITS provide one way forward for typed link analysis, they are limited to particular link classification schemes, namely those that list a number of separate categories or those that specify a number of independent dimensions. Since spectral methods make fundamental assumptions about the weights of edges in the graph (i.e., that they are non-negative), citation classification schemes that include negative valences will continue to be a problem.

While there has been some research into clustering such graphs, there has been no attempt to find richer community structures, as is done in HITS. The classification of author attitude, also known as *sentiment analysis*, is a research topic that is currently receiving significant interest (see, for example, Pang, Lee and Vaithyanathan [67]). Since sentiment is fundamentally relational, it is likely that Web links will be classified in this way.

Aside from valence, other aspects of the classification scheme presented in Section 2.3 create problems for HITS-like methods. As we have seen, generalizing the adjacency matrix to an adjacency tensor, where the third dimension contains information about the types allows the specification of multiple types for a given edge. However, this does not account for classification schemes in which there are multiple types with constituent parts. It is possible that using a 4-dimensional (or higher) tensor might remedy this problem, but that remains to be seen.

6.2 Link Classification

While much effort has been put into identifying the motivations underlying citation, significantly more work needs to be done to develop a true ontology of web links. As reviewed in Section 2.4, to date there have been only a few studies of web linking behaviour, and they have focused on a limited range of links. Further studies of existing web links are required.

Once an appropriate ontology of types for Web linking is available, research can begin on adapting automated citation classification techniques to the Web. Furthermore, the adoption of a link classification scheme for the Web might spur Web users to link in different ways. For example, many people prefer not to link to sites with which they disagree (or dislike). As algorithms such as PageRank are unaware of the valence of links, they are treated identically with other links. Thus, providing a link to a page, negative or positive, will result in an increased search ranking, and thus more page-views. If PageRank could take valence into account, it might provide more incentive (or at least less disincentive) to provide links to pages with opposing viewpoints.

6.3 Social Network Generation

There is a substantial literature concerning the generation of social networks from available data (for example, see [1, 51]). However, very little work has been done on generating such networks from networks of typed links. While the approach we have taken here has been quite naïve, it is possible to take significantly more sophisticated approaches to social network generation.

Possibly the best possible approach to social network generation treats every citation as evidence of a social relationship amongst the authors. When an author writes an article, that article is aimed toward an intended audience. This audience is composed of individuals from one or more of the communities to which the author belongs. When a paper is cited, we gain insight into a single individual that is in the *actual* audience of the article.

As the number of citations between two authors increases, the more certain we can be that they are aware of each others research, and thus should be connected within the social network of authors. This is especially true if there is a history of reciprocal citations between two authors. The particular community to which they belong will be indicated by the rhetorical purpose of the citation.

This leads to a Bayesian approach in which we treat each citation as evidence that citer and citee are socially related. To accomplish this, the probabilities of making a citation given a particular social relationship would be needed. While it is unclear whether enough data is available to learn such probabilities, one possible first avenue would be to learn from a known social network i.e., the collaboration graph. The downside of this approach is that the collaboration graph only includes a single social relationship (collaborates with), and thus may miss many other undocumented relationships.

There exists a substantial body of research on statistical-relational learning that addresses this problem (for a review see [35]). In a particularly relevant paper, Popescul and Ungar [69] have applied such models to scientific citation and collaboration data to recommend new citations for documents in the collection. Similarly, their model could be applied to recommending collaborations among authors. However, it remains to be seen how such an approach could be adapted to relations with complex types as presented in Section 2.3.

6.4 Conclusion

The exploration of community identification on graphs with typed links has only just begun. Unusually for recent information retrieval research, this has largely been due to a lack of data. However, as we move toward the Semantic Web, this is likely to change. Typed links do appear to provide additional information about scholarly communities, and it is likely that this is true for the Web as well. As semantic links become more prevalent, it is likely that new algorithms will be needed to exploit them as well.

APPENDICES

Appendix A

Citation Classification Schemes

A.1 Garzone and Mercer

Negational Type Categories :

1. Citing work totally disputes some aspect of cited work.
2. Citing work partially disputes some aspect of cited work.
3. Citing work is totally not supported by cited work.
4. Citing work is partially not supported by cited work.
5. Citing work disputes priority claims.
6. Citing work corrects cited work.
7. Citing work questions cited work.

Affirmational Type Categories :

8. Citing work totally confirms cited work.
9. Citing work partially confirms cited work.
10. Citing work is totally supported by cited work.
11. Citing work is partially supported by cited work.
12. Citing work is illustrated or clarified by cited work.

Assumptive Type Citations :

13. Citing work refers to assumed knowledge which is general background.
14. Citing work refers to assumed knowledge which is specific background.
15. Citing work refers to assumed knowledge in an historical account.
16. Citing work acknowledges cited work pioneers.

Tentative Type Categories :

17. Citing work refers to tentative knowledge.

Methodological Type Categories :

18. Use of materials, equipment, or tools.
19. Use of theoretical equation.
20. Use of methods, procedures, and design to generate results.
21. Use of conditions and precautions to obtain valid results.
22. Use of analysis method on results.

Interpretational/Developmental Type Categories :

23. Used for interpreting results.
24. Used for developing new hypothesis or model.
25. Used for extending an existing hypothesis or model.

Future Research Type Categories :

26. Used in making suggestions of future research.

Use of Conceptual Material Type Categories :

27. Use of definition.
28. Use of numerical data.

Contrastive Type Categories :

29. Citing work contrasts between the current work and other work.
30. Citing work contrasts other works with each other.

Reader Alert Type Categories :

31. Citing work makes a perfunctory reference to cited work.
32. Citing work points out cited works as bibliographic leads.
33. Citing work identifies eponymic concept or term of cited work.
34. Citing work refers to more complete descriptions of data or raw sources of data.

A.2 Radoulov

Confirms This component is present when the citing paper somehow confirms or validates the cited work. An example of this case is “The assignment of disulfides in the C-terminal domain experimentally validates the primary disulfide pattern predicted for NTR modules ([B49]).”

Supports When the citing work supports some aspect of the cited work. Example: “This protein has been identified previously as a nuclear serine/threonine kinase that interacts with the NK homeodomain transcription factor ([B46]), acts as a corepressor for the NK homeodomain, and cooperates with Groucho and HDAC-1 in enhancing transcriptional repression ([B47]).. Here, support is shown by implicitly agreeing with the previous results.”

Illustrates/Clarifies One work clarifies or illustrates something from a different work. Example: “For example, FIX Q50P has been studied by two different groups ([B43] , [B44]).”

Interprets results When one work is used to interpret results of another. Example: “Because EGF1 of activated protein C has a major loop inserted at a position corresponding to FIXa residue 54, it seems unlikely that this part of EGF1 in FIXa makes a direct contact with FVIIIa ([B15] , [B19]).”

Extends model A model is either extended or created from finding in the cited work. Example: “In this model, the key interacting regions of FIXa and FVIIIa can be aligned as previously reported with only minor reorientations ([B13] , [B32]).”

Contrasts When two works are compared. Example: “Surprisingly, mutation of the first two leucines in the LXXLL motif decreased steroid binding capacity and transcriptional activity without altering receptor levels, cell-free steroid binding affinity, or hsp90 binding ([B25]).”

Mentions in Passing The work is cited as a perfunctory reference. Example: “The mammalian BNaC/ASIC branch of the superfamily contains four genes, encoding at least six isoforms: BNaC1 (also known as BNC1, MDEG, and ASIC2) ([B2]) and its differentially spliced isoform, BNaC1 (MDEG2) ([B17]); BNaC2 (ASIC or ASIC1) ([B4] , [B18]) and its differentially spliced isoform, BNaC2 (ASIC) ([B19]); DRASIC (ASIC3 or TNaC) ([B20]); and ASIC4 (SPASIC) ([B24] , [B25]).”

Future Research Points to future research. Example: “An open question is whether the described disassembly of transcriptional regulatory complexes by p23 requires ATP ([B61]), as the requirement of hsp90 or hsp70 for the effect of p23 remains to be elucidated.”

Uses Use of a method, equation, product, etc. Example: “To study the NF-Y-TFIID connections, we employed the mouse MHC class II Ea promoter system ([B51] , [B52]).”

General Background Background that is not necessarily needed to understand the citing paper. Example: “Hitherto, the search for paxillin-binding proteins has involved either yeast 2-hybrid screens ([B8]) or GST-fusion protein pull-down assays ([B6] , [B10] , [B12] , [B26] , [B30]).”

Specific Background Background that is specific for the citing article. Example: "The p110 isoforms of PI 3-kinase played significant roles in cell migration, and differential activation of specific p110 isoforms is responsible for particular signaling events in different cell types ([B32] , [B33])."

Historical This is also a background component, but is mentioned chronologically, Example: "Earlier reports have shown that pV and tumor necrosis factor induced NFB activation in Jurkat cells, and only pV-induced activation of NFB is inhibited by wortmannin ([B21])."

Pioneering Work Citing work of pioneers in the field. This is another category that is very difficult to annotate without having an in-depth knowledge of the field. Example: "Recent evidence indicates that Sina, together with phyllo- pod, promotes the ubiquitin/proteasome-dependent degradation of tramtrack, a negative regulator of neuronal differentiation ([B29] , [B30])."

Related work/Bibliographic Lead The author either describes related work or gives leads for further reading. Example: "Consistent with previous reports ([B35] , [B37]), myc-tagged Siah-2 was found to be expressed at a relatively low level in transfected PC12 cells, perhaps as a result of self-regulating its own stability (see "Discussion")."

Concept A use of a model, definition, hypothesis. Example: "This staining showed strong colocalization with EEA1 (Fig. F5, C and F), which is consistent with the idea that mVps4 regulates the morphology and the transport functions of endosomes ([B54])."

Method Use of method. Example: "Sequence analyses show that Hrs, Eps15, STAM1, and STAM2 contain UIMs ([B55])."

Product Use of a product or material. Example: "To do this, we used a recently described phage system that displays a highly diverse and random assortment of short peptides fused to the C terminus of the M13 gene-8 major coat protein ([B9])."

Data Use or analysis of data. E.g. "As has been found with virtually all previously examined ligands for type 1 PDZ domains ([B3]), the C-terminal residue (position 0) was found to be hydrophobic."

Direction This component represents the three possible directions of a citation: (i) the citing paper describes or uses material from the cited paper (most common type); (ii) two works are compared to each other, i.e., a direction between two cited papers; (iii) and the final, and most interesting, citation direction is from a cited paper to the citing paper. This last citation direction is not very common, but it does occur in papers that are published almost simultaneously. An example of such a citation is: "Work on applying machine learning techniques for automatic citation classification is currently underway (Teufel et al., 2006)".

How sure? This component represents the scale of the annotators certainty with his classification. Although this was originally meant to judge whether the labelled citation should be included in testing or training sets, it can also be used to keep track of the strength of the given relationship between the papers.

Appendix B

MeSH and Section Scores

	Community				
Section	1	2	3	4	5
Introduction	0.21	0.50	0.15	0.29	0.41
Methods	0.88	0.41	0.07	0.26	0.28
Results	0.33	0.59	0.68	0.66	0.64
Discussion	0.25	0.48	0.72	0.64	0.58
Other	0.06	0.04	0.01	0.02	0.02

Table B.1: Section values for all communities

Term	Score	Term	Score
Animals	3.97	Models, Genetic	0.40
Drosophila melanogaster	2.38	Expressed Sequence Tags	0.40
Genome	2.08	Physical Chromosome Mapping	0.40
Genes, Insect	1.88	Insect Proteins	0.40
Drosophila	1.49	MicroRNAs	0.30
Computational Biology	1.19	Heterochromatin	0.30
Sequence Analysis, DNA	1.19	Genes	0.30
Humans	0.99	Ligases	0.30
Databases, Genetic	0.89	In Situ Hybridization	0.30
Drosophila Proteins	0.89	Gene Expression Regulation, Developmental	0.30
Molecular Sequence Data	0.79	Caenorhabditis elegans	0.30
Membrane Proteins	0.60	Algorithms	0.30
Euchromatin	0.60	Gene Expression Regulation	0.30
Multigene Family	0.60	Chromatin	0.30
Databases, Factual	0.50	Ubiquitin-Protein Ligases	0.30
DNA, Complementary	0.50	Terminology as Topic	0.30
DNA Transposable Elements	0.50	Species Specificity	0.30
Gene Library	0.50	Transcription, Genetic	0.30
Evolution, Molecular	0.50	Receptors, Notch	0.30
Conserved Sequence	0.40	Phenotype	0.30
Contig Mapping	0.40	Transcription Factors	0.30
Mice	0.40	Nerve Tissue Proteins	0.30
Base Sequence	0.40	Software	0.30
Cloning, Molecular	0.40	Database Management Systems	0.20
Cluster Analysis	0.40	Forecasting	0.20

Table B.2: Top 50 MeSH term scores for hubs of community 1.

Term	Score	Term	Score
Sequence Alignment	3.79	Amino Acid Motifs	0.25
Software	3.03	Systems Integration	0.25
Algorithms	2.78	Structural Homology, Protein	0.25
Amino Acid Sequence	2.02	Vocabulary, Controlled	0.25
Molecular Sequence Data	1.77	Vibrio cholerae	0.25
Proteins	1.77	Models, Molecular	0.25
Internet	1.26	Membrane Proteins	0.25
Sequence Analysis, Protein	1.26	Management Information Systems	0.25
Databases, Protein	1.01	Fungal Proteins	0.25
Humans	1.01	Ribonucleases	0.25
Genomics	0.76	Phylogeny	0.25
Databases, Factual	0.76	Nucleic Acids	0.25
Animals	0.76	Mutation	0.25
Sensitivity and Specificity	0.76	Eukaryotic Cells	0.25
Reproducibility of Results	0.76	Conserved Sequence	0.25
Data Display	0.5	Archaeal Proteins	0.25
Databases, Genetic	0.5	Interleukin-1	0.25
Sequence Analysis	0.5	Globins	0.25
Bacterial Proteins	0.5	Genome, Bacterial	0.25
Benchmarking	0.5	Gene Expression Profiling	0.25
Computational Biology	0.5	Research Design	0.25
Sequence Homology, Amino Acid	0.5	Repetitive Sequences, Amino Acid	0.25
Evolution, Molecular	0.5	Protein Structure, Secondary	0.25
User-Computer Interface	0.5	Sequence Homology, Nucleic Acid	0.25
Quality Control	0.5	Sequence Homology	0.25

Table B.3: Top 50 MeSH term scores for authorities of community 1.

Term	Score	Term	Score
Oligonucleotide Array Sequence Analysis	8.57	Middle Aged	0.94
Humans	7.87	Cell Cycle	0.82
Gene Expression Profiling	5.4	Signal Transduction	0.82
Female	2.47	Mice	0.82
Saccharomyces cerevisiae	2.35	Nucleic Acid Hybridization	0.82
RNA, Messenger	1.88	Genome, Human	0.7
Gene Expression Regulation, Neoplastic	1.88	Neoplasms	0.7
Animals	1.76	Fungal Proteins	0.7
Transcription, Genetic	1.64	Escherichia coli	0.7
DNA, Complementary	1.53	Survival Analysis	0.7
Cluster Analysis	1.53	Mutation	0.7
Breast Neoplasms	1.53	Neoplasm Proteins	0.7
Gene Expression	1.29	Fibroblasts	0.7
Gene Expression Regulation	1.29	Algorithms	0.7
Male	1.29	Molecular Sequence Data	0.7
Saccharomyces cerevisiae Proteins	1.17	Base Sequence	0.7
Tumor Cells, Cultured	1.17	Genome	0.59
Adult	1.17	Cell Line	0.59
Gene Expression Regulation, Fungal	1.06	Prognosis	0.59
DNA-Binding Proteins	1.06	Trans-Activators	0.59
Multigene Family	0.94	Variation (Genetics)	0.59
Transcription Factors	0.94	DNA, Bacterial	0.59
Phenotype	0.94	DNA, Fungal	0.59
Genes, Fungal	0.94	Aged	0.47
Genome, Fungal	0.94	DNA, Neoplasm	0.47

Table B.4: Top 50 MeSH term scores for hubs of community 2.

Term	Score	Term	Score
Oligonucleotide Array Sequence Analysis	15.95	Information Storage and Retrieval	2.28
Humans	15.67	DNA, Complementary	2.28
Gene Expression Profiling	10.82	Molecular Sequence Data	2.28
Saccharomyces cerevisiae	8.83	Genomics	1.71
Saccharomyces cerevisiae Proteins	5.13	Base Sequence	1.71
Transcription Factors	3.99	Mice	1.71
Gene Expression Regulation, Neoplastic	3.99	Signal Transduction	1.71
Animals	3.99	Genome, Human	1.71
Female	3.7	Binding Sites	1.71
Genome, Fungal	3.7	Algorithms	1.71
Gene Expression	3.42	Sequence Analysis, DNA	1.42
Genes, Fungal	3.42	Nuclear Proteins	1.42
Gene Expression Regulation	3.13	Neoplasms	1.42
Breast Neoplasms	2.85	Databases, Factual	1.42
Transcription, Genetic	2.85	Nucleic Acid Hybridization	1.42
Multigene Family	2.56	Chromosomes, Fungal	1.42
Software	2.56	Trans-Activators	1.42
DNA-Binding Proteins	2.56	DNA, Fungal	1.42
Databases, Genetic	2.56	Adenocarcinoma	1.42
Gene Expression Regulation, Fungal	2.56	Cell Cycle	1.42
Internet	2.28	Lymphoma, B-Cell	1.42
RNA, Messenger	2.28	Computational Biology	1.42
Tumor Cells, Cultured	2.28	Male	1.42
Cluster Analysis	2.28	Neoplasm Proteins	1.42
Fungal Proteins	2.28	Phenotype	1.42

Table B.5: Top 50 MeSH term scores for authorities of community 2.

Term	Score	Term	Score
Humans	14.08	Cells, Cultured	2.17
HIV-1	8.66	Mutation	2.17
Transcription, Genetic	8.3	Cyclin-Dependent Kinase Inhibitor p21	2.17
Gene Products, tat	7.58	Cyclin-Dependent Kinases	2.17
tat Gene Products, Human Immunodeficiency Virus	7.22	HIV Long Terminal Repeat	1.81
Hela Cells	5.05	Protein Binding	1.81
Cell Line	4.69	HIV Infections	1.81
Human T-lymphotropic virus 1	4.33	CDC2-CDC28 Kinases	1.81
Molecular Sequence Data	4.33	Virus Replication	1.81
Gene Products, tax	3.61	RNA Polymerase II	1.81
Cell Cycle	3.61	Virus Integration	1.81
Transcription Factors	3.61	Nuclear Proteins	1.44
Cyclins	3.25	Gene Expression Regulation	1.44
Trans-Activation (Genetics)	3.25	G1 Phase	1.44
Animals	2.89	Mice	1.44
Amino Acid Sequence	2.89	Kinetics	1.44
Phosphorylation	2.89	Cell Cycle Proteins	1.44
Transfection	2.53	Base Sequence	1.44
Gene Expression Regulation, Viral	2.53	Cyclin E	1.44
Cyclin-Dependent Kinase 2	2.53	Cell Division	1.44
Promoter Regions (Genetics)	2.53	Recombinant Fusion Proteins	1.44
Apoptosis	2.17	S Phase	1.44
Chromatin	2.17	Signal Transduction	1.44
T-Lymphocytes	2.17	Tumor Cells, Cultured	1.44
Binding Sites	2.17	Oligonucleotide Array Sequence Analysis	1.44

Table B.6: Top 50 MeSH term scores for hubs of community 3.

Term	Score	Term	Score
Transcription, Genetic	1.48	Cyclin-Dependent Kinase Inhibitor p21	0.39
Gene Products, tat	1.35	Cyclin-Dependent Kinases	0.39
tat Gene Products, Human Immunodeficiency Virus	1.29	HIV Long Terminal Repeat	0.32
Hela Cells	0.9	Protein Binding	0.32
Cell Line	0.84	HIV Infections	0.32
Human T-lymphotropic virus 1	0.77	CDC2-CDC28 Kinases	0.32
Molecular Sequence Data	0.77	Virus Replication	0.32
Gene Products, tax	0.64	RNA Polymerase II	0.32
Cell Cycle	0.64	Virus Integration	0.32
Transcription Factors	0.64	Nuclear Proteins	0.26
Cyclins	0.58	Gene Expression Regulation	0.26
Trans-Activation (Genetics)	0.58	G1 Phase	0.26
Animals	0.51	Mice	0.26
Amino Acid Sequence	0.51	Kinetics	0.26
Phosphorylation	0.51	Cell Cycle Proteins	0.26
Transfection	0.45	Base Sequence	0.26
Gene Expression Regulation, Viral	0.45	Cyclin E	0.26
Cyclin-Dependent Kinase 2	0.45	Cell Division	0.26
Promoter Regions (Genetics)	0.45	Recombinant Fusion Proteins	0.26
Apoptosis	0.39	S Phase	0.26
Chromatin	0.39	Signal Transduction	0.26
T-Lymphocytes	0.39	Tumor Cells, Cultured	0.26
Binding Sites	0.39	Oligonucleotide Array Sequence Analysis	0.26

Table B.7: Top 50 MeSH term scores for authorities of community 3.

Term	Score	Term	Score
Animals	9.99	Models, Genetic	1.00
Drosophila melanogaster	5.99	Expressed Sequence Tags	1.00
Genome	5.24	Physical Chromosome Mapping	1.00
Genes, Insect	4.74	Insect Proteins	1.00
Drosophila	3.75	MicroRNAs	0.75
Computational Biology	3.00	Heterochromatin	0.75
Sequence Analysis, DNA	3.00	Genes	0.75
Humans	2.50	Ligases	0.75
Databases, Genetic	2.25	In Situ Hybridization	0.75
Drosophila Proteins	2.25	Gene Expression Regulation, Developmental	0.75
Molecular Sequence Data	2.00	Caenorhabditis elegans	0.75
Membrane Proteins	1.50	Algorithms	0.75
Euchromatin	1.50	Gene Expression Regulation	0.75
Multigene Family	1.50	Chromatin	0.75
Databases, Factual	1.25	Ubiquitin-Protein Ligases	0.75
DNA, Complementary	1.25	Terminology as Topic	0.75
DNA Transposable Elements	1.25	Species Specificity	0.75
Gene Library	1.25	Transcription, Genetic	0.75
Evolution, Molecular	1.25	Receptors, Notch	0.75
Conserved Sequence	1.00	Phenotype	0.75
Contig Mapping	1.00	Transcription Factors	0.75
Mice	1.00	Nerve Tissue Proteins	0.75
Base Sequence	1.00	Software	0.75
Cloning, Molecular	1.00	Database Management Systems	0.50
Cluster Analysis	1.00	Forecasting	0.50

Table B.8: Top 50 MeSH term scores for hubs of community 4.

Term	Score	Term	Score
Animals	4.23	Models, Genetic	0.42
Drosophila melanogaster	2.54	Expressed Sequence Tags	0.42
Genome	2.22	Physical Chromosome Mapping	0.42
Genes, Insect	2.01	Insect Proteins	0.42
Drosophila	1.59	MicroRNAs	0.32
Computational Biology	1.27	Heterochromatin	0.32
Sequence Analysis, DNA	1.27	Genes	0.32
Humans	1.06	Ligases	0.32
Databases, Genetic	0.95	In Situ Hybridization	0.32
Drosophila Proteins	0.95	Gene Expression Regulation, Developmental	0.32
Molecular Sequence Data	0.85	Caenorhabditis elegans	0.32
Membrane Proteins	0.63	Algorithms	0.32
Euchromatin	0.63	Gene Expression Regulation	0.32
Multigene Family	0.63	Chromatin	0.32
Databases, Factual	0.53	Ubiquitin-Protein Ligases	0.32
DNA, Complementary	0.53	Terminology as Topic	0.32
DNA Transposable Elements	0.53	Species Specificity	0.32
Gene Library	0.53	Transcription, Genetic	0.32
Evolution, Molecular	0.53	Receptors, Notch	0.32
Conserved Sequence	0.42	Phenotype	0.32
Contig Mapping	0.42	Transcription Factors	0.32
Mice	0.42	Nerve Tissue Proteins	0.32
Base Sequence	0.42	Software	0.32
Cloning, Molecular	0.42	Database Management Systems	0.21
Cluster Analysis	0.42	Forecasting	0.21

Table B.9: Top 50 MeSH term scores for authorities of community 4.

Term	Score	Term	Score
Animals	18.12	Cluster Analysis	2.13
Mice	15.19	Transcription Initiation Site	2.13
DNA, Complementary	10.92	Reverse Transcriptase Polymerase Chain Reaction	2.13
Humans	8.26	DNA Primers	1.87
Transcription, Genetic	7.73	Exons	1.87
Gene Library	5.86	Genomics	1.87
Genome	5.06	Genes	1.87
Gene Expression Profiling	5.06	Genome, Human	1.87
RNA, Messenger	4.80	Protein Structure, Tertiary	1.87
Cloning, Molecular	4.80	Sequence Alignment	1.87
Promoter Regions (Genetics)	4.26	Multigene Family	1.87
Oligonucleotide Array Sequence Analysis	4.26	Reproducibility of Results	1.60
Computational Biology	4.00	Cell Line	1.60
Base Sequence	4.00	Databases, Nucleic Acid	1.60
Expressed Sequence Tags	3.73	Transcription Factors	1.60
Databases, Genetic	3.73	Open Reading Frames	1.60
Gene Expression Regulation	3.46	Genes, Plant	1.60
Sequence Analysis, DNA	3.46	Nucleic Acid Hybridization	1.60
Chromosome Mapping	2.93	Membrane Proteins	1.60
Arabidopsis	2.66	Mice, Inbred C57BL	1.60
Proteins	2.66	RNA	1.60
Alternative Splicing	2.40	Evolution, Molecular	1.60
Molecular Sequence Data	2.40	Organ Specificity	1.60
RNA, Untranslated	2.40	Variation (Genetics)	1.60
Proteome	2.40	DNA	1.60

Table B.10: Top 50 MeSH term scores for hubs of community 5.

Term	Score	Term	Score
Molecular Sequence Data	11.07	Saccharomyces cerevisiae	1.90
Amino Acid Sequence	10.95	Viral Proteins	1.79
Evolution, Molecular	10.71	Mice	1.79
Animals	9.52	Models, Molecular	1.67
Humans	8.57	Multigene Family	1.67
Sequence Homology, Amino Acid	6.90	Fungal Proteins	1.67
Phylogeny	6.79	Introns	1.55
Sequence Alignment	6.31	Genes, Archaeal	1.55
Conserved Sequence	5.83	Sequence Homology, Nucleic Acid	1.55
Bacterial Proteins	4.64	Binding Sites	1.55
Proteins	4.17	Software	1.55
Eukaryotic Cells	3.93	Signal Transduction	1.55
Genome, Bacterial	3.57	Genomics	1.43
Bacteria	3.45	Gene Duplication	1.43
Computational Biology	3.21	Base Sequence	1.43
Protein Structure, Tertiary	3.10	Species Specificity	1.31
Genome	3.10	Algorithms	1.31
Databases, Factual	3.10	Escherichia coli	1.31
Models, Genetic	2.62	Protein Folding	1.31
Archaeal Proteins	2.50	Caenorhabditis elegans	1.31
Archaea	2.38	Transcription Factors	1.19
Genes, Bacterial	2.38	Variation (Genetics)	1.19
Genome, Archaeal	2.14	Adenosine Triphosphatases	1.19
Gene Transfer, Horizontal	2.14	Genome, Human	1.07
Evolution	2.02	Saccharomyces cerevisiae Proteins	1.07

Table B.11: Top 50 MeSH term scores for authorities of community 5.

References

[1] Lada A. Adamic and Eytan Adar. Friends and neighbors on the Web. *Social Networks*, 25(3):211–230, 2003. 40

[2] Rakesh Agrawal, Sridhar Rajagopalan, Ramakrishnan Srikant, and Yirong Xu. Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th International World Wide Web Conference, WWW2003*, pages 529–535, 2003. 3

[3] Yuan An, Jeannette Janssen, and Evangelos E. Milios. Characterizing and mining the citation graph of the computer science literature. *Knowledge and Information Systems*, 6(6):664–678, 2004. 20

[4] Brett W. Bader and Tamara G. Kolda. Algorithm 862: MATLAB Tensor Classes for Fast Algorithm Prototyping. *ACM Transactions on Mathematical Software*, 32(4), 2006. 30

[5] Brett W. Bader and Tamara G. Kolda. Efficient MATLAB computations with sparse and factored tensor. Technical Report SAND2006-7592, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, 2006. 30

[6] Brett W. Bader and Tamara G. Kolda. MATLAB Tensor Toolbox Version 2.2. http://csmr.ca.sandia.gov/ tgkolda/TensorToolbox/, 2007. 30

[7] Dana B. Barr, Doug Landsittel, Marcia Nishioka, Thomas Kent, Brian Curwin, James Raymer, Kirby C. Donnelly, Linda McCauley, and P. Barry Ryan. Statistical issues: Barr et al. respond. *Environ Health Perspect*, 114(12):A689–A690, 2006. 13

[8] Tim Berners-Lee and Mark Fischetti. *Weaving the Web : the original design and ultimate destiny of the World Wide Web by its inventor*. HarperSanFrancisco, San Francisco, 1999. 2, 3

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. 36

[10] Sergey Brin and Larry Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:107–117, 1998. 23

[11] Deng Cai, Zheng Shao, Xiaofei He, Xifeng Yan, and Jiawei Han. Community mining from multi-relational networks. In *Proceedings of the 2005 European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'05)*, 2005. 24

[12] Deng Cai, Zheng Shao, Xiaofei He, Xifeng Yan, and Jiawei Han. Mining hidden community in heterogeneous social networks. In *Workshop on Link Discovery: Issues, Approaches and Applications (LinkKDD-2005)*, 2005. 24

[13] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the 7th World Wide Web conference*, 1998. 22

[14] S. Chakrabarti, B.E. Dom, S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and Jon Kleinberg. Mining the Web's link structure. *Computer*, 32(8):60–67, 1999. 1

[15] Alvin Chin and Mark Chignell. A social hypertext model for finding community in blogs. In *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia*, pages 11–22, 2006. 4, 13

[16] Peter Christen, Tim Churches, and Markus Hegland. Febrl – A parallel open source data linkage system. In *Advances in Knowledge Discovery and Data Mining*, volume 3056 of *Lecture Notes in Computer Science*. Springer, 2004. 33

[17] Heting Chu. Taxonomy of inlinked Web entities: What does it imply for webometric research? *Library and Information Science Research*, 27(1):8–27, 2005. 16

[18] Daryl E. Chubin and Soumyo D. Moitra. Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science*, 5(4):423–441, 1975. 10

[19] Susan E. Cozzens. What do citations count? The rhetoric-1st model. *Scientometrics*, 15(5-6):437–447, 1989. 1

[20] Diana Crane. *Invisible Colleges*. The University of Chicago Press, Chicago, 1972. 5

[21] Aron Culotta, Ron Bekkerman, and Andrew McCallum. Extracting social networks and contact information from email and the web. In *Proceedings of the First Conference on Email and Anti-spam (CEAS 2004)*, 2004. 3

[22] Scott Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990. 36

[23] Patrick Doreian and Andrej Mrvar. A partitioning approach to structural balance. *Social Networks*, 18(2):149–168, 1996. 25

[24] Elizabeth Duncan, F. Anderson, and Ray McAleese. Qualified citation indexing: Its relevance to educational technology. In *Proceedings of the 1st Symposium on Information Retrieval in Educational Technology*, pages 70–79, 1981. 10

[25] P. Elias, A. Feinstein, and C. Shannon. A note on the maximum flow through a network. *IEEE Transactions on Information Theory*, 2(4):117–119, 1956. 20

[26] David T. Felson and Leonard Glantz. A surplus of positive trials: weighing biases and reconsidering equipoise. *Arthritis Research & Therapy*, 6(3):117–119, 2004. 14

[27] B. Finney. The reference characteristics of scientific texts. Master's thesis, The City University of London, 1979. 10

[28] Gary W. Flake, Steve R. Lawrence, C. Lee Giles, and Frans M. Coetzee. Self-organization of the web and identification of communities. *IEEE Computer*, 35(3):66–71, 2002. 20, 24

[29] Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41, 1977. 20

[30] Carolyn O. Frost. The use of citations in literary research: A preliminary classification of citation functions. *Library Quarterly*, 49:399–414, 1979. 10

[31] Eugene Garfield. Can citation indexing be automated? In M. E. Stevens, V. E. Giuliano, and L. B. Heilprin, editors, *Statistical association methods for mechanized documentation*, pages 189–192. National Bureau of Standards, Washington, DC, 1965. 10

[32] Eugene Garfield. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities.* John Wiley and Sons, New York, 1979. 1

[33] Mark Garzone and Robert E. Mercer. Towards an automated citation classifier. In *Advances in Artificial Intelligence: 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2000*, pages 337–346, 2000. 10, 11

[34] Mark A. Garzone. Automated classification of citations using linguistic semantic grammars. Master's thesis, The University of Western Ontario, 1997. 5, 10, 11

[35] Lise Getoor and Benjamin Taskar, editors. *Introduction to Statistical Relational Learning.* MIT Press, 2007. 41

[36] David Gibson, Jon Kleinberg, and Prabhakar Raghavan. Inferring web communities from link topology. In *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, 1998. 4, 5, 22

[37] Michelle Girvan and Mark E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826, 2002. 5, 20

[38] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th International World Wide Web Conference, WWW2004*, pages 491–501, 2004. 3

[39] Mark S. Handcock, A. E. Raftery, and Jeremy M. Tantrum. Model based clustering for social networks. Technical Report Working Paper no. 46, University of Washington, 2005. 18

[40] T. L. Hodges. *Citation indexing: Its potential for bibliographic control.* PhD thesis, University of California, Berkeley, 1978. 10

[41] Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002. 18

[42] Christian Jacquemin. *Spotting and Discovering Terms through NLP.* MIT Press, Cambridge MA., 2001. 36

[43] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: a review. *ACM Computing Surveys (CSUR)*, 31(3):264–323, 1999. 19

[44] M. A. Jaro. Probabilistic linkage of large public health data file. *Statistics in Medicine*, 14:491–498, 1995. 33

[45] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25, 1963. 6, 19

[46] Hak Joon Kim. Motivations for hyperlinking in scholarly electronic articles: A qualitative study. *Journal of the American Society for Information Science*, 51(10):887–899, 2000. 16

[47] Richard Klavans and Kevin W. Boyack. Identifying a better measurement of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, 57(2):251–263, 2005. 19

[48] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999. 5, 21, 28

[49] Tamara Kolda and Brett Bader. The TOPHITS model for higher-order Web link analysis. In *Workshop on Link Analysis, Counterterrorism and Security*, 2006. 7, 26, 29

[50] Michael Krauthammer and Goran Nenadic. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6), 2004. 36

[51] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007. 40

[52] B. A. Lipetz. Problems of citation analysis: A critical review. *American Documentation*, 16:381–390, 1965. 10

[53] M. Magee. How research biochemists use information: An analysis of use of information from cited references. Master's thesis, University of Chicago, 1966. 10

[54] Katherine W. McCain and Kathleen Turner. Citation context analysis and aging patterns of journal articles in molecular genetics. *Scientometrics*, 17(1–2):127–163, 1989. 10

[55] Robert E. Mercer, Chrysanne DiMarco, and Frederick W. Kroon. The frequency of hedging cues in citation contexts in scientific writing. *Advances in Artificial Intelligence*, 3060:75–88, 2004. 23

[56] Michael J. Moravcsik. Measures of scientific growth. *Research Policy*, 2:266–275, 1973. 11

[57] Michael J. Moravcsik and Poovanalingam Murugesan. Some results on the function and quality of citations. *Social Studies of Science*, 5:86–92, 1975. 10, 25

[58] Hidetsugu Nanba, Noriko Kando, and Manabu Okumura. Classification of research papers using citation links and citation types: Towards automatic review article generation. In *Proceedings of the 11th SIG Classification Research Workshop, Classification for User Support and Learning*, pages 117–134, 2000. 11, 23

[59] Mark E. J. Newman. Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E*, 64(1):016131, 2001. 3

[60] Mark E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):404–409, 2001. 4

[61] Mark E. J. Newman. Analysis of weighted networks. *Physical Review E*, 2004. 24

[62] Mark E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8577–8582, 2006. 18

[63] Mark E. J. Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004. 20

[64] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, 2001. 18

[65] Charles Oppenheim and Susan P. Renn. Highly cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science*, 29(5):227–231, 1978. 10

[66] Larry Page, Sergey Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998. 1, 23

[67] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, pages 79–86, 2002. 39

[68] Bluma C. Peritz. A classification of citation roles for the social sciences and related fields. *Scientometrics*, 5:303–312, 1983. 10

[69] Alexandrin Popescul and Lyle H . Ungar. Statistical relational learning for link prediction. In *Proceedings of the Workshop on Learning Statistical Models from Relational Data at IJCAI-2003*, pages 81–90, New York, 2003. ACM Press. 41

[70] Jane Qiu. Scientific publishing: Identity crisis. *Nature*, 451(7180):766–767, 2008. 34

[71] Radoslav Radoulov. Exploring automatic citation classification. Master's thesis, University of Waterloo, 2008. 10, 11, 12, 30

[72] Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269, 1973. 6, 19

[73] Ira Spiegel-Rösing. Science studies: Bibliometric and content analysis. *Social Studies of Science*, 7:97–113, 1977. 10

[74] Ben Taskar, Eran Segal, and Daphne Koller. Probabilistic classification and clustering in relational data. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01)*, 2001. 18

[75] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 103–110, 2006. 11

[76] M. Thelwall. What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research*, 8(3), 2003. 16

[77] Joshua R. Tyler, Dennis M. Wilkinson, and Bernardo A. Huberman. E-mail as spectroscopy: automated discovery of community structure within organizations. *The Information Society*, pages 143–153, 2005. 4

[78] Yitong Wang and Masaru Kitsuregawa. Link based clustering of web search results. In *Proceedings of the Second International Conference on Advances in Web-Age Information Management (WAIM 2001)*, volume 2118, pages 225–236. Springer, 2001. 18

[79] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998. 36

[80] Wikipedia. Ant — wikipedia, the free encyclopedia, 2008. [Online; accessed 1-September-2008]. 27

[81] Wikipedia. Bee — wikipedia, the free encyclopedia, 2008. [Online; accessed 1-September-2008]. 27

[82] Wikipedia. Insect — wikipedia, the free encyclopedia, 2008. [Online; accessed 1-September-2008]. 27

[83] Wikipedia. List of korean family names — wikipedia, the free encyclopedia, 2008. [Online; accessed 3-September-2008]. 34

[84] Wikipedia. Wasp — wikipedia, the free encyclopedia, 2008. [Online; accessed 1-September-2008]. 27

[85] William E. Winkler. Overview of record linkage and current research directions. Research Report Series Statistics #2006-2, Statistical Research Division, U.S. Census Bureau, 2006. 33