COMBINING SOCIAL NETWORK AND SEMANTIC CONTENT ANALYSIS TO IMPROVE KNOWLEDGE TRANSLATION IN ONLINE COMMUNITIES OF PRACTICE

by

Samuel Alan Stewart

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
December 2013

Table of Contents

List of Tables

List of Figures

Abstract

Establishing online communities of practice is an important part of the knowledge translation process in the modern healthcare system, but these online communities are new entity that is inherently different from traditional communities of practice that are dependent on existing social structures. The objective of this thesis is to combine communication analysis and content analysis to delve deeper into the communications within an online community to try and determine how online communities exist, and how that information can be leveraged to improve online knowledge translation. Using a novel approach this project will map the contents of online conversations to a structured medical lexicon (MeSH), and then use the inherent relationships of that lexicon to calculate term, user and thread similarities within an online community. These similarities, combined with connection analysis results, will provide a much deeper understanding of how online communities function. The methods developed here will then be tested on two separate mailing lists, the Pediatric Pain Mailing List (PPML) and SURGINET, a mailing list of general surgeons.

List of Abbreviations and Symbols Used

AGNES	Agglomerative Nesting
BGM	Balanced Genealogy Model
BICGM	Balanced Information Content Genealogy Model
DAG	Directed Acyclic Graph
GVSM	Generalized Vector Space Model
IC	Information Content
ICD	International Classification of Diseases
IVF	In vitro fertilization
KT	Knowledge Translation
LCA	Lowest Common Ancestor
LINKS	Leveraging Internet for Knowledge Sharing
OBA	Open Biomedical Annotator
OGM	Optimistic Genealogy Measure
PARIHS	Promoting Action on Research Implementation in Health Services
PPML	Pediatric Pain Mailing List
RGM	Recursive Genealogy Measure

SE	Structurally Equivalent
SNA	Social Network Analysis
SNOMED	Systematized Nomenclature of Human Medicine
TF-IDF	Term Frequency-Inverse Document Frequency
UMLS	Unified Medical Language System
VSM	Vector Space Model

Acknowledgements

I would like to begin by acknowledging the gracious support of IDRC and the Teasdale-Corti Global Health Research Partnership Program, who have provided me with financial support over the last 5 years, without which I would not have been able to complete this thesis. I would like to thank my thesis supervisors, Dr. Raza Abidi and Dr. Allen Finley, who have provided me with immeasurable support and guidance over the course of my PhD, as well as the rest of my thesis committee: Dr. Samina Abidi and Dr. Jill Hayden. I have had the opportunity to pursue a number of different research collaborations throughout my degree at Dalhousie, with far too many collaborators to acknowledge individually, but I would like to specifically thank Dr. David Hamilton for providing me the venue to pursue these collaborations.

My family has been a vital component of my education not just during my PhD but for the last 25 years, and I would like to thank Ben, Nick, Kath and Dave for their continued support of my unending journey through academia. I would also like to thank my wife Maia, who has found the time to support and encourage my sometimes wandering focus while herself pursuing academic challenges far greater than mine, and without whom I would not have been able to deal with the challenges I have faced over the last 5 years.

I would like to finish by extending an additional, sincere thank you to Dr. Raza Abidi. Raza has gone far beyond the responsibilities of a thesis supervisor in his mentorship throughout my degree. He has supported me through all aspects of my PhD education, he has provided the opportunity for my travels to all ends of the earth in pursuit of academic success, and his overall guidance in all things academic and educational has transformed me into the person I am today. I would not be who I am today without him, and for that I am eternally grateful.

Chapter 1

Introduction

In an evidence-based medical world, there is an expectation that point-of-care decisions will be informed by established healthcare knowledge, yet research suggests that the body of healthcare knowledge is largely under-utilized. [2] This under-utilization of knowledge is leading to poorer healthcare and sub-optimal treatments. [19, 66, 82] 30 to 40% of patients are not receiving treatment supported by evidence-based medicine, and up to 25% receive unnecessary or potentially harmful care. [31, 73] To ensure that new knowledge is being used at the point of care, knowledge translation (KT) strategies must be implemented both by clinicians and by healthcare organizations as a whole.

The Canadian Institutes of Health Research (CIHR) define KT as "a dynamic and iterative process that includes synthesis, dissemination, exchange and ethically-sound application of knowledge to improve the health of Canadians, provide more effective health services and products and strengthen the health care system." [80] Clinicians utilize KT in their daily practice when they read clinical practice guidelines, participate in journal clubs, or consult peers about the best mode of treatment for a specific patient. Unfortunately, there are barriers to KT that are beyond the control of the individual. [30] To address this issue of moving knowledge into practice formal KT frameworks must be implemented at an institutional level. The focus of these frameworks is to incorporate evidence-based knowledge into daily practices, and have been shown to be effective in a variety of environments.

One of the challenges these KT frameworks face is how to work with knowledge that is not evidence based. Knowledge exists in a number of modalities: Explicit knowledge is the evidence-based knowledge that exists in the literature, experiential knowledge is the knowledge gleaned from years of working within the community, and tacit knowledge is the innate knowledge of clinicians, the intuitive knowledge of how things work. Most KT frameworks focus on moving explicit knowledge into practice, but do not spend sufficient time or resources addressing the issue of sharing experiential or tacit knowledge. Though this knowledge is not formally codified or validated it is a vital component of the medical

community. Incorporating experiential KT into existing KT frameworks to supplement the traditional KT processes would improve the KT process overall and significantly improve patient care.

Incorporating the theories behind Communities of Practice [93] into KT frameworks can provide the tools for facilitating experiential or tacit KT. A community of practice is an environment in which people gather around a common subject to share ideas and experiences with the goal of improving daily practice. It can compliment explicit KT strategies by providing informal communication avenues, allowing clinicians to share their experiences and seek advice about specific clinical situations.

Communities of practice, and KT in general, are unfortunately hampered by the temporal and geographical barriers that prevent face-to-face communication, and these problems are exacerbated in multidisciplinary medical subjects and in remote areas, where clinical experts are both rare and dispersed throughout the community. Web 2.0 tools can provide the means to share information when traditional conversations are not feasible. The LINKS model [1] provides a framework for facilitating KT using web 2.0 tools. Incorporating web 2.0 tools into the KT process is vital to ensuring quality care.

Using web 2.0 tools to establish communities of practice within medical environments has the potential to improve the KT practices of the healthcare community in ways that traditional KT methods cannot. Web 2.0 tools can create larger and more focused communities by connecting disparate groups from far reaching locations. Asynchronous communication systems, such as email and discussion forums, allow community members to communicate without having to coordinate their schedules, eliminating the temporal challenges that often hinder KT. Web-based KT can be more efficient: face to face conversations may not provide a lasting imprint on either party, particularly with respect to specific clinical problems. Being able to recall and review conversations about specific cases while facing that clinical problem allows the clinician to extract knowledge objects from the conversation and use them in their daily practice.

Web-based KT practices provide a record of the knowledge being shared and translated, with insights into the patterns of sharing across the community. These archives create a 2-mode structure between people and knowledge, allowing for a three-part analytic framework: connecting people to people, connecting knowledge to knowledge, and connecting people to knowledge.

1.1 Facilitating Web 2.0 Communities

Establishing virtual communities of practice is an essential part of facilitating KT within a modern medical community. A community of practice is defined in part by the connectivity between its members, therefore establishing connections between members is paramount to ensuring a strong community, and subsequently a sound KT process. As the size of the community grows it becomes increasingly challenging to find members that share the same KT objectives with you, therefore mechanisms need to be in place for finding members of the community that share your interests.

Tools exist for leveraging network structure and network connections for finding similarities between users. Most of these methods, however, are based in the Social Network Analysis (SNA) or graph theory literature, and focus on the structure of the network of connections between users, rather than the content of the messages themselves. The communications within a virtual community of practice contain important experiential and tacit knowledge, and extracting and representing this knowledge can supplement the traditional SNA methods by incorporating knowledge-based relationships between users into the traditional connectivity measures.

1.2 Research Approach and Methods

The objective of this thesis is to improve the online KT process through a combination of social network and content analysis. Through the LINKS model [1] we have identified several key areas that need to be addressed when developing an online KT community. The questions, methods and expected outcomes are presented in table 1.1 (an explanation of the LINKS model is presented in section 2.3.4 in the background chapter).

The two key components of the LINKS model that will be addressed by this thesis are establishing a culture of collaboration for KT and directing knowledge with respect to the users' context.

A culture of collaboration is essential to building a successful KT community. This thesis will investigate methods to establish a culture of collaboration through four different analytic approaches: (a) Isolate detection will identify what knowledge seeking activity within a community goes unanswered, and determining what could be the cause of the lack of engagement. (b) Response analysis will look at the thread reply patterns to determine the

Objective	Method	Expected Outcome
Establishing a Culture of Collaboration	Isolate Detection	Identifying knowledge seeking behaviour that goes unanswered, why it happens, how to fix it
	Response Analysis	Determining the participation behaviours of the community members, identifying knowledge seekers
	Centrality	Identifying the most active users, the community leaders
	Connection Clustering	Identifying user subgroups, finding similar users
Directing Knowledge Content	Knowledge Maps	Mapping the content of the conversations to common subjects
	Thread Clustering	Identifying content subgroups, determining popular subjects, linking similar threads
	Content Clustering	Identifying similar users; finding potential subgroups

Table 1.1: A summary of the research objectives of the thesis

activity levels of both the community as a whole and individual users, to get a broad sense of how the community performs its KT activities. (c) Centrality represents a set of SNA metrics designed to identify the most central members of a community, the leaders that are at the centre of the KT activities. Identifying these key users is important to the development and maintenance of a KT community, as they can greatly affect all other members. (d) Connection clustering looks at the leaders at a macro level, attempting to identify a core group of users within the community (through core-periphery analysis) or identifying potential sub groups of users based on their communication patterns (through 1-mode and 2-mode clustering).

Understanding the knowledge content being shared within the community is the second objective of this research, as this is key to gaining objective insight into the knowledge of the community members. Our approach is to exploit semantic mapping of the content of the messages to medical lexicons. Exploring the relationships between these mapped terms can provide greater insight into the content of the conversations, and the community as a whole. Content-based clustering of the users and threads can tell us what the most popular subjects are within the community, where potential subgroups may arise, and provides a second user clustering based on content instead of connections. The content mappings will also provide

a method for detecting similarity between users, which can be used to increase connectivity by connecting users to other like-minded individuals.

Finally, the knowledge content and the culture of collaboration methods can be combined, which provides the third objective of this thesis. We will investigate how content analysis and collaboration analysis can be combined to further our understanding of the KT process. Semantic explorations of the isolates can help explain what types of questions are not being answered within the community. Comparing the connection-based and content-based clusters can provide additional insight into potential subgroups within the community, and the content mappings can inform what may be causing the connection clusters. The Balanced Information Content Genealogy Model will be presented in section 3.3.8.2 as a novel method for determining similarity between users, and applying SNA methods to this model can provide further insight into content expertise within the community. Figure 1.1 presents the analytic methods for the thesis and how the combination of content and connection analysis can be combined to further our understanding of the KT process.

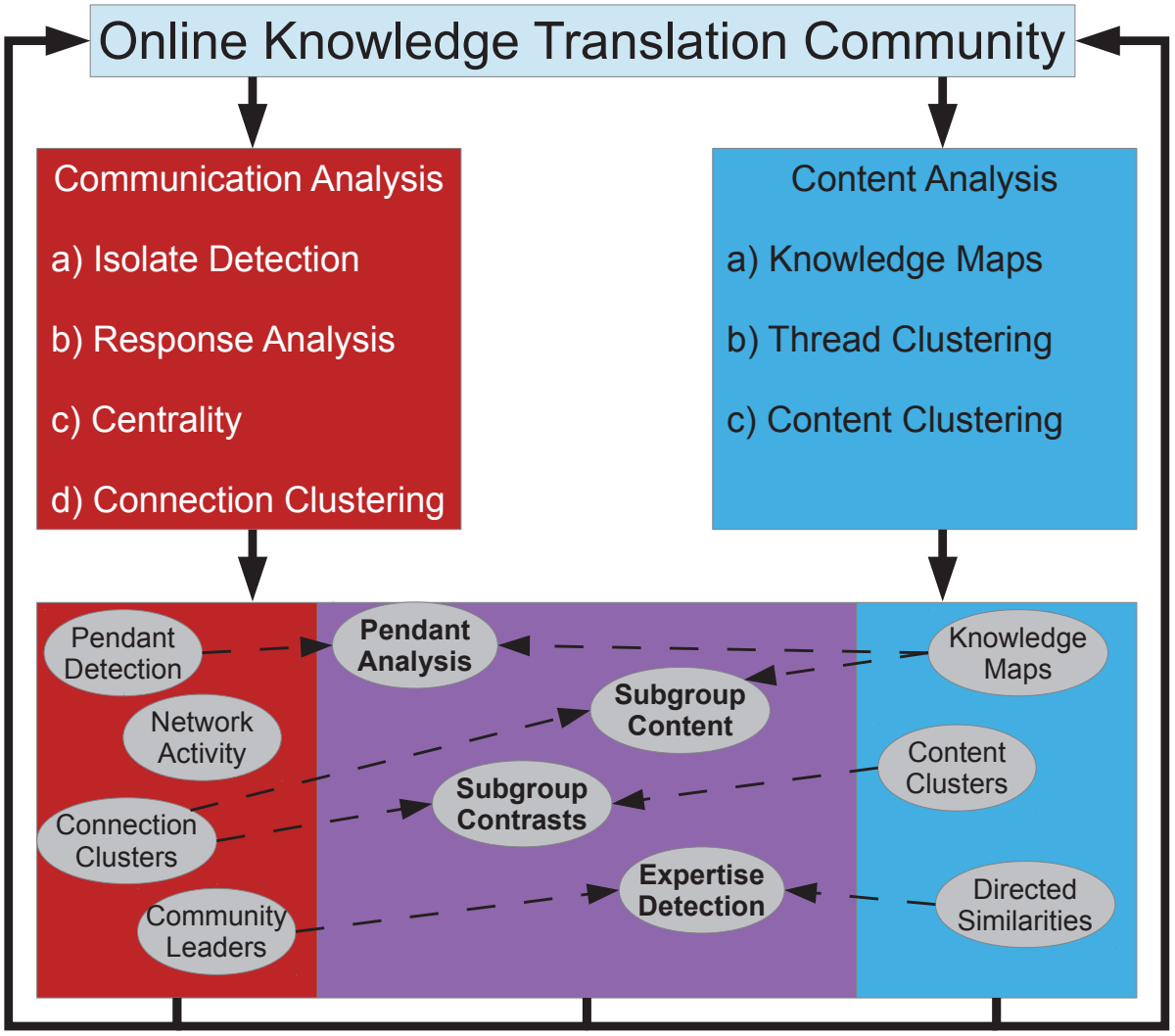


Figure 1.1: A broad overview of the connection based and content based analytic methods

1.3 Analyzing Knowledge Translation

The methods developed within this thesis will be applied to two different medical mailing lists. The Pediatric Pain Mailing List (PPML) is a community of 938 clinicians from around the world who meet online to discuss issues pertaining to pediatric pain. I have extracted the archives of the PPML from 2009-02-02 to 2013-02-03, during which time 2505 messages were shared on 783 threads by 460 community members. The SURGINET dataset is a community of 865 clinicians from around the world that use the mailing list to discuss general surgical issues. The community is much more active, sharing over 17,000 messages on 2,111 threads by 231 users during the period 2012-01-01 to 2013-04-05.

The knowledge maps will provide a summary of the content that is discussed within each community. For the PPML an investigation of the issues surrounding non-response will reveal a potential algorithm for preventing future messages from being ignored by the rest of the community. SNA metrics will help identify the most active users in both communities, and will identify the core of each community, those users that are at the centre of the KT activities. The content mappings for both lists will provide knowledge based representations of both users and threads, which will be used to find similar users, to find clusters of threads based on message content, and to attempt to define content expertise.

Chapter 2 will present the background and literature review for the thesis, including a short summary of KT frameworks (including the LINKS model) and communities of practice, medicine 2.0 research and semantic mapping tools. Chapter 3 will present the analytic methods for the thesis in three sections: Social Network Analysis methods, content analysis methods and integration of the network and content analytic methods. Chapter 4 will apply the analytic methods to the two test mailing lists, and chapter 5 will outline the conclusions of the thesis as well as future work.

Chapter 2

Background

This chapter will investigate the principles surrounding the use of online tools for KT within the medical community. It will investigate two KT frameworks (Diffusion of Innovation [71] and the PARIHS model [59, 72]), and will explore the LINKS model [1] and Communities of Practice [94] as tools for performing KT online. It will then investigate the semantic mapping literature, and in particular the Metamap [5] and Mgrep [74] systems for mapping unstructured medical text to formal medical lexicon.

2.1 Web 2.0

Web 2.0 was defined in 2004 as "a set of economic, social, and technology trends that collectively form the basis for the next generation of the internet, a more mature, distinctive medium characterized by user participation, openness and network effects." [86] Other definitions focus specifically on the improved communication web 2.0 can provide via social networking. [33] In its most simplest terms the move from web 1.0 to web 2.0 is an increase in interaction: Web 2.0 is the interactive web, in which content is created and modified by users rather than by the website administrators themselves.

Web 2.0 can be encapsulated by the simplest of interactive web tools: email, mailing lists, news groups, chatrooms, online discussion forums and online bulletin boards have been around since the early stages of the internet, yet their ability to facilitate communication and share information between users puts them into the realm of web 2.0. More modern and technically advanced web 2.0 tools include instant messaging tools, social networking sites, and blogs. These tools provide real-time interactivity, and provide the means for users to present their own views and opinions to a wide audience, sharing information via social media.

Critics of web 2.0 argue that, since the interactive web is not built on any new technologies, it is not a revolution but merely a natural evolution of the internet itself. [86] Other critics argue that the older examples of web 2.0 tools do not truly fall within the realm of

web 2.0 because they fail to function as modern forms of social media.

This thesis considers web 2.0 in its broadest sense, as any internet-based tool that can be used to facilitate communication between users. Without excluding social media as an integral part of web 2.0, the focus is on the use of the internet as a tool to improve communication between users.

The idea of using the principles behind web 2.0 to facilitate social interaction and KT is a well studied area, dating back before the invention of the term itself. Wellman and colleagues [92] explored the burgeoning world of internet communities, and how the principles of social interaction online could be used in personal and workplace interactions. He explored primitive tools such as email, list servers and usenet groups, but established their potential for improving communications by bridging physical boundaries.

2.2 Medicine 2.0

The modern internet is flooded with discussion forums and online communities around health. Some of the largest examples include Patientslikeme (www.patientslikeme.com), Hello Health (hellohealth.com), IVF clinic [84] and Parkinson Net (www.parkinsonnet.nl). A review of the medicine 2.0 literature in 2004 found over 24000 health-related discussion groups within Yahoo! groups alone. [28]

Van de Belt and colleagues [86] performed a systematic review of the medical literature in 2010 to try and find formal definitions for the terms medicine 2.0 and health 2.0. They found 46 unique definitions of the terms, the majority of them focusing on health 2.0 over medicine 2.0. Their results were ultimately not conclusive: Most definitions focused on the relationships and communication patterns between patients and healthcare professionals, but some did not include the use of web 2.0 (or the use of the internet at all) in their definition.

Like web 2.0, medicine 2.0 is ultimately a very loosely defined term that may represent a revolution in the field or may just be the natural evolution of healthcare in an increasingly connected world. This thesis will consider medicine 2.0 in much the same way it considers web 2.0, as a tool for facilitating communication between clinicians, patients and all other healthcare stakeholders.

In a formal evaluation of the effects of medicine 2.0 interventions, Eysenbach and colleagues [28] attempted to review the efficacy of discussion forums as a medical intervention, but found a dearth of quality papers evaluating discussion forums. They found 45 papers

representing 38 studies, of which only 6 were pure internet-based interventions, the rest including a discussion forum as part of a larger study. One of the conclusions from the authors was that there is no robust evidence on the health benefits of virtual communities. As the number and size of virtual health communities increases, it is vital to understand the implications of these communities, therefore research into their effects must be done.

One important finding of the Eysenbach review was the suggestion that virtual communities succeed when there is an "intrinsic desire" to communicate with each other and share health knowledge and experiences, and that it was very difficult to try and create a community. [28] This finding is confirmed in more modern experiences of using discussion forums to facilitate education and KT. Students in an anatomy class that had 8% of their grade linked to their participation in a discussion forum actively engaged through the forum, and 83% of the students found the boards useful, improving their team building and critical analysis skills. [14] This finding was replicated by Kuhn et al [47] who found that a moderated pretest discussion forum as a tool for facilitating communication between nursing students significantly improved students' grades. Valaitis et al [85] designed a discussion forum to facilitate the establishment of a virtual community of practice for community health nurses. For a disparate community with a dearth of quality information [85] a discussion forum provided a key KT tool for the participants, providing them with a way to connect to their peers. "The development of effective CoPs is dependent upon the ability of individuals in the community to critically interpret, respond and share information with colleagues." [85]

In contrast, when participation is neither required (via grades) nor requestion by the community (for KT), participation wanes. In a study comparing online journal clubs to face-to-face clubs, researchers found a huge gap in participation rates between the two, "... because of the low participation in the Internet journal club." [60] Though the authors stated that the journal clubs were required there was no punishment for not participating. With no explicit inducement to participate and no intrinsic desire from the residents the forum faltered.

Using the principles suggested by Wellman and colleagues [92] medicine 2.0 seems like a natural tool to be incorporated into the formal KT frameworks that are implemented within the healthcare system, as it can allow healthcare stakeholders to communicate new knowledge across boundaries that prevent traditional communication.

2.3 Knowledge Translation Frameworks

Within the world of healthcare KT there has been much research on KT frameworks. Estabrooks and colleagues [27] provide an overview of the options available to the medical community and try to provide guidance in choosing the best one for specific situations, but fail to identify an overarching strategy; the Diffusion of Innovation Theory [71] is the closest to achieving this status. This project has decided to implement the PARIHS framework as a tool for bringing research to practice, but other frameworks are capable of providing the necessary tools to facilitate KT.

2.3.1 Diffusion of Innovations

The Diffusion of Innovations, popularized by Everett Rogers [71], attempts to explain how and why innovations are adopted within a specific community. Rogers' original work was based largely on research in agriculture and medical practice, so even though it is designed for innovation diffusion rather than knowledge diffusion, it can still provide a basis KT.

Rogers' proposes four main elements that influence the spread of a new idea [71]:

- The innovation, an idea, practice or object that is perceived as new by the unit of adoption
- The communication channels through which the members of the community share the innovation
- Time, the rate by which members of the community adopt the innovation
- Social system, the set of interrelated units that are engaged to accomplish a common goal

New innovations are rarely evaluated from a scientific standpoint, but are instead evaluated subjectively within the community. For knowledge adoption in the medical community this is particularly important, as it demonstrates that the adoption of new practices is not always guided by best practices, but by the attitudes and beliefs of the community as a whole.

There are five stages to the adoption process [71]:

- Knowledge: The initial exposure to the innovation

- Persuasion: Interested individuals seek information about the innovation
- Decision: Individual decision units (people and/or groups) take the concept and weigh the advantages and disadvantages of using it, making a decision to accept or reject. This is the most difficult step to gather evidence on since it is a subjective, individual decision [71]
- Implementation: The innovation is employed, and its utility is evaluated
- Confirmation: The decision is finalized based on the evidence surrounding its use

The social system plays a key role in the diffusion, as opinion leaders and change agents exert their influence on the process. Individuals tend to choose to interact with people that are similar to them, i.e., they chose homophilic relationships. [71] Diffusion requires heterophily, as people from different backgrounds can bring new material to the community. The optimal situation for diffusion, therefore, is when two people are homophilous except for their knowledge of innovation.

The members of the community that are adopting a new innovation can be categorized into five groups: Innovators, early adopters, early majority, late majority, and laggards. The adoption rate follows an S-curve (i.e. a logistic curve or an ogive), with very few people in the innovators and laggards categories.

Though the diffusion of innovation theory is not directly applicable to all KT practices, it is influential in the way it informs other KT frameworks, such as the Research Development and Dissemination Utilization Framework [34] and the Ottawa Model of Research Use. [54] The PARIHS model is another example of a KT framework that leverages the principles of the diffusion of innovation.

2.3.2 PARIHS

The Promoting Action on Research Implementation in Health Services (PARIHS) framework is designed to guide knowledge uptake and instigate practice change in health systems [59,72]. It posits that successful implementation of research is dependent on the relationships between evidence, change context and the method of facilitation. [45]

Under the PARIHS framework evidence is composed of research, clinical experience and patient choice, i.e., it is not restricted to strictly evidence-based knowledge, but incorporates

the experiential and tacit knowledge of both clinicians and their patients. Explicit knowledge is easily extracted from published literature, including journal articles, textbooks, clinical practice guidelines, et cetera, but identifying and using experiential knowledge is more difficult. The extraction and leveraging of clinical experience is done through self-reflection on clinical practices, discussion with peers and critique.

The context of the KT environment is a key component of the process. The KT culture, the leadership, the evaluation methods and the receptivity of the community all keenly influence the KT process. The culture represents the status quo in terms of both practice and knowledge implementation strategies. [59] Leadership within the community is a vital part of the framework: strong leaders define clear roles, effective team work and a practical and clear organizational structure. [45] The evaluation dimension of the framework establishes a need for change by providing baseline measures of knowledge implementation practices, and then provides mechanisms for evaluating the effectiveness of the KT intervention.

The PARIHS framework operates under an active, multifaceted facilitation process. Education outreach, reminder systems, audits and feedback are all components of facilitation. Facilitators are unique members of the PARIHS team, distinct from both change agents and champions. They can either be internal members of the community or external, and their role is to enable KT rather than direct it. They do not necessarily need content knowledge, but they must be experts in working with others, managing conflict and enabling others to change. They need to have drive, enthusiasm, strong communication skills and credibility. [72]

2.3.3 Communities of Practice

Quality knowledge management and KT requires collaboration between different members of the healthcare team. A knowledge sharing team that includes physicians, nurses, pharmacists, researchers, patients and their families brings a larger and more heterogeneous body of knowledge together, allowing the members of the team to acquire new knowledge that is not available from their peers. Establishing a community of practice can provide this environment, allowing healthcare practitioners to meet and discuss healthcare issues. A community of practice is defined as a group of people who share a common practice, and who interact with each other to learn to do it better, [94] and is formally defined by three characteristics: the domain, the community and the practice. [93]

The domain is the subject area that defines the group, such as a medical specialty. Bringing people together around this common domain to learn from one another establishes a community of learning around the subject. The community does not need to necessarily work together on a daily basis, nor do they need to meet face-to-face; their common interest in learning from one another and their willingness to communicate around this topic is sufficient to establish the community. The sharing of common interests within the community is necessary, but not sufficient in establishing a community of practice. The members of the community must take the knowledge being shared in the group and use it to improve their respective practice. The idea behind a community of practice is to build a strong knowledge base by connecting different practitioners, who share their knowledge and leverage their new knowledge to improve their own practice.

This third aspect, of taking shared knowledge and implementing it in clinical care, requires a trust. Within the community there must be an inherent trust in the other members in order for them to be confident in using the advice of their colleagues in their own medical practice. There is a proliferation of information available to clinicians, but much of it is unvalidated, and cannot be directly trusted without validation. The validated knowledge is largely explicit knowledge, and it is often too time consuming to extract the salient information for a specific clinical example. A clinician that has a trusted peer group can quickly ask them about specific treatments, or what part of a clinical practice guideline is pertinent to their situation, and can then use that information to improve their care with confidence that the information they are receiving is correct and appropriate.

2.3.4 LINKS Model

The LINKS model provides a conceptual framework to help establish online communities of practice for specialized knowledge sharing using web 2.0 tools. [1] The LINKS model identifies the key determinants of an online knowledge sharing environment in order to systematically conceptualize and implement a purposeful health knowledge sharing environment for an online community of practice. The LINKS model characterizes healthcare knowledge sharing solutions at three interrelated levels: Conceptual, operational and compliance, as demonstrated in figure 2.1. The conceptual level stratifies knowledge sharing into three dimensions: the knowledge modality, the knowledge sharing context, and the knowledge sharing medium. The operational level addresses technical infrastructure issues pertaining to establishing a

culture of collaboration between the stakeholders. The compliance level addresses the underlying issue of perceived trust in the system. As the layout of figure 2.1 illustrates, the three levels are not hierarchical in nature, rather they are inter-related, and each level must be addressed in order to implement a successful knowledge sharing environment.

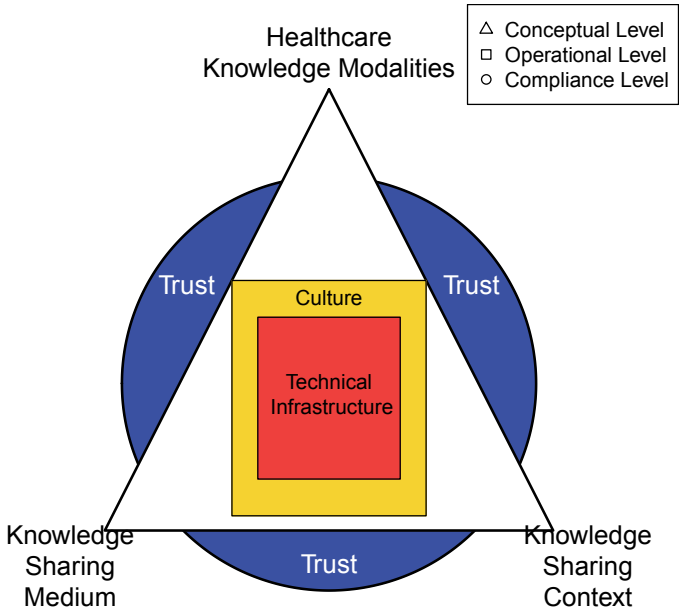


Figure 2.1: The LINKS model, adapted from [1]

Table 2.1 presents a summary of the LINKS model, illustrating the different levels and their constituent elements with respect to the reported research. For more detailed information about the LINKS model see its original methodological paper. [1]

Level	Element	Description
Conceptual Level	Knowledge Modality	The knowledge modality characterizes the type of knowledge being shared. Tacit, explicit, experiential and social knowledge are the key knowledge modalities that are typically shared through an online knowledge sharing environment
	Knowledge Sharing Context	The knowledge context aims to define the topics being discussed, the motivation for knowledge sharing, the temporal relevance of the knowledge sharing and the orientation of the discussion stakeholders
	Knowledge Sharing Medium	The medium determines the range of methods that can be employed to share knowledge. Implementation of each medium imposes a different set of operational considerations. Examples include face-to-face environments, virtual meeting tools, synchronous and asynchronous messaging systems, et cetera.
Operational Level	Technical Infrastructure	The technical infrastructure characterizes the technologies and strategies to be used to develop and deploy the knowledge sharing environment. It is imperative that the technical implementation of the project adequately addresses the conceptual level, in order to ensure maximum trust and engagement in the system.
	Culture of Collaboration	The culture of collaboration defines the ecosystem in which the online community of practice engages, collaborates and perpetuates knowledge sharing. Different community members have different levels of expertise, expectations and experiences, and the community must be designed to facilitate knowledge sharing between these different members.
Compliance Level	Trust	A community of practice engages and shares knowledge when there is a sufficient degree of trust in (a) the veracity of the knowledge being explicated and shared; and (b) the pedigree of the member who is explicating and sharing knowledge. The aim of the compliance level is to institute mechanisms to establish trust in the knowledge sharing exercise so that the community engages in a free flow of knowledge sharing. In addition, for individuals to freely share knowledge there is a need to instil trust in the eventual use of the knowledge, i.e., that the knowledge will be used for the right purposes and in the right manner.

Table 2.1: Summary of the LINKS model, adapted from [1]

2.4 SNA for Understanding Online Communities

Barry Wellman and colleagues were some of the first to propose SNA as a tool for exploring online communities as social networks, describing methods for better understanding how people communicate online. [91] These ideas of using social network analysis to understand online communications were further explored, [17, 35, 36] and the principles for understand online communities were established.

Aviv et al. [7] explored the use of Asynchronous Learning Networks for knowledge construction by studying the communication patterns within the community using social network analysis. They looked at basic social network analysis metrics along with clique analysis and clustering in an attempt to differentiate between structured and unstructured forum design and its effect on overall learning. Similar work has been pursued by several other research projects [3, 16, 21, 22, 26, 44, 50, 56, 57, 65, 67, 69, 75, 77, 83, 88, 97, 98] who used a variety of SNA methods to understand their communities. Beyond simple SNA metrics such as centrality measures (see section 3.2.2.3 for a detailed explanation) there were a variety of more advanced metrics that looked deeper into the structure of the community. Clique analysis was used as an attempt to define clusters of users [50, 88], while other projects approached clustering using regular or structural equivalence [67, 83], core-periphery analysis [16, 97], or k-means clustering [22]. Moving onto modelling techniques, ERGMs or p* models [75], other regression methods [3, 77, 88], or structural equation models [83] were used to try and gain insight into what drives a community.

One of the ideas that continually arises in the literature on analyzing online communications is the need to analyze the content of the messages. Many projects pursue this goal using content analysis. [7, 50, 56, 67, 88, 97, 98] Content Analysis typically involves manually processing the messages from the community to code them based on a formal text-coding schema, and then studying the results of the coding. Other projects leveraged natural language processing techniques [21, 22, 50], which analyzed the text automatically. Few projects incorporated the content analysis into their social network analysis, and those that did [22, 26, 50, 67, 88] did so only as a stratification factor.

This project is not interested in manual content analysis, as the time-consuming process of parsing each message manually makes it an impractical tool for large databases. Natural language processing techniques, including WordNet [61] and the approach described by Davoodi et al [22] are potential options, but they do not take full advantage of the existing

semantic mapping techniques that exist within the healthcare community.

2.5 Semantic Mapping and Knowledge Acquisition

In an evidence-based medical world, it is vital that knowledge be available to clinicians at the point of care. Unfortunately, the lack of organization, proper indexing, aging information sources and poor distribution have been shown to negatively affect a clinician's access to pertinent information. [19, 66, 82] The use of formal semantic languages is a key step in improving clinician access to medical knowledge, by providing a unified indexing of the existing medical knowledge.

Clinicians need to be able to leverage the semantic languages, however, in order to make full use of the formal indexing. Leroy and Chen [49] developed a system that processes general medical queries and returns a set of medical keywords from Unified Medical Language System (UMLS). Cimino et al [15] designed a system that maps clinician queries to a set of generic queries based on UMLS keywords. Developing tools that can provide semantic terms to accompany existing, unstructured text can provide a valuable resource, by providing clinicians with the formal semantic terms that pertain to the text they are currently processing. Emails, electronic medical records, news articles and clinical notes can all provide vital information if they are properly processed to extract the relevant information and provide it in a formatted, systemic manner.

The process of mapping free text to formal medical lexicons (and specifically to the UMLS) has been an objective of the medical research community for a long time. The value of having formal medical representation of ideas combined with the challenge of performing the task manually has made research into automated approaches very valuable. This problem is often linked to MEDLINE, which is manually indexed by Medical Subject Heading (MeSH) terms, and thus provides an objective reason to connect text to UMLS terms. MicroMeSH [55] was one of the first attempts to do this, by providing a simple system to expand search queries to MEDLINE and provide a tool where users could browse the MeSH tree around the terms they searched.

CHARTLINE [62] processed free text of medical records and connected them to relevant terms in the MeSH lexicon via a direct mapping. This process was improved by SAPHIRE [37], which explored the idea of processing free text and cleaning it by mapping terms to their synonyms. This was a valuable addition to the literature, as it normalized the process of

mapping *women* to *woman*. This process was taken up by Nadkarni et al [63] who used this synonym mapping along with a part of speech tagger to better identify the structure of the conversations and attempt to identify specific words and phrases in the text. PhraseX [76] also used this kind of synonym parser to analyze the mapping of MEDLINE abstracts to the UMLS metathesaurus, in order to evaluate the contents of UMLS itself. Other, similar approaches include KnowledgeMap [23] and IndexFinder. [100]

The current, gold standard in the industry is Metamap, though another product, called Mgrep [74] provides a very similar service. The creators of the Open Biomedical Annotator [41] designed a system that leverages the results of any semantic mapping service (Metamap or Mgrep) and the ontology relations within the lexicon to produce a more complete semantic mapping.

2.5.1 Metamap

Metamap uses a special natural language parser called SPECIALIST [5] to find all the nouns and noun-phrases in a discussion thread, and maps them to one or more UMLS terms. Each mapped UMLS term is assigned a score that is a measure of how strongly the actual term mapped to the UMLS vocabulary. The score is a weighted average of four metrics measuring the strength of the matching, with an overall range in [0,1000], with higher scores indicating a better match. The formal equation for calculating the scores is:

$$\frac{1000 \times (Centrality + Variation + 2 \times Coverage + 2 \times Cohesiveness)}{6} \tag{2.1}$$

- Centrality: An indicator of whether the matched (source) term is the head of the phrase

- Variation: A measure of the distance between the matched term and the root word. For example, if the source word is eye and the match is to the term ocular, the distance is 2, as ocular is a synonym for eye

- Coverage and Cohesiveness: Measures of how well the source term and the UMLS term match each other: if the source and UMLS terms are both “pain” then the match is perfect, but if the source term ocular matches to the UMLS term Ocular Vision then the coverage and cohesiveness are less than perfect.

Metamap's precision and recall in previous projects have varied depending on the format of the text being processed, from values as high as 0.897 and 0.930 respectively [42] to values as low as 0.56 and 0.72 [11]. The difference between the precision and recall values show that Metamap does a good job at returning pertinent MeSH terms, but also returns impertinent terms as well, i.e., its results are somewhat noisy. Projects that reported low recall and precision with Metamap acknowledged that many of the problems come from the inherently ambiguous nature of the text being processed: in processing medical residents' voice recordings, it was noted that Metamap failed to recognize abbreviations, acronyms or complex phrases that omitted key terms [12].

For our purposes, the Metamap scoring system provides a baseline measure of how well the mapped UMLS term represents the original term in the PPML discussion thread. Table 2.2 contains some sample mappings to the MeSH lexicon and their scores.

Despite the inconsistencies in the terms returned by Metamap, it provides a valuable tool for mapping unstructured messages and conversations to a structured medical lexicon. The Knowledge Linkage project [78] uses these mappings to try and provide explicit knowledge links to the experiential knowledge being shared within the community.

2.5.2 Open Biomedical Annotator and MGrep

The Open Biomedical Annotator [41] was developed to automate the process of providing keywords to datasets that are available on the web. Their process was to take the metadata from the datasets, pass them through a semantic mapping engine (either Metamap or Mgrep) and then post-process their output using ontological relationships. The amount of expansion allowed can be controlled by the user.

1. *is_a* transitive closure: Terms in the UMLS have a tree-like structure, so most terms have 1 or more parents. Expanding up the tree provides more general semantic terms than the specific term returned.

2. Semantic Distance: Sibling relationships also exist within the tree. Two semantic terms are somewhat similar if they have a shared parent, less similar if they share a grandparent (a parent of a parent), et cetera.

3. Ontology Mapping: Using the UMLS mapping tools the MeSH term returned can be augmented by the SNOMED or ICD-9 term as well. The terms from other lexicons

can provide new information.

The authors of the Open Biomedical Annotator performed an experiment to compare MetaMap to Mgrep [74] in terms of accuracy and speed. They found that Mgrep performed slightly better in terms of precision and was much faster (1/5th of a second compared to 8 minutes). The authors concluded that, because they were looking for real-time implementation, Mgrep was was a better option for them, and thus The Open Biomedical Annotator was implemented using Mgrep.

The details of how Mgrep works are not completely clear, and publications on it have been limited to conference posters [20]. The authors of the Open Biomedical Annotator claim that it "implements a novel radix-tree-based data structure that enables fast and efficient matching of text against a set of dictionary terms" [41]. The scoring algorithm as well is not completely explained, though it performs a similar expansion scoring to Metamap, where partial matches and derived matches receive lower scores that perfect matches. Mgrep is not distributed itself, but is accessed via the OBA: performing a mapping with the OBA without using the ontological expansions results in a strictly Mgrep-based mapping. Table 2.2 contains some sample mappings from Mgrep.

The report stated that when music therapy is used, the babies required less pain medication. Does anyone know of any published reports of empirical research demonstrating the effect?					
Metamap Terms			**Mgrep Terms**		
Source	**MeSH Term**	**Score**	**Source**	**MeSH Term**	**Score**
music therapy	Music Therapy	1000	Music	Music	10
			therapy	therapy	10
the babies	Infant	966			
less pain medication	Pain	660	Pain	Pain	10
less pain medication	Pharmaceutical Preparations	827			
of any published reports	Publishing	694	Report	Report	16
			Research	Research	10
of empirical research	Empirical Research	1000	Empirical Research	Empirical Research	10

Table 2.2: Sample message and its associated MeSH mappings from both Metamap and Mgrep

2.6 Conclusion

It is clear that both web 2.0 and medicine 2.0 are somewhat nebulous terms within the research community. I will take the term web 2.0 as a blanket term to describe those interactive online tools that facilitate communication. Medicine 2.0, therefore, refers to the online tools that are used to facilitate communication about health between clinicians. I am restricting the discussion to clinician-communities because I want to focus on communities of practice and KT frameworks, and though the online communities that bring patients together can be a valuable tool in the healthcare process, they do not fall within the same domain. It is important to draw the clear distinction between these two types of communities, between open and anonymous online communities and virtual communities of practice that are embedded within formal KT frameworks.

Social network analysis was proposed as a tool to analyze and understand online communities over 15 years ago. It provides the means to understand online relationships in a

way that traditional analysis cannot, as it can adapt to the dependent nature of the interpersonal relationships that challenges traditional statistical methods. Many projects have explored various SNA methods for analyzing online communities, but these methods have largely ignored the content of the messages when investigating the structure of the network.

Semantic mapping techniques can provide the means for "understanding" the content of the messages within the community. Though many tools have been developed to map free text to structured medical lexicons, Metamap and MGrep seem to be the two best utilities to date, so they should be investigated going forward. The methods outlined within the Open Biomedical Annotator [41] should also be explored. Once a tool is chosen, incorporating the mappings from the tool into the SNA methods will be key in improving our understanding of online communities.

Chapter 3

Methods

This chapter will outline the methods for measuring the culture of collaboration and directing the knowledge context of the community. The first section will present network analysis, including Social Network Analysis (SNA), methods, the second section will present content analysis methods, the third will evaluate the content analysis methods, and the fourth will show how the network analysis and content analysis methods can be used in combination to further our understanding of the community overall.

3.1 Definitions

Throughout the rest of this thesis certain terms will be used with respect to the analysis that may require explanation to understand their purpose within the analytic framework. The terms defined below are how they should be interpreted withing the context of this thesis.

Community Leaders are the users within the community that drive the KT practices. They are not explicitly defined based on roles that are defined by the community (such as mailing list administrators), they are implicitly identified as community leaders based on their active participation within the community.

Expertise is defined as being knowledgeable about a specific subject that other users are interested in. The knowledge within a message is extracted through semantic mapping techniques, so it is difficult to differentiate knowledge seeking from knowledge sharing behaviour, but the general thought process is this: User B is interested in a subject, and user A has spoken about that subject before, therefore user B is interested in user A. We assign the name expertise to this relationship, even though it may reflect the true nature of the relationship. The relationship may be asymmetric if user A is interested in a myriad of topics that user B is not

Correlation and Similarity are measures of how much two users/threads overlap, based on the semantic mappings of their messages. The assumption in this definition is that

a user's interests are represented by the content of their conversations. Similarity is measured on a $[0, \infty)$ scale and correlation is a scaled version on a $[0, 1]$ scale

Semantic Similarity is the similarity between two words based on their inherent meaning. Within the MeSH lexicon there is an inherent structure between words, and the semantic similarity calculations capture this relationship. MeSH only has one type of relationship, an hierarchical *is_a* relationship, so that is the semantic similarity calculated in this thesis.

Contextual Similarity is the similarity between words that is not capture in their meaning, but in their application. Within a specific context words may be inherently related in ways that are not captured in their generic interpretations. When words occur together consistently it is assumed that there is a strong relationship between them, and this relationship is captured using contextual similarity.

3.2 Establishing Culture of Collaboration

The nature of online KT is markedly different from face-to-face KT, with larger and more focused communities, anonymous or quasi-anonymous communication (depending on the medium), asynchronous communication and the existence of communication archives. Given the importance of KT and its potential for improving clinical care it is of interest to get insights about the knowledge sharing dynamics of the virtual community, as it can provide detailed insight into the collaboration practices of the community members. This section will investigate methods to establish a culture of collaboration within an online community. The first step will be isolate detection, followed by response analysis. SNA will provide centrality measures designed to identify the leaders that are at the centre of the KT activities, and connection clustering will look at these leaders at a macro level, attempting to identify a core group of users within the community and potential sub groups of users based on their communication patterns.

3.2.1 Building a Network From a Mailing List

An online KT community is comprised of users, and the messages they share form threads. To properly understand the collaboration structure within the community it is important to

move beyond the individual users and study the community as a whole. SNA utilizes the principles of graph theory to represent communication networks in terms of actors (nodes) and ties between actors (edges) [32, 90]. Traditional statistical analysis focuses on actors as independent units, and analyzes them in terms of their personal attributes. SNA instead focuses on the structures that emerge out of the relations between actors, and not on the actors themselves.

The structure of the network is a key component of the analysis process, and there is no accepted standard in terms of how to design the network to properly represent the conversations within online discussion forums or mailing lists. Within social network analysis the network is represented as an *adjacency matrix*, in which the entry at row i column j represents the communication from node i to node j. There are three general attributes of the adjacency matrix that define how to represent the network.

Directed vs Undirected: One would expect that communication is directed. When user i sends and email to user j, there should be a directed tie from i to j. The challenge in using this representation, however, is determining the target of the communication when all communication is public. When someone posts the 8^{th} reply in a conversation on a mailing list, who should their post be directed to? The person directly before them? The first person? All 7 people that posted before? The idea of an undirected communication network is that a tie from user i to j indicates that they have shared a conversation before. In the analysis of mailing list data a thread is thought of as a community, and messages are announcements to all members of the community. In this sense the ties in the adjacency matrix represent shared interest rather than directed communication. An undirected matrix is an upper triangular matrix (i.e. a symmetric matrix), as a tie from i to j is the same as a tie from j to i.

Binary vs Valued: The values in a network can either be binary, in which case they can represent a yes/no value, or they can be valued. In a communication network a value would represent the number of messages sent between i and j, or the number of threads shared between the two users. In reality the decision between binary and valued is dependent on the structure of the data and the values in the network. Valued data often does not provide more information than binary, and binary data has more analytic options within the SNA literature. This project will largely employ valued networks, which will be dichotomized if necessary in certain situations.

1-Mode vs. 2-Mode Networks: In the majority of network analysis nodes represent a class of people and ties represent some sort of social construct that connects them: friendship, advice, work, et cetera. This is referred to as a *one-mode network*, as the connections are between a single class of nodes. *Two-mode networks*, in contrast, represent two different classes of nodes, and ties exist only between classes. These most commonly occur when one set of nodes represents people, and the second set represents events, and ties go strictly from one mode to another (indicating that a person has attended an event). A two-mode network that represents people and the events they attend is sometimes called an *affiliation* network. With online conversations a two-mode network can be constructed, in which the users are one mode and the threads they are communicating on are another mode, and ties exist between users and threads if they have posted to and/or read the thread. A one-mode transformation of the two-mode network can be made, in which a tie between users indicates that they have communicated on a thread together.

In order to analyze a network we need to process the communication data and create a network structure. We will first consider the network data as a series of *threads*, or conversations around a specific subject. A thread normally begins with a user making a post about a specific subject, or asking a specific question, which elicits comments from other members of the community. We can form a two-mode network between the comments and the users, where a tie between the user and the thread indicates that the user has commented on that thread. Figure 3.1 presents a sample thread and the network that evolves from it.

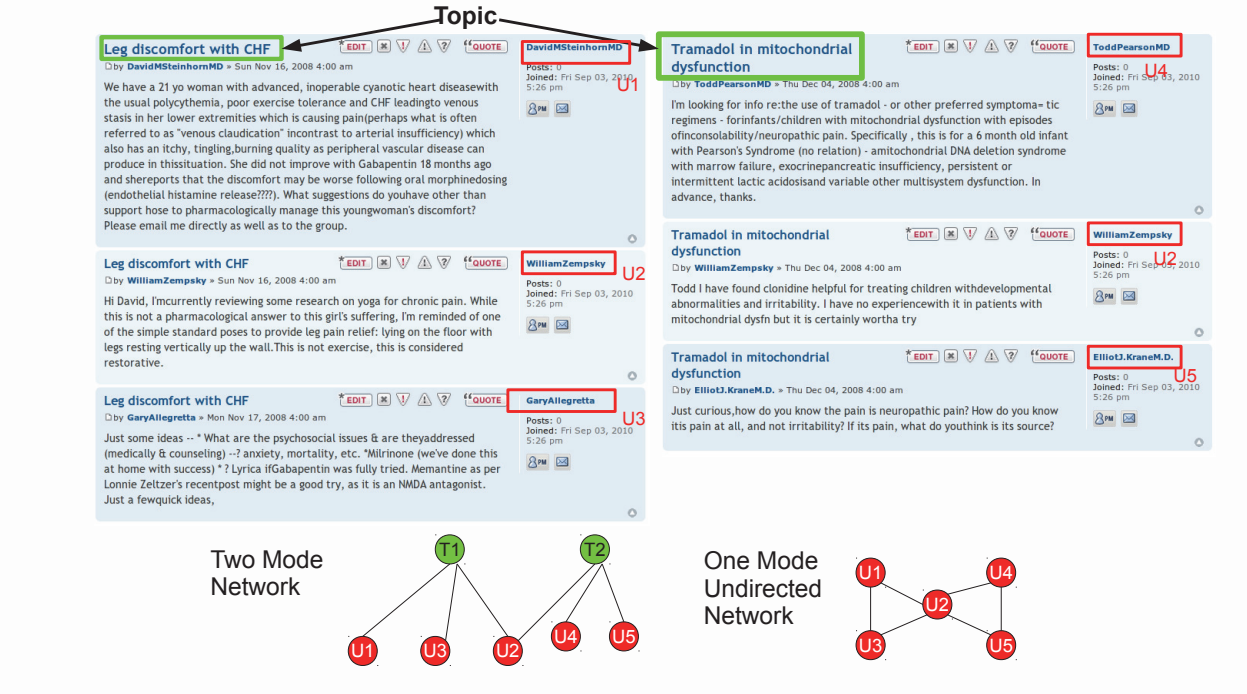


Figure 3.1: An example of how a network is built from a discussion forum thread. For mailing lists the same process can be done, where threads are determined via subject lines.

3.2.2 Measuring Community Member Activity

Simple statistical summaries can provide basic understanding of activity levels within the community. Number of posts, number of threads, posts per thread, et cetera, can provide insight into how much the community is used. Looking close at the thread-level activity can give some insight into the nature of the community. We will investigate isolates and response rates to identify when messages go unanswered, and we will use centrality metrics to identify the community leaders.

3.2.2.1 Isolates

Isolate threads are messages to the community that do not receive a reply. They can be a problem, as a question that goes unanswered can leave the poster feeling alienated and disconnected, and may result in that user leaving the community. At the same time, many isolate threads represent threads that warrant no response: announcements, advertisements and even spam are posts that are not expected to receive a reply, and are very different from unanswered questions. Identifying isolates is important, but differentiating unanswered

questions from job advertisements is a key step in understanding the network.

Isolates can be categorized into three groups: broadcasts, errors, and pendants. Broadcasts are messages that receive no reply because they do not deserve one, such as administrative messages, conference announcements, job advertisements, et cetera. Errors are processing and submission errors that cause threads to become disconnected, and are largely caused by subject-line manipulation and email-client processing problems. Pendants are the true problems within the community, as they represent knowledge seeking behaviour that goes unanswered.

What we are keenly interested in is what the effect of pendants are on the community. There are two basic questions: is there a way to predict a message will be a pendant, and is there an effect on the user if their message is a pendant. For the first question the worry is that new users, or users who are not recognized names within the community, are less likely to receive responses, as this may demonstrate a level of elitism amongst the existing community members. For the second question we want to investigate a users messaging rate both before and after their pendant thread, to ensure that not receiving a reply did not cause them to remove themselves from the community. We will investigate these two problems by looking at pendants rates based on previous activity level, in order to ensure that messages are not being left unanswered based on the activity level (i.e. the perceived prestige level) of the person asking the question.

3.2.2.2 Measuring Thread Participation Levels

Thread participation levels provide an overview of the community activity levels, and can provide insight into how the users see the community. Fast responses rates are a positive finding, and suggest that other community members are monitoring the incoming messages to provide instant feedback. Slower responses are not necessarily a problem within the community, but represent a more thoughtful and less clinically relevant community.

For threads that receive at least one reply there are two time intervals that are of particular interest: the time between the initial post and the first reply, and the time between the initial post and the last reply. Time to first reply will be used to evaluate the overall health of the community, and to try and help with the detection and prediction of pendants. Time to last reply provides insight into the attention span of the community, detailing how long ideas last within the community members.

There is potential to leverage response times to develop a pendant detection strategy for the community. For a community manager trying to prevent pendant threads the quantiles of response times can provide the means for determining when a new message may become a pendant. The actual time interval is dependent on the nature of the community, but the general structure of the algorithm is:

1. Find the time-to-first reply for all threads in the community (this could be a rolling window defined by time or number of threads)

2. Find the 90^{th} or 95^{th} quantile for these times

3. For any future thread that goes beyond that timepoint, attempt to initiate a response

This is a relatively low-effort process that could easily ensure that all threads are receiving replies in a timely manner, and fits well within the facilitation component of KT frameworks such as the PARIHS model [59, 72].

The duration of a thread can also provide insight into the nature of both the thread and the community in general. If most threads receive the bulk of their replies quickly (where quickly is defined subjectively by the nature of the community) then that may be evidence of an active community, and also one that is constantly monitoring the contents of the community for new information. If threads instead stretch out over long periods of time then the community is providing a different kind of resource: instead of providing up to the minute information, it is a place where community members can go and discuss interesting ideas, but not to go to find answers to specific, clinical questions. These issues speak to the culture of collaboration within the online community, and how the community members are using the online space for KT.

The total number of replies can also provide some insight into the duration of a particular thread. Question and answer threads occur when a community member poses a question, and another member responds in, hopefully, a timely manner with a direct answer to that question. After the response there are no more posts on the thread, as no further discussion is warranted. These threads tend to have a short duration and few replies. In contrast are discussion threads, which start with a question or a statement that is designed to elicit a discussion rather than a definitive answer. These threads tend to be longer, with many community members responding with their opinions. The division of the community into

these two broad categories provides general insight into its utility: if the members are participating because they are interested in having meaningful clinical discussions then we would expect more long threads, while more short threads would demonstrate that the community members are there to get answers to clinical questions.

Moving beyond simple investigations of the thread data, centrality measures can provide valuable insight into the overall structure of the network, along with helping identify the more significant contributors to the community.

3.2.2.3 Identifying Community Leaders

Identifying active users is key to developing and understanding an online community. These users are the ones who control the flow of knowledge within the community, and are in a position of power when it comes to what knowledge is shared between community members. Analytic methods that can identify the active users are essential to furthering our understanding of how the community functions.

Centrality measures can provide insight into the most important actors in the network. For the 2-mode network, *degree centrality* measures the number of ties an individual node has, i.e., it is a count of the number of threads a user communicates on. In the 1-mode user network degree presents the number of other users a single user has communicated with.

Closeness centrality extends the idea of degree centrality beyond a single step: It considers an actor central to the network if they can reach all other nodes in the network in as few steps as possible. A node is "close" to another node if it can reach that node in very few steps, i.e., by traversing very few ties within the network. For both the 1-mode and 2-mode networks a high closeness indicates that a user can quickly connect to another user through shared threads.

Betweenness centrality deems nodes central if they are hubs of information. Where closeness deems a node central if it can quickly reach other nodes, betweenness deems a node central if it is used as path between other nodes. A node has a high betweenness score if it falls on the shortest path of many other nodes. The higher the betweenness score, the more integral the actor is to facilitating communication between other actors in the network. As with closeness, betweenness is the same in the 1 and 2 mode networks, with the exception of how they are normalized.

All four measures can be normalized to a [0,1] scale for simpler interpretation, see Wasserman [90] and Hanneman [32] for the technical calculations of these values, and Borgatti et al [9] for the adaptation to 2-mode networks.

In section 3.5.4 we will investigate how SNA can be used to understand the BICGM similarity developed in section 3.3.8.2, so it is worth presenting some directed centrality measures for analyzing that network. With directed networks the same three centralities can be adapted to directed relationships, but only degree centrality will be of use to this project. Degree centrality can be split into two categories: in-degree is the number of connections to a user, and out-degree is the number of connections from a user.

Within directed networks there is a concept similar to centrality, and that is the idea of *prestige*. Since SNA is most often concerned with the idea of receiving ties, prestige is a way to measure incoming ties in the network. The simplest form of prestige is in-degree, which counts the number of ties in the network directed at a specific user. Along with in-degree there are other prestige measures that can provide additional insight into the directed network.

Proximity prestige extends the idea of degree prestige from all users that connect directly to user i to all users that can reach user i, and measures the average distance that those users need to travel to connect with the target user. It can be thought of as a directed version of closeness.

The idea of rank prestige is that users are prestigious if they are near other users that are prestigious, much like the idea of coreness (see below). It is a recursive definition that seems difficult to calculate, but if the adjacency matrix is restricted in certain ways then eigen-value decomposition can be used to find rank prestige. The details of the process will not be explained here, but are well explored in Wasserman and Faust's book. [90] This process can be related to the idea of hub-authority analysis [46]. In hub-authority analysis nodes that are the target of many ties are identified as "authorities" in that they must be the source of much information, and nodes that are the source of many ties are identifies as "hubs" from which links are directed to valuable information. If the adjacency matrix for the directed network is A then the first eigenvector of $A'A$ and AA' are the authority and hub values respectively.

3.2.2.4 Knowledge Translation Activities

The centrality indicators provide insight into the active members of the community and its overall structure, but we want to investigate the users farther, in order to distinguish what KT roles the users play within the community. Three specific roles are going to be investigated: knowledge seekers, facilitators and content experts.

Knowledge seekers are those that are using the list to further their knowledge. In some online communities of practice, such as online discussion forums, we would have access not only to who has contributed to the conversations but also who has read them. In this scenario we can identify knowledge seekers as those that consume the knowledge from the community but do not contribute to it. Previous research we have done has investigated this in a foreign language community [78], but such analysis is not possible here, as there is no recorded record of which community members consume content from the mailing list. We therefore will identify knowledge seekers as those that initiate conversations. Most knowledge-based messages within a mailing list are questions about a specific problem or paper, and result in a discussion about the problem. The initiators play an important role in this process, as their questions are what draws the vital information out of the content experts.

The second role we will investigate are facilitators, those users that encourage further conversation by their posts. These people respond early in the thread and in a timely manner, and their replies spur further conversation. They are often more active members of the community, and their interest in a subject results in other users engaging the conversations. Facilitators play a vital role in the PARIHS framework [59, 72] and other KT frameworks, as they are the ones that encourage knowledge uptake and connect seekers to experts.

The final role we will investigate are content experts. These are users that finish conversations by providing answers to specific questions. They respond to question threads with concrete answers that often end the thread, or else only result in responses by other content experts.

3.2.3 Identifying Collaboration Groups: Connection Clustering

Within an online community subgroups may be expected to form. These groups may be separate from the community as a whole, or they may represent the leaders within the community, but either way identifying potential subgroups within the community is vital to understanding the KT patterns within the community. Connection clustering will look

at the network at a macro level, attempting to identify potential sub groups of users based on their communication patterns (through 1-mode and 2-mode clustering) and identifying a core group of users within the community (through core-periphery analysis).

3.2.3.1 Detecting Connection Subgroups

In a study of the evolution of online communities, the "death" stage of the cycle has been partially attributed to the segmentation of the community into disparate subgroups. [39] The problem with segmentation is that, if you are only interested in a specific sub-topic of a community then the rest of the messages become "noise", and eventually the noise will overwhelm the meaningful content, causing users to ignore all content from the community. It is vital, therefore, to try and stay ahead of this segmentation by using methods for automatically detecting subgroups within the community, in order to better serve them, perhaps by splitting your community into smaller communities with a more focused topic. SNA provides a function for detecting subgroups in the form of blockmodeling.

A blockmodel is a partitioning of the network into exclusive, non-overlapping groups, such that most communications are within groups rather than between them. Traditional SNA uses structural equivalence blockmodeling for the undirected 1-mode networks as well as a more generic hierarchical agglomerative method similar to section 3.3.9, and generalized blockmodeling for the 2-mode network.

Formally, two nodes are structurally equivalent (SE) if they have the same ties to all other nodes in the network. If two nodes are SE then one can replace the other without interfering with the flow of information. In reality true SE is rare, so approximate SE needs to be measured. There are many different methods used for approximating structural equivalence, the most simple of which is Hamming distance. The Hamming distance between two nodes is the number of ties that would have to have to change in order for the nodes to be SE.

Regardless of which SE measure is used, a SE matrix is developed, which records the SE between all the actors or threads. This SE matrix can be thought of as a distance matrix, at which point the hybrid clustering methods from section 3.3.9 can be used. There are two general problems with SE in this scenario, however. The first problem is that, since the data is very sparse, similarity between active users is rare, and the SE found in the network will usually only be between users with few posts. SE also fails to capture the difference in messaging rates between users. A more interesting metric, rather than SE, would be number

of threads shared. This is captured directly from the 1-mode actor network. Taking the number of threads shared as a similarity metric results in a user clustering similar to that in section 3.3.9, only these clusters are based on communication patterns rather than the content of the communications itself.

In a 2-mode network the question of clustering can go beyond finding what users are communicating on threads together, but what users are communicating on *which* threads together. Generalized blockmodeling [25] can provide the answer to this question, by partitioning the network such that the users and the threads are clustered concurrently. The general idea is to partition the rows and the columns of the matrix into groups such that the groups are as pure as possible (either all connected or all disconnected). Generalized blockmodeling is performed using a local optimization procedure [25, 99].

3.2.3.2 Identifying the Core of the Community

In an online community past experience suggest that the bulk of the communication is performed by a minority of the users. The "Pareto principle" is a theorem that states that 80% of the work is done by 20% of the population. The exact size of the "core" of an online community varies by application: some studies have found the core to comprise upwards of 50% of the users [64, 87, 88] while other studies have found the numbers to be smaller [8]. Regardless of the specific sizes, the principle is that online communities are expected to have a core group of users that contribute the majority of the knowledge to the conversations. This core group of users can be identified using core-periphery analysis.

Core-Periphery analysis assumes that there is a core set of nodes at the centre of the network, and a periphery set of nodes that connect to that core. [9, 10] It can be used to identify the community members that are at the centre of the one and two-mode networks. For the undirected 1-mode member network a measure of "coreness" can be calculated that is a measure of how central the member is to the network. This coreness can be thought of as another measure of centrality. With appropriate row/column normalization this coreness centrality can be determined as the first eigenvector of the adjacency matrix of the 1-mode network.

3.2.4 Summary

It is vital to understand how KT is performed within an online community. This section investigated methods for evaluating online communities to better understand their collaboration patterns. It detected response and thread duration times, and presented a simple and effective algorithm for preventing threads from becoming pendants that receive no response. it looked at centrality measures to provide insight into the leaders within the community, and through clustering and core-periphery analysis it developed methods for identifying potential subgroups of interest.

The problem with all of these methods is the lack of insight into what is being said within the messages. In a knowledge-based online community it is imperative that we look beyond the communication patterns to study the actual content of the messages. The next section will look at mapping the content of the messages to a formal medical lexicon, where it can then be used to better understand the users and threads in the community.

3.3 Directing Knowledge Content

Understanding what knowledge is being shared within the community is vital to directing the KT practices of the community members. This section will use semantic mappings of the content to a formal medical lexicon to better understand the content being shared using knowledge maps. The relationships between these mapped terms can provide greater insight into the content of the users and threads through content-based clustering. The content mappings will provide a method for detecting similarity between users, which can be used to increase connectivity by connecting users to other like-minded individuals.

3.3.1 Knowledge Maps

A knowledge-based online community is usually centred around a medical topic: Pediatric Pain or General Surgery are two examples from this thesis. Within those fields, however, there is a vast range of potential subjects that may be of interest to the community. Monitoring the specific content being shared by the users can provide insight into what the community members are interested in, and may provide mechanisms for guiding users toward less popular subjects that the community administrators want to discuss, or for recruiting new users that may provide valuable insight into particular content. This thesis will present

this content using Knowledge Maps.

Knowledge maps provide a detailed summary of the general subject areas that the community expresses an interest in. The semantic terms returned by Metamap [6] (or by any semantic mapping program) provide a knowledge-based representation of the messages, and therefore of the community as a whole. Summaries of the mapped terms that make full use of the inherent relationships within the medical taxonomy can provide detailed insight into the content being shared.

This project will use mappings to the MeSH lexicon. MeSH is designed in a hierarchical structure, such that terms may have one or more parents and/or one or more children within a tree-like structure (a directed acyclic graph). At their root there are 16 different groups of terms (noted by letters *A-N, V* and *Z*) that represent very broad groups of terms around a single idea. Root *A* is "Anatomy", and all the medical terms within that tree are related to the physical body parts. Root *D* is "Chemicals and Drugs", and represents the chemical components used in medicine, including all natural and synthetic medications. Combining these 16 roots and their immediate children can provide a broad representation of the community in terms of what knowledge is most interesting to the community as a whole.

3.3.2 Content-Based Similarity

In the online KT process establishing similarities between threads or members of the community is important to facilitating future KT. Linking like-minded individuals within the community based on the similarity of their communications or identifying threads similar to the content a user is currently reading can improve the knowledge base of users through leveraging the existing archives of the community. This process exploits a semantic mapping of the content of the online discussions with a standard medical lexicon (MeSH) using Metamap [6]. Metamap standardizes the unstructured messages, and provides a semantically-enriched representation of the message content, allowing us to aggregate across messages to represent users with the content of their messages.

Once the mappings are established different algorithmic approaches to identifying potential relationships between users or threads based on the content of their online communications will be used. We will explore Generalized Vector Space Models (GVSM) [95] and the

Balanced Genealogy Measure (BGM) [29] as two different approaches to calculate these similarities. Each method needs to leverage the inherent relationships within the MeSH terms, resulting in two separate, novel approaches to calculating user or thread similarity. We will then use those similarity calculations to identify potential clusters within the community.

3.3.3 Notation

A user $\overrightarrow{v_i}$ is represented by a vector of their semantic terms, t_{ij}.

$$\overrightarrow{v_1} = [t_{11}, t_{12}, \ldots, t_{1n}]_{1\times n} \tag{3.1}$$

All the users can be combined into a single matrix, V.

$$V = \begin{bmatrix} t_{11} & t_{12} & \ldots & t_{1n} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ t_{k1} & t_{k2} & \ldots & t_{kn} \end{bmatrix}_{k\times n} = \begin{bmatrix} \overrightarrow{v_1} \\ \overrightarrow{v_2} \\ \vdots \\ \overrightarrow{v_k} \end{bmatrix} \tag{3.2}$$

The user's semantic terms are extracted from their messages. These terms need to be scaled in order to avoid biasing issues, so we'll define a user $\overrightarrow{u_i}$ by his/her scaled mesh terms a_{ij}, and the matrix of users U.

$$\overrightarrow{u_1} = [a_{11}, a_{12}, \ldots, a_{1n}]_{1\times n} \tag{3.3}$$

All the users can be combined into a single matrix, U.

$$U = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{k1} & a_{k2} & \ldots & a_{kn} \end{bmatrix}_{k\times n} = \begin{bmatrix} \overrightarrow{u_1} \\ \overrightarrow{u_2} \\ \vdots \\ \overrightarrow{u_k} \end{bmatrix} \tag{3.4}$$

Let C be a matrix of similarities between semantic terms. c_{ij} is the similarity between two terms. Note that, since each user was represented by a $1 \times n$ vector there are n terms

in the corpus overall, therefore the correlation matrix is $n \times n$.

$$C = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nn} \end{bmatrix}_{n \times n} \tag{3.5}$$

A thread can be represented by a vector of semantics terms in the same manner as a user. Let $\overrightarrow{h_i}$ be the i^{th} thread in the community.

$$\overrightarrow{h_1} = [t_{11}, t_{12}, \dots, t_{1n}]_{1 \times n}$$

All threads can be combined into a single matrix, H.

$$H = \begin{bmatrix} t_{11} & t_{12} & \dots & t_{1n} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ t_{m1} & t_{m2} & \dots & t_{mn} \end{bmatrix}_{m \times n} = \begin{bmatrix} \overrightarrow{h_1} \\ \overrightarrow{h_2} \\ \vdots \\ \overrightarrow{h_m} \end{bmatrix} \tag{3.6}$$

Unlike the user-term representation, there is no-need to perform message-level scaling of the term-representations for the thread-term matrix H. The purpose of the message level scaling in the user representation was to avoid the potential for particularly long messages to bias the representation of a user. Since the knowledge in a thread is drawn from the messages contributed to it, longer messages represent more knowledge contributed to the conversation. The scaling for users was because users could potentially be interested in several different knowledge areas, but this is not the case for threads, which should be grouped around a common subject.

Let $H_2 = HC$ be the semantically scaled matrix. The purpose of the H_2 matrix is to incorporate the semantic and co-occurrence correlations into the thread-term representation in matrix H. The value of the representation H_2 is that the columns of the matrix, the MeSH terms, can now be thought of as being quasi-independent, which makes it easier to apply clustering methods to them. Below is a simple example of the calculation of H_2.

$$H = \begin{bmatrix} 3 & 0 & 5 & 2 & 1 \\ 2 & 1 & 0 & 0 & 0 \\ 4 & 3 & 1 & 0 & 1 \\ 0 & 0 & 0 & 3 & 0 \\ 4 & 6 & 3 & 2 & 8 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad C = \begin{bmatrix} 1.0 & 0.9 & 0.1 & 0.0 & 0.0 \\ 0.9 & 1.0 & 0.5 & 0.0 & 1.0 \\ 0.1 & 0.5 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 & 1.0 \end{bmatrix} \quad H_2 = \begin{bmatrix} 3.5 & 6.2 & 5.3 & 2.0 & 1.0 \\ 2.9 & 2.8 & 0.7 & 0.0 & 1.0 \\ 6.8 & 8.1 & 2.9 & 0.0 & 4.0 \\ 0.0 & 0.0 & 0.0 & 3.0 & 0.0 \\ 9.7 & 19.1 & 6.4 & 2.0 & 14.0 \\ 0.0 & 1.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}$$

Note that a user-term semantically scaled matrix could also be created in the same manner, denoted $U_2 = UC$.

3.3.4 Scaling

The idea of scaling is to modify user-term representation such that the most common terms do not bias the overall representation. A user or thread with a high density in a particular term is clearly more represented by that term, but if it is a rare term then it must be a more powerful representation than common terms. Scaling will occur at two levels: at the user level, and at the corpus level.

Users are comprised of a set of messages. Each message is then mapped to a set of semantic terms, which are aggregated to create the user vector $\overrightarrow{v_i}$. One of the major challenges in this representation is the difference in message size within a specific user. If a user contributes, for example, 5 messages to the community, and one of those messages is significantly larger than the others, then it will dominate that user's overall mapping. Consider figure 3.2, which shows the overall mapping score, i.e., the sum of all the mappings, for each message for a sample of users, with breaks in each bar indicating an individual message. User 963 has 8 messages, however half of his total mappings come from his two largest messages. This means that, if the mappings are just added up, the user will be disproportionately represented by these two messages.

There are three possible options we will investigate for message-level scaling: boolean scaling, normalization and log scaling. In order to explain the scaling some notation needs to be introduced. Let user v_i have messages $m_{ij}, j \in [1, g_i]$. For each message a score, s_{ijk}, is recorded, $k \in [1, n]$. Note that, since n is the total number of mapped terms in the dataset,

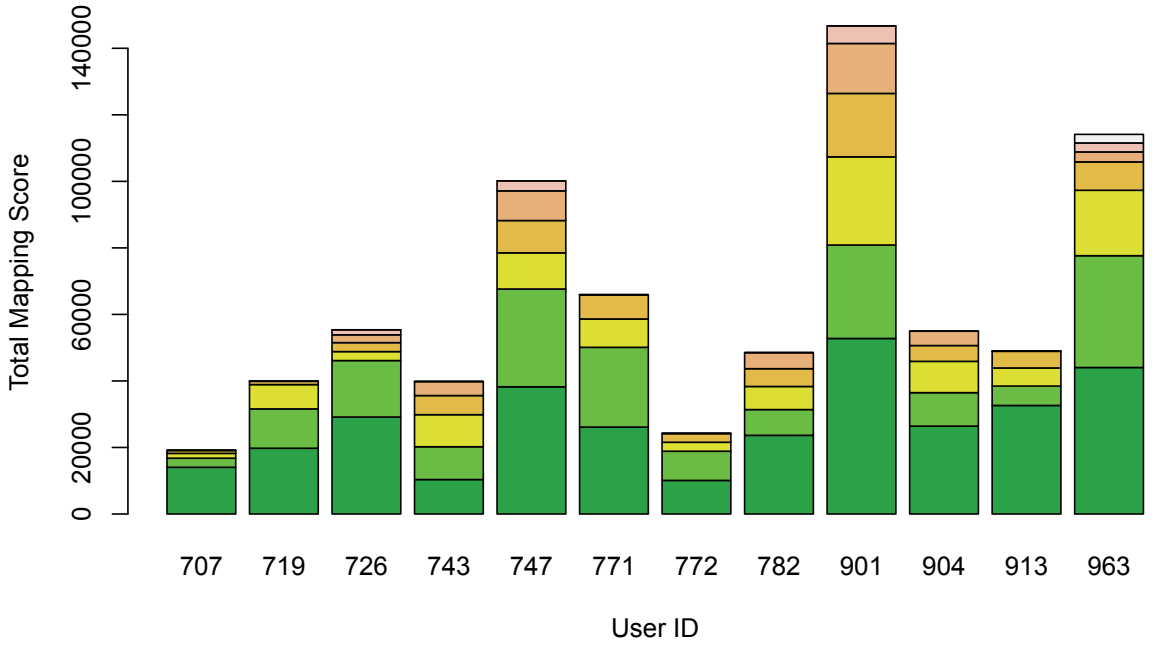


Figure 3.2: The total message scores for a sample of users from the PPML. The breaks within each bar represent a separate message.

most $s_{ijk} = 0$. Figure 3.3 shows the mapping for a particular user with three messages.

Relating back to the notation from section 3.3.3, the representative terms for a user are aggregated across all the messages, as demonstrated in the equation below.

$$t_{ij} = \sum_{l=1}^{g_i} s_{ilj}$$

We are going to look at three other methods for determining the t_{ij} values, starting with boolean scaling. Boolean scaling reduces all message-level scores to a 1/0, where a 1 indicates that the message was mapped to that term, and a 0 indicates that it did not. The formal equation is given in equation (3.7). In this case a user is represented not by the strength of their messages, but by the number of different messages that map to the same term. If a user maps to the same term over multiple messages then this should be represented in the user's scores, but for users with few messages or short messages, boolean scaling could have

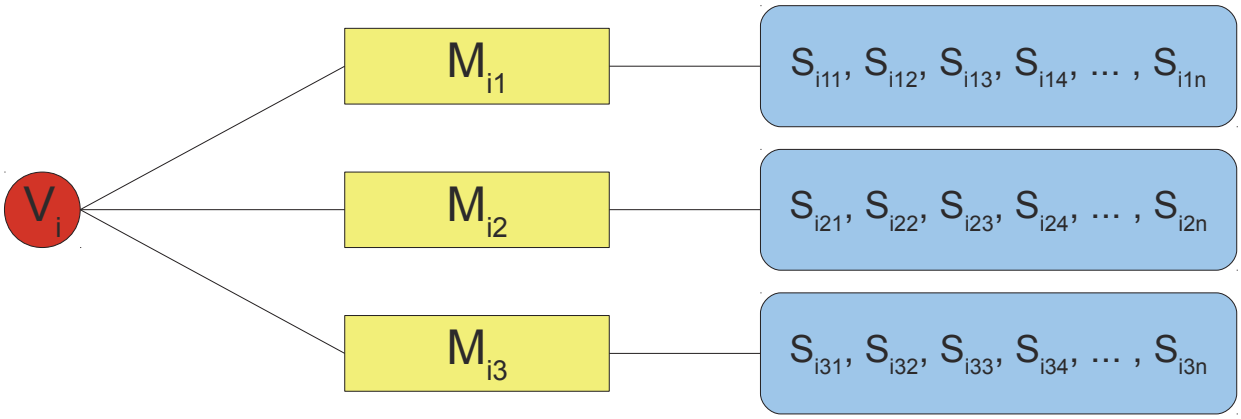


Figure 3.3: The mappings for a single user stratified by message

the effect of clouding the results by assigning all terms essentially the same score.

$$t_{ij}^{b} = \sum_{l=1}^{g_i} I(s_{ilj} > 0) \tag{3.7}$$

Normalization adapts the idea of boolean scaling, but within a message it assigns a value on a $[0, 1]$ scale that measures how important that term was to that specific message. This leverages the boolean scaling principle of multiple messages with the same term being more important, but adapts to the idea that, within a message, not all terms are necessarily equal. Equation (3.8) has the formula for normalization scaling.

$$t_{ij}^{n} = \sum_{l=1}^{g_i} \frac{s_{ilj}}{max_{k\in[1,n]} s_{ilk}} \tag{3.8}$$

Finally, log scaling is taking the log-transform of each message score. This has the effect of reducing the overall difference between the highest and lowest scoring terms, while also rewarding terms that occur in multiple messages, as evidenced in equation (3.9). The log terms are a log-transform of the product of terms occurring across multiple messages. Higher scoring terms that appear across multiple messages will score much better than high-scoring terms that occur in only a single message, while still rewarding those terms that score higher. Note that, because $log(0)$ is undefined, the term score is only calculated over the mapped terms, and not over all terms. If iterating over all terms is the only option, using $log(1 + s_{ilj})$

would also work.

$$t_{ij}^{l} = \sum_{l \in [l,g_i] | s_{ilj} > 0} log(s_{ilj}) = log\left(\prod_{l \in [l,g_i] | s_{ilj} > 0} s_{ilj} \right) \tag{3.9}$$

Of the three methods, normalization seems the most appropriate. Boolean scaling is an option, but for users with only 1 or 2 messages in the community they end up with little differentiation between their terms, and it seems like log transforms are strictly better, as, within a single message, they maintain the differentiation between users. For comparing log-scaling to normalization, consider figure 3.4, which shows the scores for 7 active users from the PPML, sorted by overall score. The figure demonstrates the similarity between counts and log-scaled scores, while maintaining some differentiation between terms that have the same count-level.

Log-scaling still leaves a large gap between the largest couple of mapped terms and the rest of the terms, however. It is effective in increasing the importance of terms that map multiple times, but it should be noted that normalized terms usually spike at the same points.

I believe that message-level normalization is the most appropriate scaling for the scores. If a user is represented by what they talk about, then they should be equally represented by all their messages. Assigning any extra weight to messages that are long implies that a longer message is inherently more interesting to the writer than short messages, where in reality the length of the message is dictated by far more than interest in the subject matter. Message-level normalization significantly rewards mappings to terms from multiple messages. The risk with this method is the potential for short messages with few mappings can bias the representation of a user, but we should be able to expect that terms returned by the mappings are accurate representations of the messages, so even if a message is short its content is valuable.

Moving on from user-level scaling to corpus level scaling, the most common form of scaling is Term Frequency - Inverse Document Frequency (TF-IDF) scaling. We will first define the term frequency, $tf(t_{ij}, v_i)$. There are two potential definitions we will investigate: the normalized frequency (equation 3.10) and the log-scaled frequency (equation 3.11).

$$tf(t_{ij}, v_i) = \frac{t_{ij}}{max(t_{ik} | k \in [1, n])} \tag{3.10}$$

In the normalized equation (3.10) each user or thread is scaled such that their individual

scores are on a [0,1] scale, with higher scores indicating more use. This weights all users or threads on the same scale.

$$tf(t_{ij}, v_i) = \begin{cases} 0 & if\ t_{ij} = 0 \\ 1 + log(t_{ij}) & otherwise \end{cases} \tag{3.11}$$

In the log-transformed equation (3.11), each score is scaled down by the log transform. This reduces the effect of larger terms, while still maintaining the score differences between two users.

The inverse document frequency, $idf(t_i, V)$, is the ratio of the size of corpus (either the number of users or the number of threads) divided by the number of users or threads that mapped to a specific term. For rare terms this number will be large, indicating that it is a more informative term than the more common terms.

$$IDF(_i, V) = log\left(\frac{|V|}{|\{v_j \in V; j \in [1,k]|t_i \in v_j\}}\right) \tag{3.12}$$

The final scaled value is a product of tf and idf.

$$a_{ij} = tf(t_{ij}, v_j) \times idf(t_i, V) \tag{3.13}$$

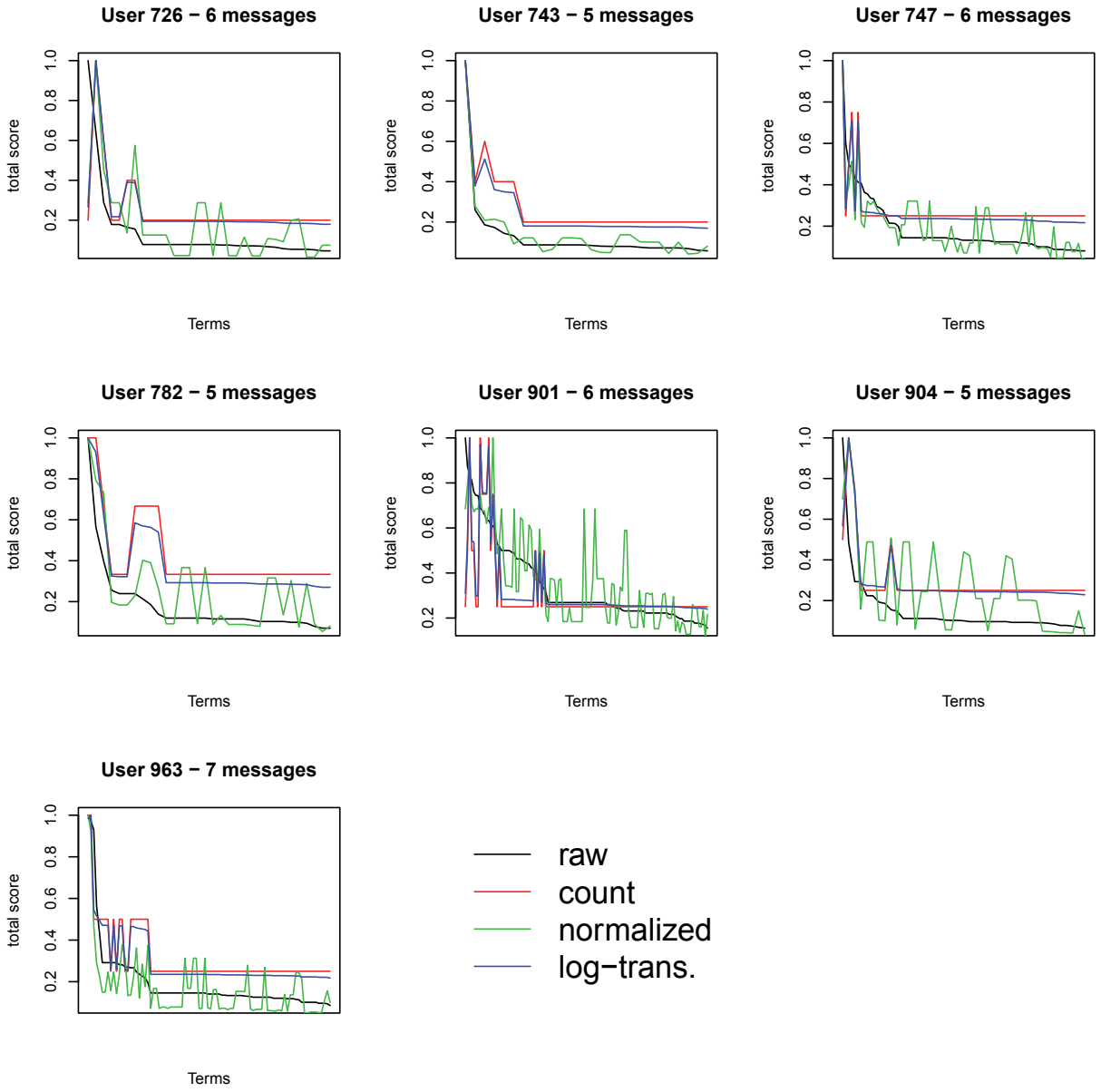


Figure 3.4: The scores for a user, scaled to a [0,1] scale, as scores will only be compared relative to one another.

3.3.5 Directed Similarities: The Balanced Genealogy Measure

The objective of the directed similarity measures is to find a way to calculate the similarity between users that a) incorporates the inherent relationship between the MeSH terms into the calculation, and b) captures the asymmetric nature of the relationship. For two users A and B, A may be interested in B more than B is interested in A, especially if B is a content expert in a number of fields and A is a junior community member only interested in a specific area. Ganesan and colleagues [29] suggest two adaptations for calculating asymmetric similarity between objects whose elements are semantically related. The Balanced Genealogy Measure (BGM) and the Recursive Genealogy Measure (RGM) provide two different approaches to finding similarities between users based on their representations by taxonomically related terms. The BGM method seems more appropriate for our data than the RGM, so it will be explored in this thesis.

The Balanced Genealogy Measure (BGM) is an iterative method that is an adaptation of the authors' own Optimistic Genealogy Measure (OGM). Consider the vector representation of two users, u_1 and u_2. Their mappings produce two subtrees of the full MeSH tree, T_1 and T_2. We want to determine how similar the two subtrees are.

1. For each leaf l_1 in T_1 *visited in optimal order* (more on this later), find the leaf l_2 in T_2 that has the lowest common ancestor in the tree. This node can be defined as $LCA_{T_1,T_2}(l_1)$.

2. Increment l_2's *match count*, i.e., the number of times that this term has been used as a match for a leaf from T_1.

3. Define two equations

$$optleafsim_{T_1,T_2}(l_1) = \frac{depth(LCA_{T_1,T_2}(l_1))}{depth(l_1)} \tag{3.14}$$

$$leafsim_{T_1,T_2}(l_1) = optleafsim_{T_1,T_2}(l_1) \times \beta^{matchCount(l_2)-1} \tag{3.15}$$

The value of $optleafsim_{T_1,T_2}(l_1)$ is the ratio of how well the leaf l_1 matches the tree T_2. If l_1 is present in T_2 then the $optleafsim_{T_1,T_2}(l_1)$ will be 1, and if l_1 has no ancestors present in T_2 then $optleafsim_{T_1,T_2}(l_1)$ will be 0. The $leafsim_{T_1,T_2}(l_1)$ value scales the optimal value by the number of times the leaf it matched to has been used as a match

within the tree. As explained in the original paper, [29] if multiple elements from one subtree continually match to the same element in the second tree then the two trees are less similar.

4. Finally, the similarity between the two users is calculated using equation 3.16, where $W()$ is the weighting equation for the semantic terms (TF-IDF, for example).

$$sim_{BGM}(u_1, u_2) = \frac{\sum_{l_1 \in U_1} leafsim_{T_1,T_2}(l_1) \times W(l_1)}{\sum_{l_1 \in U_1} W(l_1)} \tag{3.16}$$

The BGM algorithm requires the leaves of the tree be visited in *optimal order*. This means that they are iterated through such that the similarity between users is as high as possible. The general approach is to find matches iteratively based on their *optleafsim* value, i.e., first find direct matches, then matches with a common parent, then a common grandparent. The proof is left to the original work [29], but this approach results in an optimal similarity measure.

The algorithm is dependent on a single coefficient, $\beta \in [0, 1]$ that measures the penalty for multiple matches. $\beta = 1$ suggests that there is no penalty, while $\beta = 0$ suggests that a node in T_2 can be used at most 1 time in matching to T_1.

Once we investigate semantic correlations and context-based information through Information Content (section 3.3.7), section 3.3.8.2 will outline how the BGM method must be adapted to non-leaf mappings and issues of homonymity, along with extensions to incorporate context specific information, resulting in a novel correlation calculation.

3.3.6 Symmetric Similarities: Vector Space Models

Detecting similar users or threads within a community is vital to increasing connectivity between users. The connection clustering in section 3.2.3.1 provides a connection-based method for detecting similar users, but these methods fail to incorporate the content of the messages shared. In this section we will investigate a content-based method for detecting similarity between users that incorporates the relationships between the MeSH terms used to represent the users and/or threads.

For each user or thread, think of their set of semantic terms as representing them in n-dimensional space. With any two vectors, their similarity/difference could be measured by the angle created between the vectors. Vector Space Similarity uses vector and trigonometry

theory to leverage the cosine of the angle between two vectors to measure their similarity, given in equation (3.17).

$$sim_{VSM}(u_1, u_2) = \frac{\sum_{i=1}^{n} a_{1i} a_{2i}}{\sqrt{\sum_{i=1}^{n} a_{1i}^2 \sum_{i=1}^{n} a_{2i}^2}} = \frac{u_1 \bullet u_2}{||u_1|| \times ||u_2||} \tag{3.17}$$

The problem with equation 3.17 is that it assumes orthogonality of the mapped terms, i.e., when we represent a user as a vector of their MeSH terms we make an assumption that the terms are independent of one another, which is clearly not the case for MeSH terms, or in fact for any medical taxonomy. Common approaches are to apply some sort of Singular Value Decomposition or eigenvector scaling to normalize the term matrix U or H such that the new representation is orthogonal, but those methods do not make full use of the inherent relationships between the dimensions of the user-vectors. Rather than trying to factor or the dependence between terms we will incorporate the term relationships into the calculation. Using the term similarity matrix C, we can calculate user or thread similarity using the *Generalized Vector Space Model*.

The Generalized Vector Space Model (GVSM) [95] adapts the VSM to deal with the problem of non-orthogonality between the terms used to represent documents. VSM represent each v_i as a unit vector in n-dimensional space. The GVSM representation maintains the independence of the representative vectors, but defines the term vectors in 2^n space to account for all possible correlations. It turns out that, though the theoretical representation is in 2^n space, the calculation of these vectors is not needed, only the correlation between the terms. Equation (3.18) presents the vector form of the equation.

$$sim_{GVSM}(u_1, u_2) = \frac{u_1 C u_2'}{\sqrt{u_1 C u_1'}\sqrt{u_2 C u_2'}} \tag{3.18}$$

The GVSM allows the use of non-orthogonal representations of users or threads, but requires a term-correlation matrix C. The next section will investigate semantic methods and co-occurrence methods for calculating the correlation matrix C, as well as introduce our own custom method for combining the two approaches.

3.3.7 Calculating Term Similarity

There are several ways that two terms can be related. Because we are mapping to a taxonomy, there is a natural taxonomic relation between terms. Much research has been done on calculating correlations between terms within a taxonomy, and section 3.3.7.1 below outlines some of the methods that have been used. As well, there are co-occurrence relations between terms. Two terms that are not semantically related may still be correlated if they occur within the same messages, or within the same threads, or are used by the same user. Section 3.3.7.2 will explore the co-occurrence measures that can contribute to the correlation, and section 3.3.7.3 will explore the extension of the co-occurrence methods to semantic-based methods. Finally, section 3.3.8.1 will provide a method for combining semantic and co-occurrence based correlations into a single correlation framework.

3.3.7.1 Correlations Within a Taxonomy

Within a taxonomy it is somewhat intuitive that terms that are close to one another within the tree are somewhat correlated, but how is that correlation quantified? Consider the tree in figure 3.5. Nodes 1, 2 and 3 are similar because they share a common ancestor, as do nodes x, y and z. How do we quantify that similarity? Are nodes 1, 2 and 3 more similar than a and b, because they are deeper in the tree? And how do you deal with nodes like node 2, that have multiple parents (at multiple depths) within the tree?

The literature on semantic distance within ontologies can provide the means to calculate these semantic similarities. The two approaches that are most used within the literature are edge-based and information content based, along with efforts to combine the two to create more sophisticated methods.

Rada and colleagues [68] first explored edge-based methods when they proposed measuring similarity between two nodes within a taxonomy based on the minimum distance between them (i.e their geodesic). His method is intuitive and simple, but there are inherent problems with it. Not all edges within a taxonomy are created equal, some nodes are naturally "closer" to one another than others. Li and colleagues [51] explored modifying the shortest path calculations between two nodes by combing path length, depth of their lowest common ancestor (LCA) and the local density at the LCA.

Resnik [70] looked at information-based methods for calculating similarity using Information Content. Let $p(c)$ be the probability of encountering an instance of the taxonomy c.

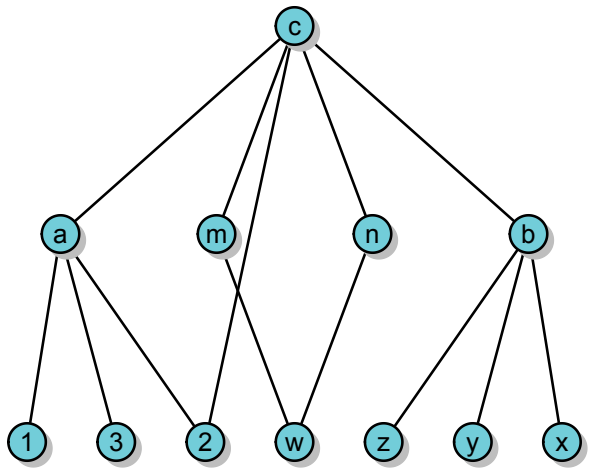


Figure 3.5: A sample of a subtree within a taxonomy

This probability function is monotonic increasing, so if c_1 is a concept that has a parent c_2 then $p(c_1) \leq p(c_2)$ and $p(root) = 1$. The information content of a node is the negative-log transformation of this probability.

$$IC(c) = -log(p(c)) \tag{3.19}$$

As the probability of encountering a term within the taxonomy increases then the relative value of the information it provides decreases. If we define $S(c_1, c_2)$ as the set of common ancestors for two nodes c_1 and c_2 then the similarity between c_1 and c_2 is the maximum information content over that set (see equation 3.20), i.e., it is the Information Content of their lowest common ancestor. Note that this is a similarity and not a correlation, because it has a maximum value of IC_{MAX}, which is variable and dependent on the implementation.

$$sim(c_1, c_2) = max_{c \in S(c_1, c_2)} IC(c) \tag{3.20}$$

One of the major challenges in using information-based methods is the calculation of $p(c)$. Resnik suggested parsing a large corpus related to the taxonomy to determine the probabilities of encountering any individual term within the taxonomy, and then propagating the results up the tree to give higher nodes a higher probability. In our corpus (the community

messages) we are considering threads as the major unit of study, therefore if we let t_i be semantic term i then a natural interpretation of $p(t_i)$ is the probability of encountering term i in a specific thread.

$$P(t_i) = \frac{number\ of\ threads\ with\ t_i}{number\ of\ threads}$$

And the information content of a term would be given by equation (3.19). Note that, in order to maintain the mathematical requirements of IC, $P(t_i)$ will count all occurrences of t_i or any children of t_i. Consider figure 3.6 for a simple example of how the IC will be calculated for terms in the taxonomy.

With usable information content measures at each of the nodes, the next step is determining what equation to use for calculating similarity between terms. Resnik's methods would provide the simplest method, as given in equation (3.20). Other researchers have explored adaptations to this method that expand it to incorporate depth and density information. Jiang et al [40] used information content to provide edge-weights for edge-based counting methods. Let $E(p)$ be the number of child links for a node p, let $\overline{E}$ be the average number of child links per internal node, and let $d(p)$ be the depth of node p. The weight of the link between a child c and its parent p is given in equation (3.21). The equation is dependent on two parameters: $\alpha \geq 0$ controls the effect of the depth of the parent node on the weight, and $\beta \in [0, 1]$ controls the effect of local density on the weight.

$$wt(c, p) = \left(\beta + (1 - \beta)\frac{\overline{E}}{E(p)}\right)\left(\frac{d(p) + 1}{d(p)}\right)^{\alpha}[IC(c) - IC(p)] \tag{3.21}$$

The overall distance between two nodes would be the sum of the weights along the shortest path between the two nodes. Taking the inverse of the distance gives a similarity metric (as in equation (3.22)). Once again, this is a similarity metric rather than a correlation, as its minimum value is $1/(2 \times IC_{MAX}) + 1$, so it is once again variable and context-specific.

$$Dist(t_1, t_2) = \sum_{c \in path(t_1, t_2)} wt(c, parent(c))$$

$$cor(t_1, t_2) = \frac{1}{1 + Dist(t_1, t_2)} \tag{3.22}$$

If we set $\beta = 1$ and $\alpha = 0$ to remove the depth and density effect, equation (3.22) can be reduced to the equation below (note that $LCA(t_1, t_2)$ denotes the *lowest common ancestor*

of the two terms).

$$sim(t_1, t_2) = \frac{1}{IC(t_1) + IC(t_2) - 2IC(LCA(t_1, t_2)) + 1}$$

Lin's proposal [52] for a semantic similarity calculation is a scaled version of Resnik's, as given in equation (3.23). The scaling puts the measure (now a correlation measure) on a $[0, 1]$ scale, while maintaining all the desired attributes of a similarity measure.

$$cor_{SEM}(t_1, t_2) = \frac{2 \times \left[max_{c \in S(c_1, c_2)} IC(c)\right]}{IC(t_1) + IC(t_2)} \tag{3.23}$$

The previous approaches all use the lowest common ancestor of two terms to calculate similarity. This is sufficient for a strict hierarchical taxonomy, but many medical taxonomies, such as MeSH, are Directed Acyclic Graphs (DAG) rather than strict hierarchies, as nodes occur in multiple locations within the tree. For terms t_1 and t_2 there may be several disjoint paths between them, and all but the shortest (in terms of IC) are removed from the calculation. Couto et al. [18] developed methods for addressing this problem. For each of the disjoint paths between t_1 and t_2, calculate the similarity using whichever of the similarity calculations you favour, then take an average of those similarities to derive a global similarity measure between the two terms.

These semantic correlation calculations provide the means to capture the inherent relationships between concepts within a taxonomy, but they do not address the contextual relationships between terms within a community. Within a knowledge-based community there will be certain concepts that are not semantically related but will be conceptually related within the context of the community. The simplest example would be the terms *Pediatrics* and *Pain* within the PPML: These two terms do not have any semantic relationship, but within the context of the PPML they are highly related. The next section will explore methods for supplementing the semantic correlations with contextual correlations to provide a context-specific correlation structure.

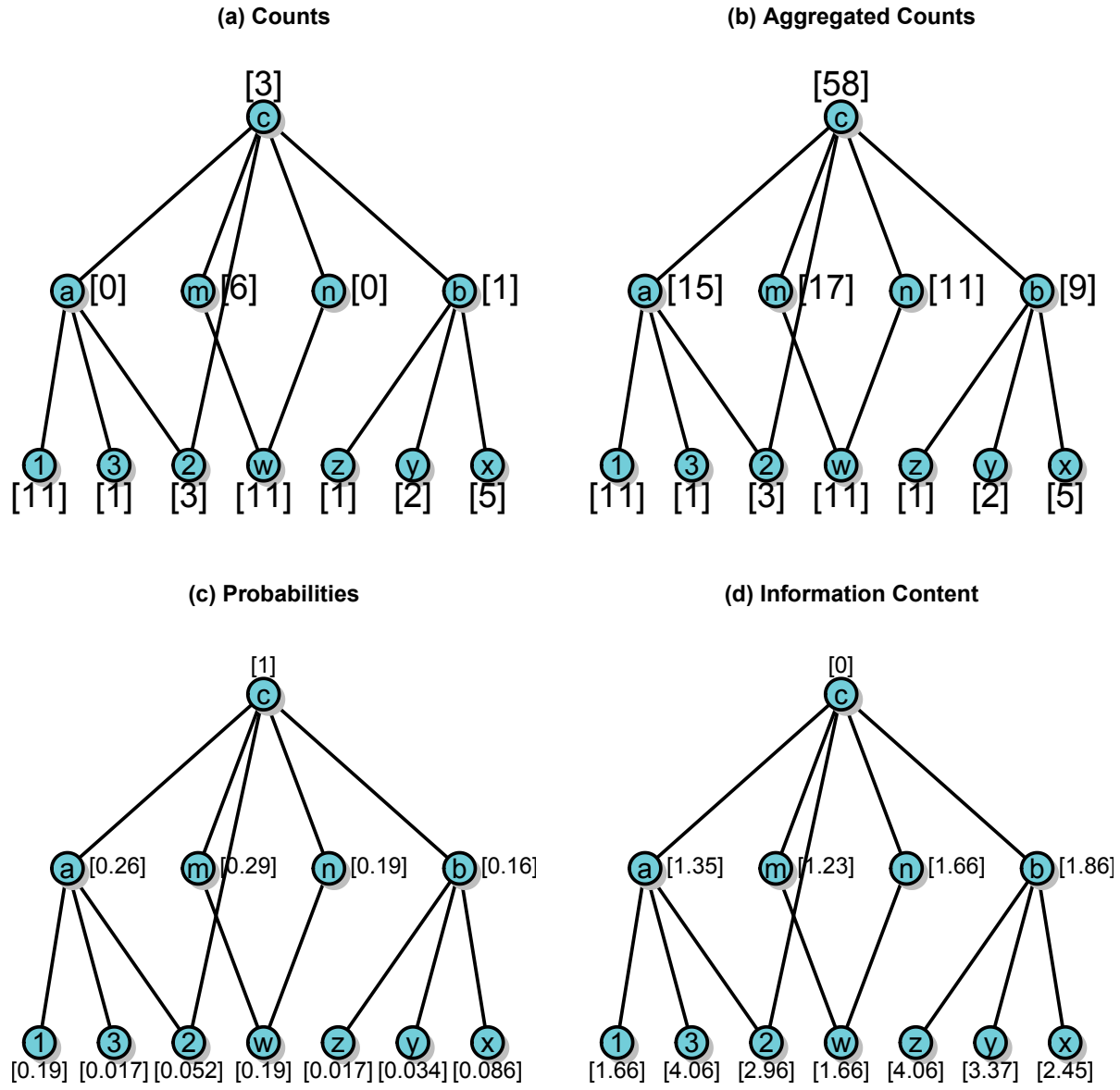


Figure 3.6: An example of how IC is calculated. The counts in figure (a) are the number of threads that each term was mapped to. These are aggregated up the tree in (b), then made into probabilities by dividing by the number of threads in (c), then converted to Information Content in (d) using equation (3.19)

3.3.7.2 Co-occurrence Calculations

Many correlation methods use term co-occurrence within the corpus to provide a measure of similarity between terms. Within the document retrieval literature, if terms occur in the same document then they are more likely to be similar than words that do not occur in the same document. When applying these methods to user similarity, the dimension across which the co-occurrences are measured is not straightforward.

There are three general ways to calculate co-occurrence within the corpus: Message-level, user-level and thread-level. Message-level correlation measures how often two terms co-occur within the same message. This is the smallest and most powerful sense of correlation, as, if two terms are consistently occurring together within a message then they must be very similar. This method may be somewhat strict, however, as messages themselves have few mappings compared to the number of messages overall, so this may result in low correlation measures with many 0's.

User-level correlation measures how often a single user uses terms over the course of their messages. If a user is using terms together consistently then they must also be similar. This is probably the least effective of the three methods. Users may be interested in multiple subjects, so calculating similarity across different threads within a user would create a correlation between subjects that may not be related.

Thread-level correlation measures how often terms occur together within a thread. If a thread is thought of as the embodiment of the knowledge of the community about a specific subject then co-occurrence within that thread should be representative of the two terms being related to a common idea. This is probably the best of the three methods: it encapsulates message-level correlation as well, while being the best measure of correlation within our idea of a knowledge object.

In order to calculate the similarity between terms we need to introduce additional notation for the threads. Let matrix M be an $n \times h$ term-thread matrix. The rows of matrix M represent the individual semantic terms t_i, and the columns of the matrix represent the threads, h_j. $m_{ij} = 1$ if term t_i is present in the thread, and 0 if the term is absent. Note that this is a binary matrix and not a valued matrix, and can be calculated from the thread-term matrix H, as $M = I(H^T > c)$, or the transpose of the H matrix dichotomized. Since the mappings have values, they need to be dichotomized by comparing them to some threshold value c. An appropriate threshold value will be determined in the experimental stage, but

an initial value of $c = 0$ is sufficient for now.

The co-occurrence similarity between two terms can be calculated using Jaccard's Distance, which is simply the intersection of the term vectors divided by their union (see equation (3.24)).

$$cor_J(t_i, t_j) = \frac{|\overrightarrow{t_i} \cap \overrightarrow{t_j}|}{|\overrightarrow{t_i} \cup \overrightarrow{t_j}|} = \frac{\overrightarrow{t_i}\overrightarrow{t_j}'}{\overrightarrow{t_i}\overrightarrow{t_i}' + \overrightarrow{t_j}\overrightarrow{t_j}' - \overrightarrow{t_i}\overrightarrow{t_j}'} \tag{3.24}$$

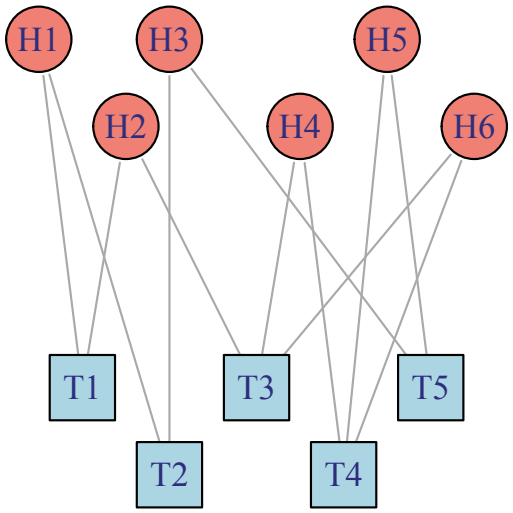
Note that Jaccard Distance is only one option, and others exist. Dice similarity [24] calculates the similarity as two times the size of the union divided by the degree of each term, while Adamic and Adar [4] use the size of the intersection while scaling each component by its own degree.

3.3.7.3 Term Correlation Using Network Analysis

Term co-occurrence is one way to calculate the non-taxonomic relations between terms. If two terms occur together in multiple threads, then there is an inherent relationship between them, but only considers direct relations between terms and not secondary relations. If term A and term B are similar because of their co-occurrence, and term B and term C are similar because of their co-occurrence, what is the similarity between A and C? Consider the example in figure 3.7, which present a simple network and the similarities between the elements.

Looking at the figure, the most similar terms should be terms 3 and 4, since they share 2 threads of their 4 total threads, and then threads 1 and 2, as they share 1 of their 3 total threads. Jaccard distance accurately captures this sense of correlation within 1 step, but what about beyond 1 step? Term 2 has no similarity to term 4, because they share no common threads. They are both very similar to term 3, however, so one would expect them to have something in common. The problem with Jaccard similarity is that there is no sense of *transitivity*, and in large, sparse matrices Jaccard similarity is going to have many more 0's in the matrix than is desired.

For terms that do not share a direct connection, their similarity can be calculated based on the shortest path between them via their neighbours. The similarity between them is the product of the similarity of the nodes on the shortest path. This is an adaptation of the ideas first proposed by Huang and Lai [38]. Let P_{ij} be a path between any two nodes i and



Adjacency Matrix

	H1	H2	H3	H4	H5	H6
T1	1	1	0	0	0	0
T2	1	0	1	0	0	0
T3	0	1	0	1	0	1
T4	0	0	0	1	1	1
T5	0	0	1	0	1	0

Jaccard Similarity

	T1	T2	T3	T4	T5
T1	1.00	0.33	0.25	0.00	0.00
T2	0.33	1.00	0.00	0.00	0.33
T3	0.25	0.00	1.00	0.50	0.00
T4	0.00	0.00	0.50	1.00	0.25
T5	0.00	0.33	0.00	0.25	1.00

Adapted Similarity

	T1	T2	T3	T4	T5
T1	1.00	0.33	0.25	**0.12**	**0.11**
T2	0.33	1.00	**0.08**	**0.08**	0.33
T3	0.25	**0.08**	1.00	0.50	0.12
T4	**0.12**	**0.08**	0.50	1.00	0.25
T5	**0.11**	0.33	**0.12**	0.25	1.00

Figure 3.7: A fictional term-thread network. The Adjacency Matrix presents the network in matrix form, the Jaccard Similarity is calculated using equation (3.24), and the adapted similarity is calculated using equation (3.25). The **bold** terms in the Adapted Similarity are the ones that are effected by the adapted calculations (compared to the Jaccard Similarity)

j: $< (t_i, t_{i1})(t_{i1}, t_{i2}), , ..., (t_{ic}, t_j) >$. Let P' be the set of shortest paths between nodes: there are often multiple paths of the same length, so each of them must be checked. The similarity equation is given in equation (3.25), and an example is given in figure 3.7.

$$cor_{NET}(t_i, t_j) = min_{P \in P'} \left\{ \prod_{t_k, t_l} cor(t_k, t_l) \right\} \tag{3.25}$$

3.3.8 Knowledge-Based Methods for Calculating User Similarity

The Generalized Vector Space Model (GVSM) and the Balanced Genealogy Measure (BGM) provide the best means for calculating user or thread similarity based on medical lexicon

representations. For the GVSM the construction of the correlation matrix is key to ensuring the success of the approach. Semantic and network-based methods can provide the means to estimate the correlation matrix, but neither uniquely captures the relationships between terms within the community. The combined method outlined below provides a means of incorporating semantic and co-occurrence relations between terms into a single correlation matrix to be used in the GVSM model.

The BGM method needs to be adapted to allow for non-leaf mappings, and to deal with issues of homonymity, neither of which were addressed in the original specification. Once those are complete, however, the algorithm needs to be improved. The current approach calculates the similarity between two terms in different subtrees based on the distance between their lowest common ancestor, but this approach suffers for the same reason that the simplest edge-based semantic similarity measures suffer, that not all edges within a semantic network represent the same distance. Resnik [70] used the ratio of the information content between the two terms, and there is potential for that to bring additional, context specific information to the BGM method. Below we outline an information content based improvement to the BGM called the Balanced Information Content Genealogy Model.

3.3.8.1 Combining Semantic and Co-Occurrence Correlation

Of the methods reviewed for calculating Information-Content based semantic similarity, equation (3.23) is our choice for a solution. It adapts the simplicity of Resnik's approach to a $[0, 1]$ scale, allowing it to be comparable to the network-based co-occurrence calculations from equation (3.25).

Previous research into term similarity calculations has focused largely on either semantic similarity or co-occurrence calculations, and has neglected to focus on combining the methods, but the case can be made that both are vital for an accurate representation of term correlation. Working with just the network-based methods ignores the inherent relationships between similar words. It may be the case that very similar semantic terms are not used together because it is implicit that when t_i is used, t_j is also relevant to the conversation. Conversely, the co-occurrence of specific terms with vastly different semantic meanings is quite likely an artifact of the specific clinical scenario. The Information Content captures the significance of individual terms within the corpus, but it does not capture the relatedness of terms within the corpus. Combining the two measures is the optimal solution for

determining the context-specific similarity between terms.

The simplest form of similarity would be to take the maximum of the two calculations. This would have the effect of supplementing the semantic similarity with co-occurrence similarity for terms that are contextually-related but not semantically related. It would maintain the $[0, 1]$ scale, and the resulting correlation matrix would have high scores for terms that are similar and low scores for terms that are dis-similar.

$$cor(t_i, t_j) = max\left[cor_{NET}(t_i, t_j), cor_{SEM}(t_i, t_j)\right] \tag{3.26}$$

This equation can be used to populate the term correlation matrix C, after which calculating user or thread similarity using the GVSM equation (3.18) would be straight-forward.

3.3.8.2 Balanced Information Content Genealogy Measure

Of the two genealogy methods, the BGM is more adaptable to non-leaf mappings, issues of homonymity and extensions to information content. The sections below outline how to adapt the current methods, and then how to extend them.

Adaptations to Non-Leaf Mappings The extension to non-leaf mappings should not drastically alter the BGM method. The structure of the induced subtrees would not change with the inclusion of internally mapped nodes. Looking at the example in figure 3.9, the tree would have the same structure whether or not the internal nodes (the diamonds) were mapped. The only question, therefore, is how to incorporate these internal nodes into the algorithm, and whether this incorporation is necessary. Looking back to the algorithm in section 3.3.5, they use the word *leaf* to iterate and match to all the non-internal nodes in the tree. Because of the structure of their trees, their leaf nodes are their mapped terms. There is no reason, however, that the BGM algorithm could not be changed from working with leaves to working with *mapped terms*. The algorithm can be modified so that it iterates over all mapped terms (internal and leaves) without inducing problems.

Homonymity The issue of homonymity within a medical taxonomy can pose a much larger problem. The challenge with homonyms is that their relationships are not constant across all their instances. A simple example is the term "cell division" in the MeSH tree, as demonstrated in figure 3.8. You can see the potential pitfalls of homonyms and when they

have variable numbers of children. The term *Cytokinesis* is only related to *Cell Division* within the context of *Cell Cycle*, and not when talking about cell division in the context of *Cell Proliferation*. Likewise, *Cell Nucleus Division* is related to *Cell Division* only within the context of *Cell Cycle* and *Genetic Processes*. This means that we cannot represent the MeSH tree as a graph with terms as nodes, because the edges from terms to their parents are not always constant, but rather are context dependent.

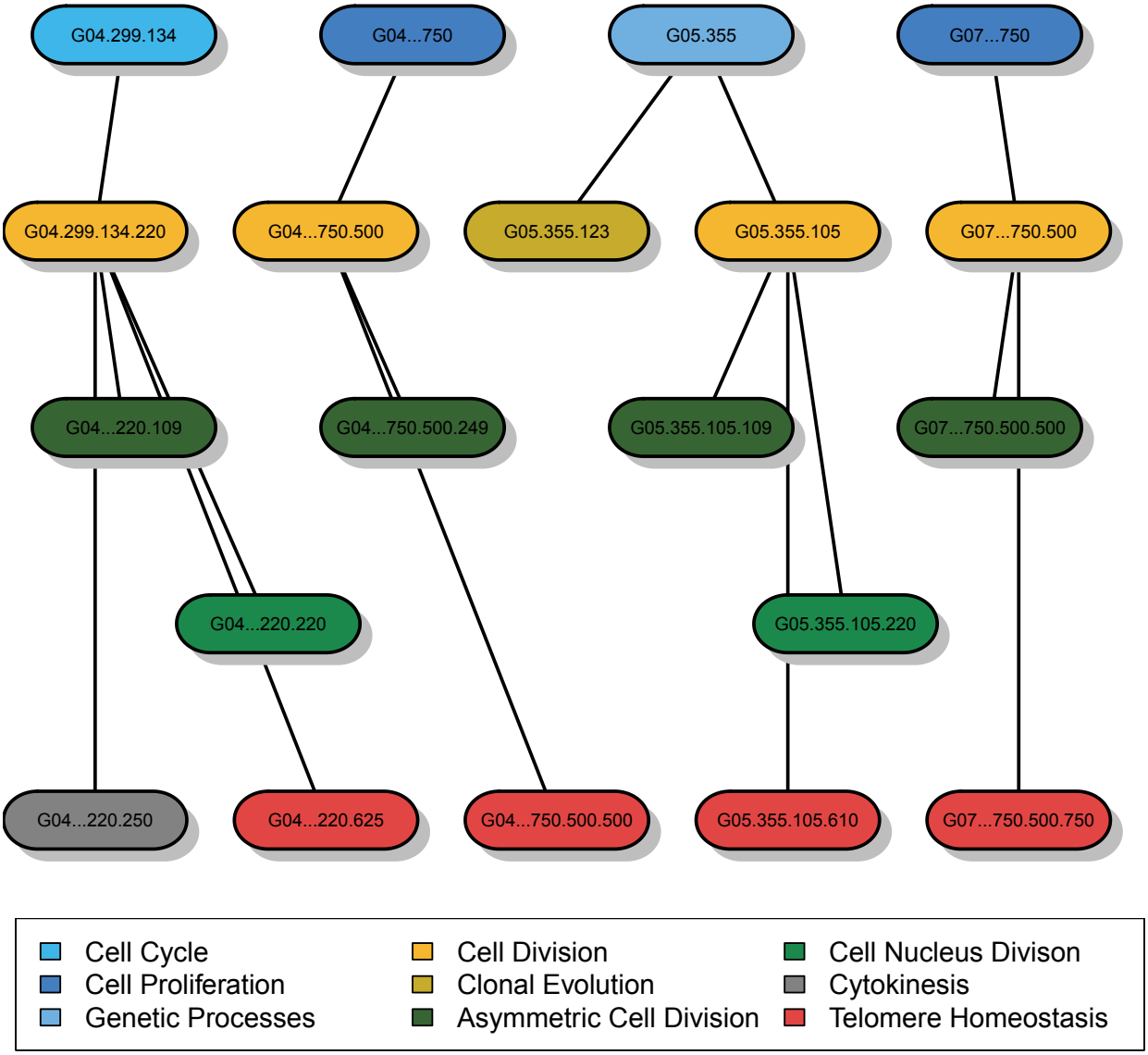


Figure 3.8: An example of the challenges that homonyms can be within MeSH

When we map to a term through a semantic mapping program (like Metamap) we must map to all possible instances of that term: if Metamap returns the term *Cell Division* then we must include all four instances of it in the subtree, as we are unable to determine which one was mapped. When building the subtree up from the mapped terms, however, we should only include terms in context. So if Metamap returns the term *Cytokinesis* then the resultant subtree contains only one reference to *Cell Division*, the one at `G04.299.134.220`. This means that *Cytokinesis* is not related to *Clonal Evolution* even though they are both related to the concept *Cell Division*.

This then raises challenges when trying to calculate the BGM for a user. The BGM requires visiting the *leaves* of the tree in optimal order, such that the final similarity is as high as possible. This is a challenge in the face of homonymity as not all the leaves of the subtree are going to be used in the final calculation. For the cell division example, whichever of the four leaves has the highest *optleafsim* value with the second tree will be the representative leaf for that term. The full tree for user A will be trimmed so that the highest scoring node for each term will be retained and the rest will be dropped.

Modifying the BGM Algorithm Because of the adaptation to allow for internal-node mappings, and the trimming due to homonymity, the optimal-visit component of the original BGM algorithm needs work. It is not clear why an *iterative* punishment strategy is needed for the BGM algorithm. The algorithm states that, for the second and third terms that map to the same leaf node, they be scaled by β^k where k is the number of matches-1, but the result of scaling all three scores by β^k should still be the same: the model should still reward matches between users that do not over-use the same term. The effect of the size of β on the algorithm overall is different, and generally larger values of β will probably be used, but the *optimal visit order* component of the algorithm seems unnecessary.

Even if the iterative punishment component is dropped, there is still a problem with the original algorithm with respect to multiple children. The original paper [29] is somewhat vague on the form of the algorithm, and there seems to be the potential for ties that their instructions do not deal with. Consider figure 3.9 as an example of trying to calculate the similarity between two users.

The major question using the BGM algorithm is, what is the correct mapping for node **G** in User A's tree. It maps to induced node **D** in user B's tree, but that induced node

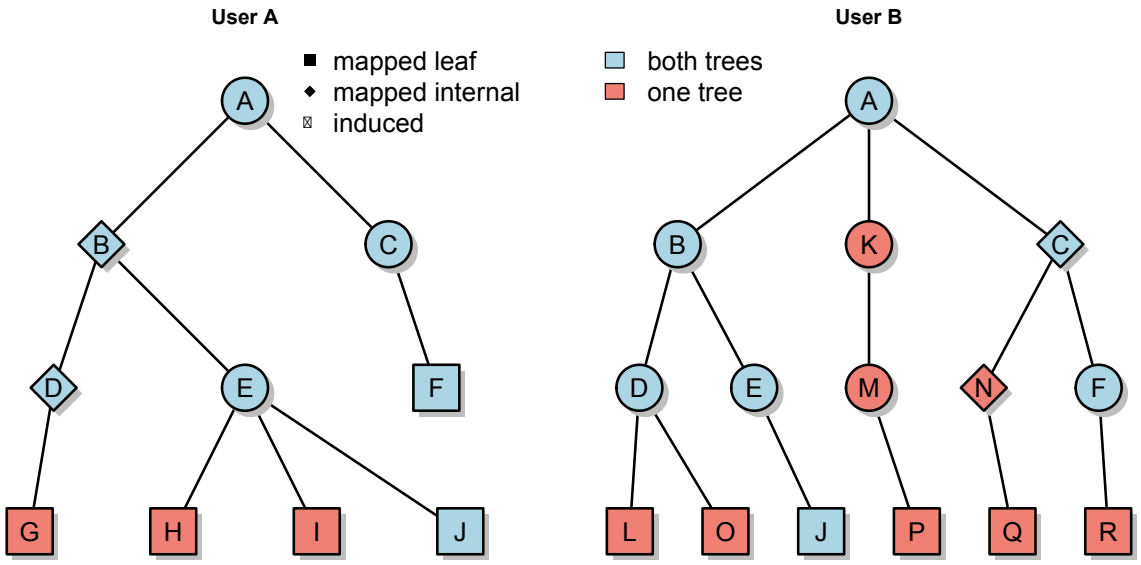


Figure 3.9: Subtrees for the mappings of two fictional users. The squares represent leaf-terms mapped to the user, diamonds represent internal nodes mapped to the user, and circles represent induced terms. Terms in blue appear in both trees, and terms in red appear in only 1 tree

has two potential leaves, **L** and **O**. This is important for counting mappings. The current algorithm counts the *leaves* mapped, but that becomes difficult when internal nodes can also be mapped. The point of counting mappings is to punish user similarities when one term in the second tree provides the mapping for several nodes in the first tree. The algorithm will be adapted such that, when multiple mapped terms contribute to an internal node that is providing a mapping, the count is split between them. The counting algorithm is as follows:

- Assign direct mappings from each node in tree A to a node in tree B
- For all mappings in tree B that are *induced nodes*, add their fractional score to their children. If an internal node in tree B has k children and was mapped to c times, add c/k to each of the children's map counts.
- For each induced node in tree B, set their map score as the sum of their mapped child nodes, scaled by the number of children. From figure 3.9, the score for node C would be $\frac{1}{3}count_N + \frac{1}{3}count_Q + \frac{1}{3}count_R$

Figure 3.10 presents the counting algorithm applied to figure 3.9 for calculating the similarity from user A to user B, and table 3.1 presents the similarity from user B to user A.

STEP 1: COUNT MAPPINGS

Direct Mappings	**user A**	B	D	G	H	I	J	F
	user B	B*	D*	D*	E*	E*	J	C

User B Map Counts	**Node**	C	L	N	J	O	P	Q	R	B*	D*	E*
	count	1			1					1	2	2

STEP 2: PROPAGATE INTERNAL MAPPINGS TO CHILDREN

Direct Mappings	**user A**	B	D	G	H	I	J	F
	user B	B*	D*	D*	E*	E*	J	C
	B-children	L,O,J	L,O	L,O	J	J	-	-

User B Map Counts	**Node**	C	L	N	J	O	P	Q	R
	count	1	$\frac{1}{3}(1)+\frac{1}{2}(2)$ $=\frac{4}{3}$		$1+\frac{1}{1}(2)+\frac{1}{3}1$ $=\frac{10}{3}$	$\frac{1}{3}(1)+\frac{1}{2}(2)$ $=\frac{4}{3}$			
	Node	B*			D*		E*		
	count	$\frac{1}{3}(\frac{4}{3}+\frac{4}{3}+\frac{10}{3})$ $=2$			$\frac{1}{2}(\frac{4}{3}+\frac{4}{3})$ $=\frac{4}{3}$		$\frac{1}{1}(\frac{10}{3})$ $=\frac{10}{3}$		

STEP 3: APPLY MODIFIED COUNTS TO MAPPINGS

Direct Mappings	**user A**	B	D	G	H	I	J	F
	user B	B*	D*	D*	E*	E*	J	C
	counts	2	$\frac{4}{3}$	$\frac{4}{3}$	$\frac{10}{3}$	$\frac{10}{3}$	$\frac{10}{3}$	1

Figure 3.10: The step by step process of calculating mapping counts for the example in figure 3.9. The opposite counts (from user B to user A) are available in table 3.1. The imputed nodes in user B's tree are denoted with * .

The final challenge with the BGM methods are the lack of context sensitivity. The calculations of term similarity are outlined in equation (3.14), but they make the same mistakes that many of the edge-based semantic similarity metrics make by assuming that all edges within a semantic network are created equal. As the semantic-methods demonstrated, the difference between ancestors within a semantic tree can be measured by looking at the ratio of their Information Content (IC). This provides a mechanism for adapting the BGM to the local context, by replacing equation (3.14) with equation (3.27). The final algorithm for the modified method is as follows.

1. For each mapped node (i.e. non-induced node) l_1 in T_1, find the node l_2 in T_2 with the highest Information Content (IC) that is an ancestor of l_1.

2. Increment l_2's match count, the number of times that this node has been used as a match

user B	**user A**[1]	**user A children** [3]	**counts**[5]
L	D	-	$\frac{15}{7}$
O	D	-	$\frac{15}{7}$
J	J	-	$\frac{8}{7}$
P	A*	(All Leaves)	2
Q	C*	F	$\frac{29}{7}$
R	C*	F	$\frac{29}{7}$
N	C*	F	$\frac{29}{7}$
C	C*	F	$\frac{29}{7}$

user A	**counts**[2]	**prop. counts**[4]	
B		$\frac{1}{7}$	$\frac{1}{7}$
D	2	$2 + \frac{1}{7}$	$\frac{15}{7}$
F		$\frac{1}{7} + \frac{1}{1}(4)$	$\frac{29}{7}$
G		$\frac{1}{7}$	$\frac{1}{7}$
H		$\frac{1}{7}$	$\frac{1}{7}$
I		$\frac{1}{7}$	$\frac{1}{7}$
J	1	$1 + \frac{1}{7}$	$\frac{8}{7}$
A*	1	$\frac{1}{7}\left(4 \times \frac{1}{7} + \frac{15}{7} + \frac{29}{7} + \frac{8}{7}\right)$	2
C*	4	$\frac{1}{1}\left(\frac{29}{7}\right)$	$\frac{29}{7}$

Table 3.1: The step by step process of calculating mapping counts for the example in figure 3.9 (The steps are the exponents in the column headers). The imputed nodes in user B's tree are denoted with * .

3. calculate $optleafsim$ for each mapped node in T_1.

$$optleafsim^*_{T_1,T_2}(l_1) = \frac{IC(l_2)}{IC(l_1)} \tag{3.27}$$

4. Scaled each $optleafsim^*$ value by the match-count according to equation (3.15)

$$leafsim_{T_1,T_2}(l_1) = optleafsim^*_{T_1,T_2}(l_1) \times \beta^{matchCount(l_2)-1}$$

5. The equation for similarity between u_1 and u_2 is given below. Let u_{1j} be the tf-idf score for term j from user 1.

$$sim_{BICGM}(u_1, u_2) = \frac{\sum_{l_1 \in u_1} leafsim_{T_1,T_2}(l_1) \times u_{1j}}{\sum_{l_1 \in u_1} u_{1j}} \tag{3.28}$$

This final method will be called the *Balanced Information Content Genealogy Method* (BICGM). It is a novel adaptation of the BGM that cleans up the unclear components of the original algorithm, adapts it to issues of internal mappings and homonyms and adds an Information Content component to provide a more context-specific calculation.

Comparison to GVSM method If we assume that $\beta = 0$ (which Ganesan originally called the Optimistic Genealogy Measure) instead of the BGM, then there are some direct relations between the GVSM using semantic similarity and the BICGM. First look at equation (3.16) reparametrized into vector notation. Let l_i represent the set of *leafsim* values for user i and let i be the identity vector, i.e. a vector of 1's. It's the simplest way to represent the sum a vector.

$$sim_{BGM}(u_1, u_2) = \frac{l_1 u_1}{l_1 i}$$

The entries in l_i are from equation (3.15). When we compare this equation to the semantic correlation between two terms (3.23), we see that it is almost the same value, except that $IC(l_2) = 0$, i.e., it's a one-sided correlation. If we propagate this assumption through the tree we can see that The BICGM method can be thought of as an asymmetric version of the GVSM. This asymmetry has its advantages: a new user may have interests completely in line with that of a content expert, but that expert may, in addition to those interests, have interests in other fields. The BICGM similarity between users changes from "how much do these two users have in common?" to "how much is user i interested in user j?" Both questions are of value, but they report slightly different things.

3.3.9 Content-Based Clusters

Content-based clustering of the users and threads can tell us what the most popular subjects are within the community, where potential subgroups may arise, and provides a second user clustering based on content instead of connections. In terms of threads, we want to know if and how the threads form into subgroups. Threads are the essential knowledge objects within the community; they are the unit within which the knowledge shared by the members is organized around a specific topic. Clustering the threads is a process of detecting the knowledge-clusters within the community and identifying potential subgroups of interest within the larger knowledge base. The general structure of the clustering process and the applied structure for this project is given in figure 3.11. The purpose of the clustering is to

identify knowledge-based and informative clusters. Whatever specific thread representation, clustering algorithm, cluster evaluation and cluster representation is used, the final set of clusters must provide meaningful information to the end users, in terms of a subset of threads that are semantically related and provide a useful subset of the overall corpus.

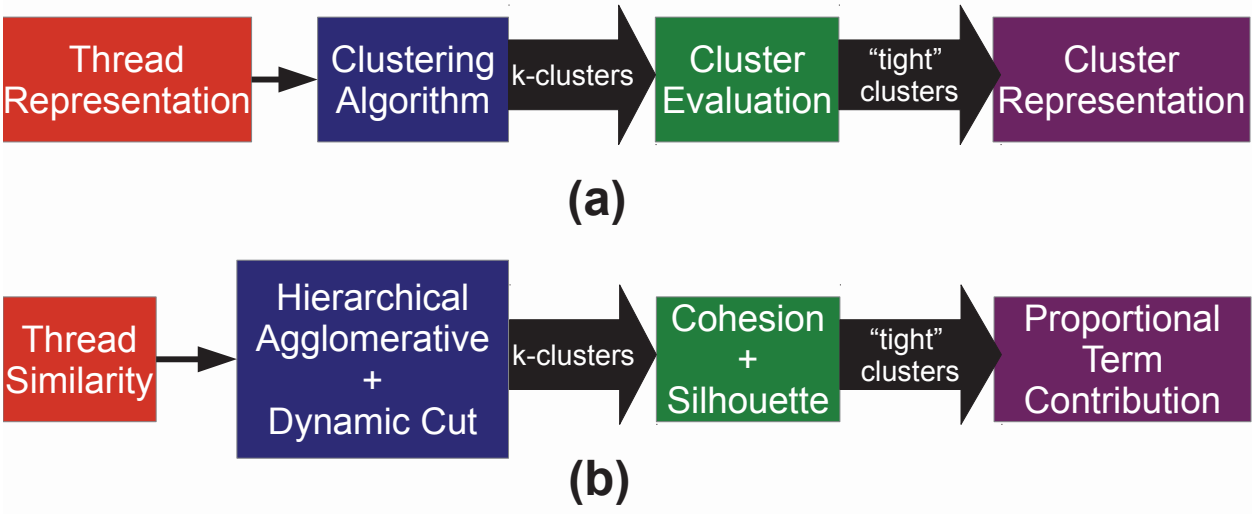


Figure 3.11: (a) the general structure of the clustering algorithm. (b) the specific algorithm decisions made

With the symmetric user or thread similarity matrices generated using the GVSM equation we have a natural "distance" measure to determine the distance between two users or threads. The clustering methods in this section assume that there exists within the data an exclusive clustering of threads, and attempts to find an optimal partitioning using hierarchical agglomerative clustering along with a dynamic tree splitting method [48]. Once we have found a suitable set of k clusters we will evaluate the clusters using silhouette coefficients to assess the quality of the clusters. Finally, for the clusters deemed sufficiently separate from the rest of the corpus we will investigate the potential semantic terms that define them by looking at proportional term contributions.

3.3.9.1 Clustering Based on Content Similarities

Applying clustering methods to the GVSM similarities provides a content based clustering that ties similar users together based on their shared content in an iterative approach. This type of hierarchical clustering differs fundamentally from other clustering methods (k-means or information maximization, for example) in that the clusters are not exclusive. Hierarchical

clustering defines a series of successive clusters (nested clusters) based on the distance between observations. The process for creating these clusters can be either top-down (divisive clustering) or bottom-up (agglomerative clustering). One advantage of hierarchical clustering is that it is based on the distances between observations and not the raw observations themselves. Any metric that can produce a distance matrix that reports the (symmetric) distances between all pairs of observations can be used for hierarchical clustering. We can use the GVSM similarities between users or threads as a distance metric.

In the bottom-up approach each observation is assigned its own cluster. At each step in the algorithm the two most similar clusters are merged to form a single cluster, and then the process is repeated. The method for calculating the distance between clusters significantly effects the resulting structure. Single-link clustering calculates the distance between two clusters as the shortest of all the distances between elements of each cluster. Complete-link clustering calculates the distance between two clusters as the longest of all the distances between elements of each cluster. Average-link clustering takes the mean of all distances between elements. Ward's method [89], which uses a minimum variance function that is similar to an error sum of squares approach.

Top-down clustering starts by assigning each observations to a single cluster, then successively splitting the clusters until atomic clusters (i.e clusters with only 1 observation) are obtained. The decisions that define the splitting are which cluster to split, and what splitting rule to use. The cluster to split is normally chosen based on some cluster evaluation metric (several of which are outlined in section 3.3.9.3), and the cluster with the worst metric is split. Once a cluster is selected, the splitting should be done to try and improve the metric as much as possible. Bisecting k-means is an example of a method that does this, by performing a k-means partitioning with $k = 2$ to split the cluster into 2 subclusters. With bisecting k-means the process is typically stopped once k clusters are created, so setting $k = m$ for our thread matrix would result in atomic clusters.

Hybrid Hierarchical Clustering [13] presents a method that combines top-down and bottom-up approaches to create optimal clusters. Top-down approaches are known to be better when the number of desired clusters is small, and bottom-up when the desired number is large, so the hybrid method attempts to combine them to find an optimal solution.

Lin and colleagues used a bisecting k-means approach to cluster citations from MEDLINE using MeSH terms [53], which was also leveraged by Yoo [96]. The AGNES algorithm [43] was

used to cluster MEDLINE citations by Struble [81] in a traditional hierarchical agglomerative clustering approach.

We will use an agglomerative clustering algorithm (specifically the AGNES algorithm from the R library `cluster` [58]) to calculate clusters within the thread and user similarity matrices. This method results in a series of clusters (a dendrogram) that must be split in order to find an exclusive clustering of the users or threads.

3.3.9.2 Defining Clusters: Splitting Dendrograms

A dendrogram is a simple tree diagram used to demonstrate the hierarchical clustering process, and splitting a dendrogram at certain points produces exclusive clusters. Figure 4.32 presents a dendrogram for the thread clustering from the PPML. The most common method of splitting a dendrogram is a static cut, where a height in the tree is chosen and the branches of the dendrogram are split at that level to produce k clusters. Hierarchical clustering is known to be sub-optimal at certain levels, however, so the clusters may have spurious memberships, especially near the cutpoint. Langfelder [48] proposed a novel algorithm called the Dynamic Hybrid Algorithm, which uses a 2-step, bottom-up approach to splitting a dendrgram into clusters.

The two step process is as follows:

1. Branches of the dendrogram that meet the following specified criteria are made into clusters

    - The cluster must meet a minimum size, N_0
    - Objects must not be too far from the cluster. Within the dendrogram this is controlled by a maximum height for joining a cluster, h_{max}
    - Clusters should be separated from surrounding clusters by a minimum distance, g_{min}
    - The "core" of the cluster (and core is defined algorithmically) must be sufficiently dense, defined by a maximum distance (equivalent to a minimum average similarity), d_{max}

2. The un-matched objects are added to the nearest cluster.

In controlling the parameters for cutting the tree the authors of the algorithm offer a variable called `deepSplit` that sets the g_{min} and d_{max} values according to pre-specified values. The values are presented in table 3.2, and are the ones used for this project. Other values were also investigated, but none yielded superior clusters, so the default 5 levels were used. The second decision made was to ignore the second step in the hybrid process, where the unclustered points are added to the closest cluster. We are most interested in finding unique clusters of threads, therefore it is not necessary that every thread be assigned to a cluster, especially if it dilutes the quality of the cluster. With the minimum cluster size at 10 threads this is not a huge issue, as most threads are assigned to a cluster.

deepSplit Value	d_{max}	g_{min}
DS_0	0.64	0.27
DS_1	0.73	0.2025
DS_2	0.82	0.135
DS_3	0.91	0.0675
DS_4	0.95	0.0375

Table 3.2: The different default values for building the hybrid cuts. The values are given as fractions of the height of the tree in order to apply consistent factors across trees of varying heights.

3.3.9.3 Evaluating Cluster Assignments

Regardless of the clustering method used, the result is a set of k distinct clusters. Evaluating the validity of those clusters is a crucial step in the clustering process. The clustering algorithms assume that a partitioning of the data into clusters exists and attempts to find it, but evaluations of the clustering provide the insight into whether these clusters are real separations of the data.

The Silhouette Coefficient is a method for evaluating the clustering at an individual level. Let a be the average similarity between an object and all the other elements of the cluster. For each other cluster, calculate the average similarity to that cluster, and let b be the largest of those values (i.e., b is the average similarity to the closest cluster). The silhouette coefficient is defined in equation 3.29.

$$s_i = \frac{a - b}{max(a, b)} \tag{3.29}$$

The silhouette coefficient ranges from [-1,1], where high values are better, and values < 0 indicate that the object is in the wrong cluster. Figure 4.35 presents the silhouette coefficient for clustering on the PPML.

Finally, Agglomerative Coefficient is an evaluation of hierarchical clustering overall, i.e. an evaluation of the dendrogram. For each node in the structure d_i is the *dissimilarity* between the node and the cluster it was first merged with, divided by the dissimilarity of the final merger. The agglomerative coefficient is the average of $1 - d_i$ across all the terms. In general high agglomerative coefficients indicate a better cluster, but the coefficient is highly susceptible to network size, and therefore should only be used for comparisons between different methods. We will use it to decided which of the four cluster distance metrics to use.

3.3.9.4 Evaluating the Content of Clusters

Clusters are useful only if the identified clusters can provide useful feedback to the community at large. After the clusters are identified and evaluated the mechanism for their clustering needs to be determined. This is an attempt to determine what the identified threads are clustering around.

Recall that, in the thread-term matrix H threads are represented by their component semantic terms, which are TF-IDF scaled. The purpose of TF-IDF scaling is to reduce the effect of common terms biasing the representation while still maintain some measure of contribution to the corpus. To evaluate the clusters we are going to looking at the *term contribution* for each term in each cluster. For each cluster we will calculate the average TF-IDF scaled score for each term. Clusters that are formed around terms should have high TF-IDF values, relative to low values in other clusters. We will study the highest term contribution values at the end of the next section to see how they can be used to define the detected clusters.

3.3.10 Summary

Exploring the content of the communications within an online knowledge-based community is imperative for properly understanding and managing the KT activities within it. Knowledge maps will provide general summaries of the knowledge being shared within the community, which can be leveraged to improve the community overall.

We have explored methods for exploiting the semantic representations of messages to calculate user and thread similarities within the online community. We have demonstrated two novel approaches to calculating these similarities: the BICGM for calculating asymmetric similarities based on semantic relationships between a user's or thread's semantic terms, and a GVSM approach that calculates symmetric similarities based on the semantic and co-occurrence correlations between semantic terms. We then explored how these similarities could be used to detect user or thread clusters within the community.

The original specifications for the BGM measure [29] were in need of improvement. We have adapted it to deal with non-leaf mappings and helped clear up the un-addressed issue of multiple child nodes for internal mappings. We have also explored the idea of incorporating the concept of information content into the calculation, in order to better measure the inherent relationships between terms within a medical lexicon. Other studies [51, 70] have demonstrated that the effect of measuring distance within a lexicon based on IC instead of depth-based or edge-based methods improves the similarity calculations, so we posit that using IC in the BGM (thereby creating the BICGM method) will only improve the similarities between users.

The GVSM requires a term-correlation matrix that measures the similarity between the terms used to represent users. The semantic correlation measures [51,70] provide an optimal method for measuring correlation between terms within a medical lexicon, but we believe that they do not capture the context-specific relationships between concepts that co-occurrence measures do. Combining the two provides a way to supplement the semantic correlations when there are relationships between terms (such as *Pediatrics* and *Pain* within the PPML) that do not have explicit semantic relationships.

The next section will evaluate the modifications to both similarity calculations to evaluate their overall effect on detecting user and thread similarity and clusters.

3.4 Comparing Content Similarity Methods

The previous section presented several approaches to similarity calculations that will need to be tested in order to determine the best approach. Message-level normalization is required in order to balance user representation across their messages, and TF-IDF weighting is the most straightforward and effective term-weighting method. Beyond those decisions there are four potential methods available for calculating user/thread similarity.

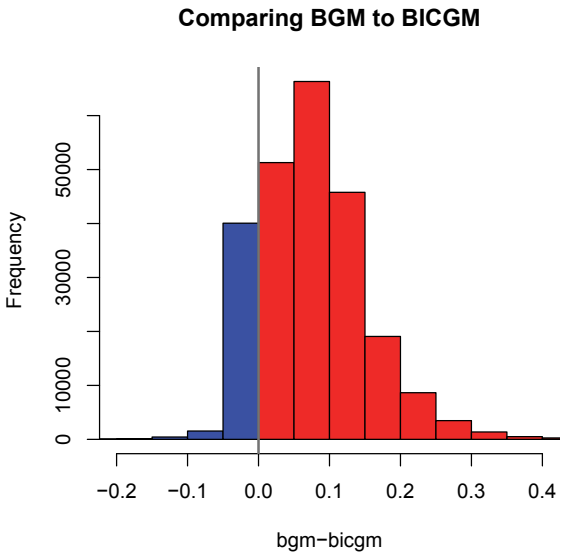
1. Balanced Genealogy Measure
2. BICGM
3. General Vector Space Model using cor_{SEM} from equation (3.23)
4. GVSM using combined correlation from equation (3.26), i.e., supplementing the cor_{SEM} method with the co-occurrence correlations.

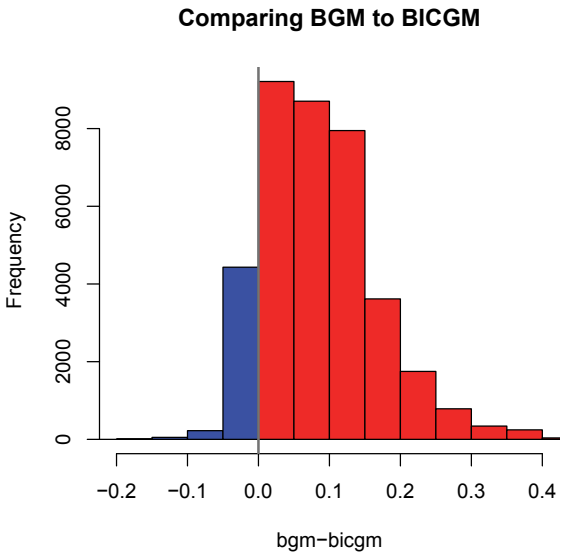
This section will compare the four different methods. The first step will be comparing the two genealogy methods (1 vs. 2), then the two GVSM approaches (3 vs. 4), and finally comparing the BICGM to the GVSM approach (2 vs. 4).

3.4.1 BGM vs. BICGM

Figure 3.12 presents the distribution of the pairwise differences between the BGM and BICGM methods for computing similarity for the PPML and SURGINET data. The BGM method tends to report higher correlations than the BICGM method, but that is not that significant of a finding, as it merely suggests that the methods function on different scales.

Figure 3.13 presents the distribution of the measurements themselves, along with their square-root transformations. After looking at the figures, I decided to use a square-root transformation of both similarity measures, in order to normalize them and put them on a more manageable scale. With the raw data there is crowding at the low end of the similarity, causing challenges in interpretation.





PPML Data			SURGINET Data		
BGM-BICGM	**Count**	**Proportion**	**BGM-BICGM**	**Count**	**Proportion**
< -0.2	176	0.0009	< -0.2	7	0.0002
(-0.2,-0.15]	93	0.0005	(-0.2,-0.15]	10	0.0003
(-0.15,-0.1]	331	0.0017	(-0.15,-0.1]	51	0.0014
(-0.1,-0.05]	924	0.0048	(-0.1,-0.05]	224	0.0060
(-0.05,0]	14592	0.0764	(-0.05,0]	4432	0.1178
(0,0.05]	31620	0.1656	(0,0.05]	9212	0.2448
(0.05,0.1]	49257	0.2579	(0.05,0.1]	8705	0.2313
(0.1,0.15]	47978	0.2512	(0.1,0.15]	7949	0.2112
(0.15,0.2]	27031	0.1415	(0.15,0.2]	3616	0.0961
>0.2	18967	0.0993	>0.2	3430	0.0911

Figure 3.12: The distribution of the pairwise differences between the BGM and BICGM similarities using the PPML data (left) and the SURGINET data (right)

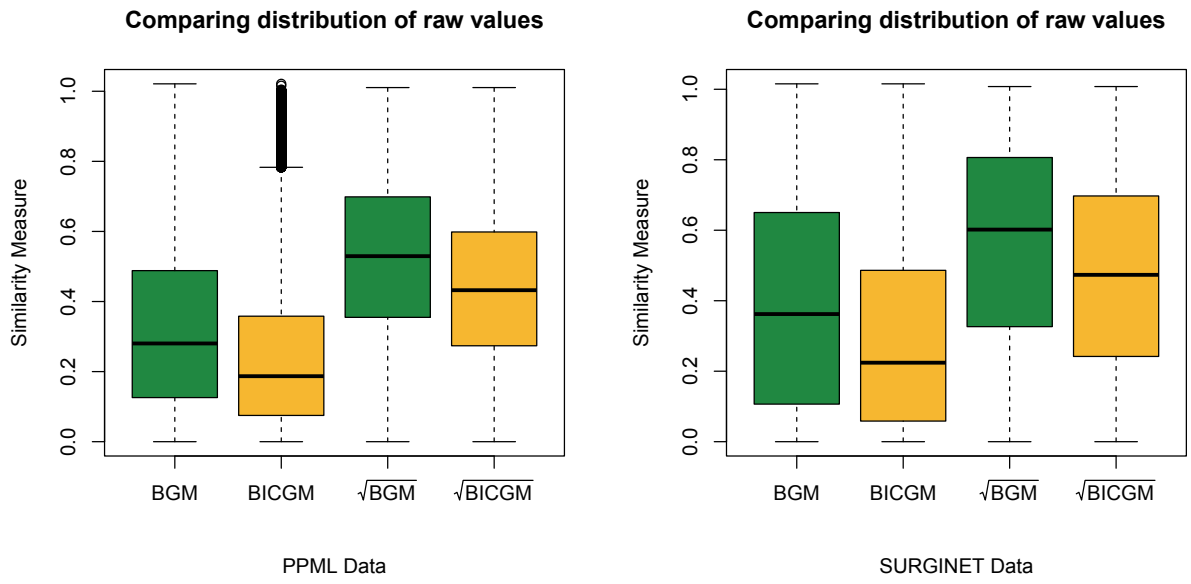


Figure 3.13: Boxplots of the BGM and BICGM values, along with their square-root transformations, for the PPML data (left) and the SURGINET data (right)

3.4.1.1 Differences in Individuals

We are going to investigate some individual pairings from the PPML in order to understand how the BGM and BICGM methods differ. The users selected are S0602 and S0605. User S0602 has a BGM similarity of 0.79 and a BICGM of 0.85 to user S0605. User S0602 participated in one thread, a discussion about the management of pain due to gum recession. His MeSH terms and corresponding mapping scores are listed in table 3.3.

Message Text	MeSH Term	Mapping Score
a really old fashioned remedy - oil of cloves?	Clove Oil	800
Gum Recession	Gingiva	694
the pain	Pain	2000
originating from the exposed dental roots	Tooth Root	553
of gum recession.	Gingiva	1992
analgesics for the gums,	Analgesics	770
gum inflammation.	Gingivitis	1000

Table 3.3: The message mappings for user S0602 from the PPML

User S0605 is one of the more active users in the community. He has communicated 19 times on 17 threads, including one thread entitled "Gum Recession From the Dental Literature". His mappings from that thread are listed in table 3.4.

Message Text	MeSH Term	Mapping Score
with oral mucosal pain	Mucous Membrane	660
due to cancer	Neoplasms	1694
cancer therapy.	Therapeutics	861
Oral topical doxepin rinse	Mouthwashes	902
Gum Recession From the Dental Literature	Gingiva	586
Oral topical doxepin rinse	Doxepin	645
analgesic effect in patients	Analgesics	604
analgesic effect in patients	Patients	604
with oral mucosal pain	Facial Pain	740

Table 3.4: The relevant message mappings for user S0605 from the PPML

Of user S0602's 6 terms, 3 are perfect matches to S0605, MeSH terms *Gingiva, Pain* and *Analgesics*. These mappings receive a matching score of 1 regardless of whether the BGM or BICGM method is used. User S0602's three other terms map through common ancestors. See table 3.5 for how the mapping worked between the two users.

S0602			S0605			Sim Scores	
IC	Location	Term	Term	Location	IC	BICGM	BGM
5.544	A14.549.167.646.480	Gingiva	Gingiva	A14.549.167.646.480	5.544	1.000	1.000
1.364	D27.505...014	Analgesics	Analgesics	D27.505...014	1.364	1.000	1.000
0.412	G11.561.600.810.444	Pain	Pain	G11.561.600.810.444	0.412	1.000	1.000
5.033	A14.549.167.900.750	Tooth Root	Dentition	A14.549.167	4.697	0.933	0.667
6.642	D20.215.784.750.186	Clove Oil	Complex Mix.	D20	3.347	0.504	0.333
6.642	C07.465.714.258.480	Gingivitis	Stomato. Dis.	C07	2.881	0.434	0.333

Table 3.5: A sample of the PPML mappings from user S0605 to user S0602. The Sim Scores in the last two columns are the two different approaches to calculating *leafsim* values, IC for the BICGM method, and depth ratio for the BGM method.

For all three imperfect matches, the IC similarity is larger than the depth similarity. This is because, though the matches are not close to each other within the MeSH tree (3, 5 and 5 steps respectively), within the context of pediatric pain they are rather similar ideas relative to the overall content of the community. The MeSH term *Tooth Root* is closer to the term *Dentition* than 3/5 steps, because within the context of *Dentition* most threads contained the term *Tooth Root*. Table 3.6 contains the aggregated mappings and thread counts for all children of the term *Dentition* within the MeSH tree.

Tree Location	Mesh Term	Thread Count
A14.549.167.646.267	Dental Cementum	1
A14.549.167.646.480	Gingiva	3
A14.549.167.860	Tooth	1
A14.549.167.900.250	Dental Cementum	1
A14.549.167.900.750	Tooth Root	5

Table 3.6: All the child mappings from the MeSH Term "Dentition"

This example illustrates the potential to use information content to shorten the edge length between similar terms that may be several steps apart within the MeSH tree. This example is of a situation where the BICGM similarity is greater than the BGM similarity, however, and as we saw in figure 3.12 this is not the norm. In the majority of cases the BGM similarity was higher than the BICGM similarity, suggesting that the IC method is increasing edge length more often than decreasing.

Users S0875 and S0636 are examples of where IC is lengthening the edges within the MeSH tree. These two users mapped to only 1 term on the mailing list. In reality they are not good candidates for indepth study (short, one-mapping messages, no further participation), but

their mappings demonstrate something important when considering BICGM vs BGM. They both mapped to the term *Chronic Pain* at location C23.888.592.612.274, which is a child of the term *Pain* at location C23.888.592.612. These two terms are only 1 step apart, so their BGM similarity would be 5/6, but the IC of the two terms are 2.11 and 0.37 for the child and parent respectively. The term *Chronic Pain* is a specific idea, whereas the term *Pain*, especially within the context of the PPML, is a very general and common term. If you were interested in chronic pain within the mailing list then looking at all users who communicated around the concept of *Pain* would return far more users than you are personally interested in. Users S0875 and S0636 have a BGM similarity of 0.913 with 239 users on the mailing list (all users with a direct mapping to Pain) and a BICGM similarity of 0.415 for the same users.

A more meaningful example of the effect of IC lengthening edges is for user S0872. For multiple users S0872 has a BGM similarity of 0.671 and a BICGM similarity of 0.337, and for all those users it is because of the same sub-tree overlap, listed in table 3.7.

S0872			**Target**			Sim Scores	
IC	Location	Term	Term	Location	IC	BICGM	BGM
0.412	G11...444	Pain	Pain	G11...444	0.331	1.000	1.000
2.617	D02....367.652	Ketamine	Chemicals and Drugs	D	0.412	0.186	0.125
4.445	A05.810.890	Urinary Bladder	Anatomy	A	0.776	0.175	0.250
3.087	G01.750.770.776	Sound	Phenomena and Processes	G	0.486	0.078	0.200
2.325	F04.096.628	Psychology	Psychiatry and Psychology	F	0.150	0.065	0.250
6.642	C23.888.592.900	Urinary Bladder, Neurologic	Neurologic Manifestations	C23.888.592	0.242	0.050	0.800

Table 3.7: The similarity mappings for S0872 to a number of target users on the PPML

These mappings are a good example of how far terms near the top of the tree can be from their roots conceptually. The term *Psychology* is 4 steps from its root, *Psychiatry and Psychology*, so it's BGM similarity is 1/4, but the tree rooted at *Psychiatry and Psychology* (node F) in the PPML is quite large, therefore the IC of the root is very low. This means that the distance between the term *Psychiatry* and its root is almost 4 times as far based on IC as based on node depth. Both similarity measures are designed to punish mappings that go through multiple ancestors, but also to punish mappings near the top of the tree, but the penalty of mapping near the top of the tree does not seem to be large enough within the

BGM model to encapsulate how different some the root nodes in the MeSH tree are from their children.

Ultimately I think that low similarity scores are appropriate for this user. Of S0872's 6 terms, four of them mapped through root nodes, and the two that did not map through a root are mapped through very common terms within the tree, *Pain* and *Neurologic Manifestations*.

3.4.1.2 Ranking User Similarities

We can find examples of how BGM and BICGM are different in individuals, and have demonstrated the potential improvement in similarities based on IC methods, but as we demonstrated in figures 3.12 and 3.13 there is a difference in the distributions between BGM and BICGM values, so value-based comparisons may not be the most appropriate. If the ultimate goal of similarity ranking is to find similar users, then that should be the ultimate comparison between the methods. The magnitude of the similarity is important, but users are going to be more interested in their top five most similar users than they are in the size of that similarity.

To this end we investigated how the two methods differed in their similarity rankings for individual users. For each user we found their top 5 most similar users based on both the BICGM and BGM methods. We then checked the top 5 and top 10 of the other method to investigate the overlap. Table 3.8 contains the overlap between the two methods for the PPML data and table 3.9 for the SURGINET data.

For the PPML, The table demonstrates that, though the methods are reporting fairly different numbers (as was evident in figure 3.12), the relative ranking of the two methods is similar. For 93% of users the top 5 most similar users to them according to BGM were in the top 10 of the BICGM method, and of those top 5 49% were the same top 5 for the BICGM, and another 39% had 4 of 5 the same.

The SURGINET results again reflect the similarities between the two methods, with significant overlap in the way they rank users. The BGM and BICGM methods are clearly different, but ultimately there is little difference in their final effect on similarity rankings.

Clearly there is a difference between the two methods. Figures 3.12 and 3.13 demonstrate that the two methods are reporting different numbers, and the investigation into individual pairs from the PPML demonstrate that the effect of using IC, of shortening some edges and

		BGM					
		0	1	2	3	4	5
Overlap with BICGM top 10	n	0	0	2	4	29	455
	proportion	0.000	0.000	0.004	0.008	0.059	0.929
Overlap with BICGM top 5	n	0	1	7	50	191	241
	proportion	0.000	0.002	0.014	0.102	0.390	0.492

		BICGM					
		0	1	2	3	4	5
Overlap with BGM top 10	n	0	0	2	4	37	447
	proportion	0.000	0.000	0.004	0.008	0.076	0.912
Overlap with BICGM top 5	n	0	1	7	50	191	241
	proportion	0.000	0.002	0.014	0.102	0.390	0.492

Table 3.8: Measuring the overlap between the most similar users per user, for both BGM and BICGM in the PPML.

lengthening others, has the potential to improve the similarity measures, at least conceptually. Ultimately, however, the two methods are reporting very similar relative rankings, in that, regardless of whether you look at your most similar users based on BGM or BICGM methods, you will get very similar rankings of users you may be most interested in.

		BGM					
		0	1	2	3	4	5
Overlap with BICGM top 10	n	0	0	0	1	7	186
	proportion	0.000	0.000	0.000	0.005	0.036	0.959
Overlap with BICGM top 5	n	0	0	1	12	73	108
	proportion	0.000	0.000	0.005	0.062	0.376	0.557

		BICGM					
		0	1	2	3	4	5
Overlap with BGM top 10	n	0	0	0	2	7	185
	proportion	0.000	0.000	0.000	0.010	0.036	0.954
Overlap with BGM top 5	n	0	0	1	12	73	108
	proportion	0.000	0.000	0.005	0.062	0.376	0.557

Table 3.9: Comparing the overlap between the BGM and BICGM methods for finding the most similar users on SURGINET

3.4.2 GVSM with Semantic Versus Network Correlations

Within the GVSM method the major challenge is how to measure term correlation. We have investigated this process in section 3.3.6, and determined that, though the semantic similarity seems the most appropriate, supplementing it with co-occurrence measures when no semantic similarity exists may help improve the correlation measures.

This section will begin by directly investigating the correlation values for individual terms, then will move on to applying both measures within a GVSM to calculate user similarity.

3.4.2.1 Semantic Versus Co-Occurrence Correlations

The distributions of the semantic and co-occurrence correlations are presented in figure 3.14 for the PPML data and figure 3.15 for the SURGINET data. It is clear that the majority of both correlation matrices are quite low. For the PPML data 96% of the semantic correlations resulted in a correlation of 0, while 88% of the pairs in the co-occurrence calculations had a correlation < 0.05, and similar patterns hold for the SURGINET data.

What is more interesting is the terms that scored highly in the co-occurrence relations and low in the semantic correlations: these are the terms that do not have inherent taxonomic relations to one another, but within the context of pediatric pain they are related via their discussion threads. Figure 3.16 contains the distribution of the differences between

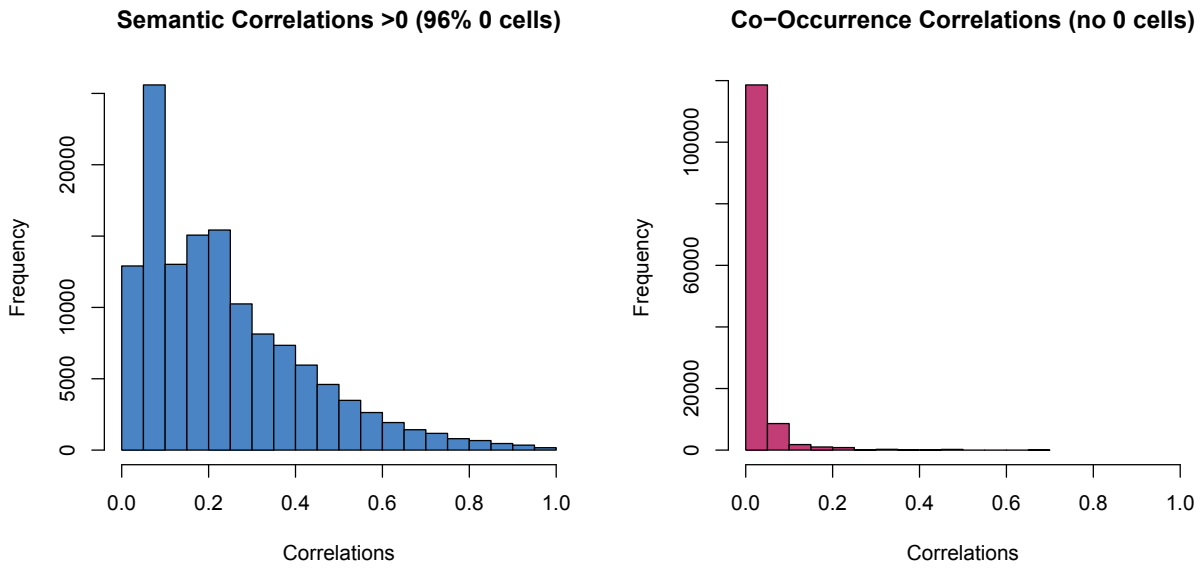


Figure 3.14: The distribution of the semantic and co-occurrence correlations on the PPML

correlation values in the two matrices for the PPML and SURGINET data, and demonstrate similar patterns. For the PPML data, table 3.10 contains the pairs that scored lowest in figure 3.16 from the PPML data, i.e, the pairs that had a high co-occurrence correlation and a low semantic correlation.

Table 3.10 presents some of the terms from the PPML that have the highest co-occurrence correlation. These correlations highlight the importance of context in the determination of term correlations. Low semantic and high co-occurrence correlation suggests that certain correlations exist within a specific context. For example, the high correlations between the terms Pain, and both Pediatrics and Child are indicative of the context in which these messages are being communicated (within a pediatric pain mailing list), but these terms do not have any inherent relationship within the MeSH taxonomy.

Some of the correlations in table 3.10 represent relationships within MeSH that are not specifically denoted in the semantic tree. Cystis, for example, has a "scope note" connecting to Urinary Bladder, and likewise Urinary Bladder has an annotation for Cystis, but they do not share an hierarchical relationship within the MeSH tree.

Mouthwash and Mucositis are an interesting pairing, and demonstrate the true value of these context-specific correlations. Mucositis is an inflammation of the mucosa, and often arises in the mouth, so it is possible that mouthwash could be used to treat it, but

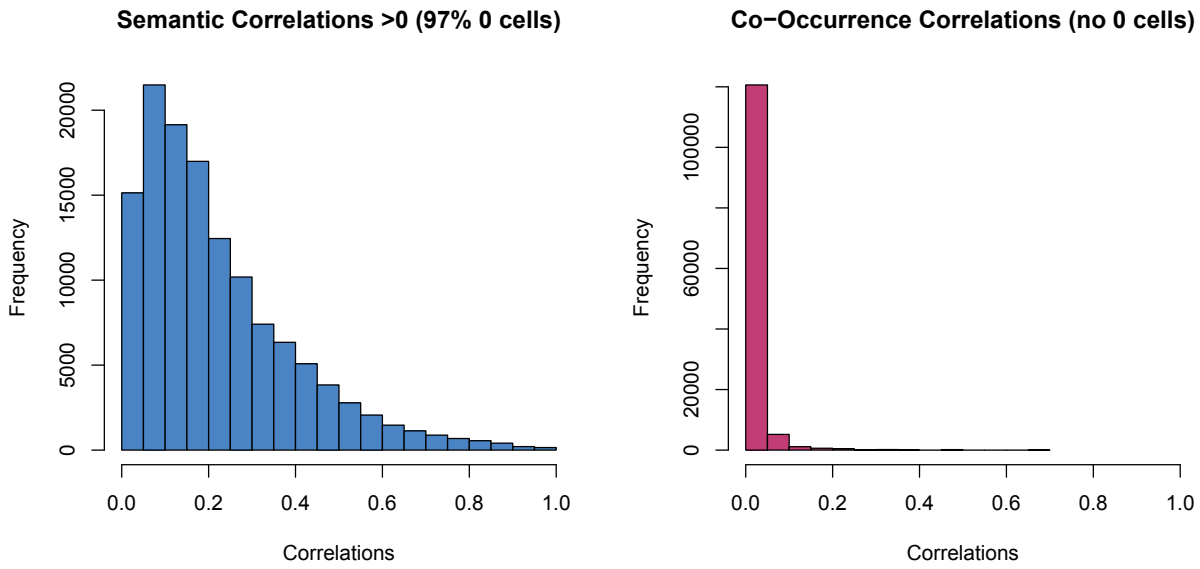


Figure 3.15: The distributions of the semantic and co-occurrence correlations on SURGINET

the relationship is not so strong that it warrants a semantic relationship within the MeSH tree. Within the communications on the PPML, however, the terms have largely appeared together. Within the entire healthcare community the MeSH term mouthwash can be related to a myriad of subjects, but within the context of pediatric pain mouthwash is more likely to be discussed in terms of management of specific diseases such as mucositis. This is a great example of how the context in which the message is communicated is a vital component of the correlation calculations. Within SURGINET the terms have no relationship, but within the context of pediatric pain they are related.

We argue that the co-occurrence correlations represent important relationships between terms that are either not present in MeSH or not explicitly annotated. Incorporating them into the correlation calculations provides additional insight into the inherent experiential knowledge within the community, which is an important facet of KT. In turn, these relationships can contribute to the overall user similarity calculations.

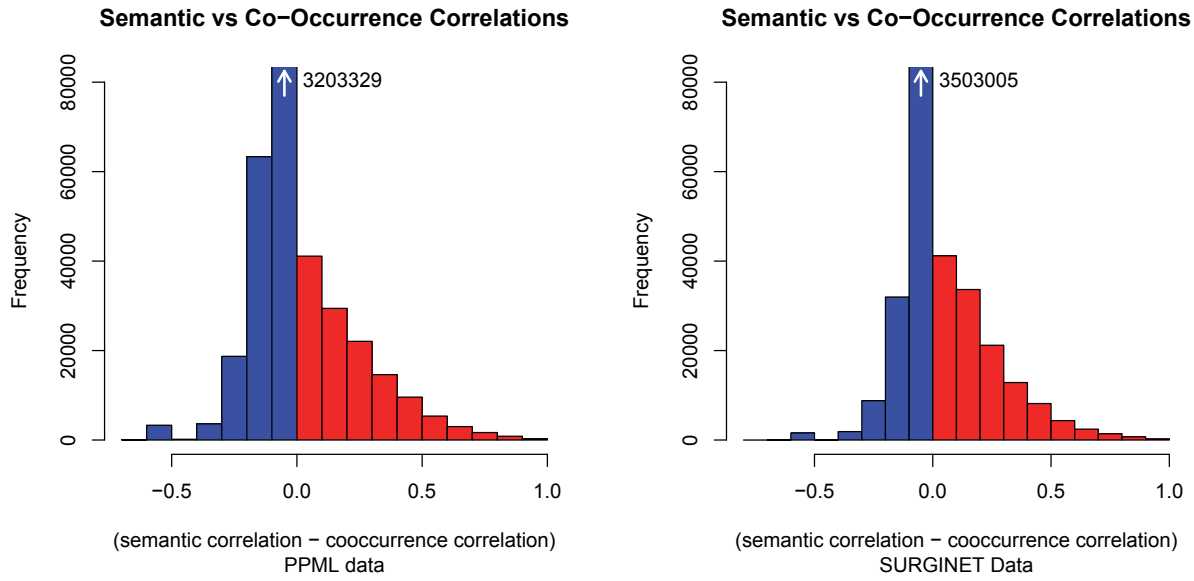


Figure 3.16: Differences between semantic and co-occurrence correlations for the PPML (left) and SURGINET (right) data. The highest bar is significantly truncated in each figure, as the vast majority of pairs have no semantic correlation and a very low co-occurrence correlation

Term 1	Term 2	Sem. Cor	Co. Cor
Pain (3360)	Pediatrics (1443)	0.000	0.516
Pain (3360)	Child (1252)	0.000	0.402
Cystitis (42)	Urinary Bladder (78)	0.000	0.455
Mucositis (34)	Mouthwashes (25)	0.000	0.429
Mucous Membrane (15)	Mouthwashes (25)	0.000	0.500
Penis (13)	Circumcision, Male (26)	0.000	0.500
Baclofen (20)	Muscle Spasticity (14)	0.000	0.455
Pancreatitis (15)	Celiac Plexus (17)	0.000	0.500
Pancreatitis, Chronic (13)	Celiac Plexus (17)	0.000	0.500
Urinary Retention (25)	Urinary Tract (3)	0.000	0.500
Scrotum (12)	Inferior Colliculi (15)	0.000	0.667
Gingiva (24)	Paint (3)	0.000	0.500
Pancreas (7)	Celiac Plexus (17)	0.000	0.600
Immunity (2)	Whooping Cough (21)	0.000	0.500

Table 3.10: A sample of the terms with the highest co-occurrence and lowest semantic correlation on the PPML

3.4.2.2 User Similarity

We calculated user similarity using two Generalized Vector Space Models, one using the combined correlation from equation (3.26), the other using only semantic correlations from equation (3.23). The distribution of user similarity measures, along with comparisons between them, are presented in figures 3.17 and 3.18 for the PPML and SURGINET data respectively.

The combined correlations result in larger similarity values for both the PPML and SURGINET users. What is of particular interest here is the individual users that have high similarity in the combined correlation calculation and low semantic similarity: these are users that are being detected due to the adaptations to semantic correlation applied in equation (3.26).

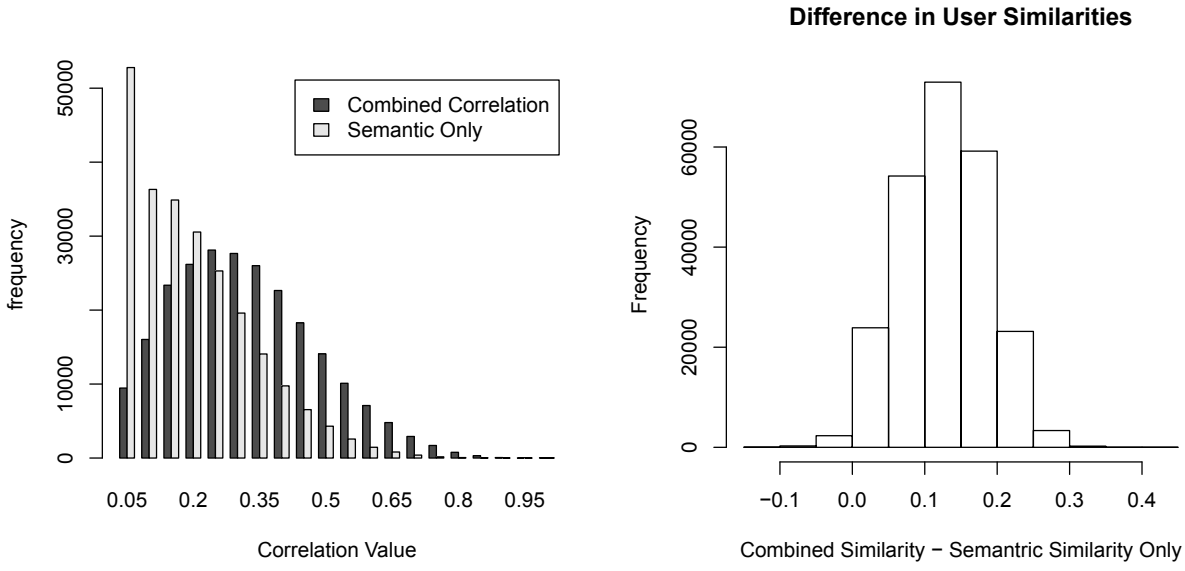


Figure 3.17: Distributions of the PPML user similarity measures calculated using a GVSM

We are going to investigate the similarity between two sample users from the PPML that fell into this group. Users S0493 and S0654 have a semantic-only similarity of 0.292 and a combined similarity of 0.456. User S0493 has 14 messages in the dataset, while user S0654 has 3. Both users seem interested in the pharmacological side of pediatric pain (table 3.11 presents a sample of their most used mesh terms). User S0493 has 239 distinct mappings, and user S0654 has 46 distinct mappings.

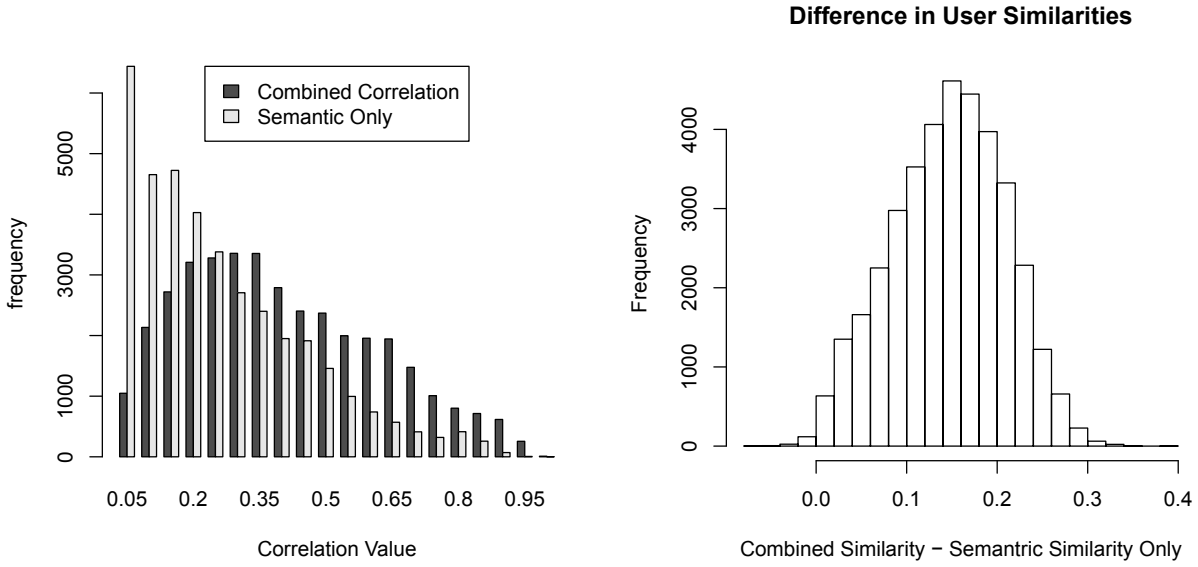


Figure 3.18: Distributions of the SURGINET user similarity measures calculated using a GVSM

The users have four MeSH terms in common (Pain, Analgesics, Pharmaceutical Preparations and Anxiety), and of the terms that they do not share there are 118 pairings between them that are semantically related (table 3.12 has a sample of those semantic pairings).

That leaves 9752 pairs of MeSH terms between the two users that have no semantic relationship. A sample of the terms from this set that have a co-occurrence correlation are listed in table 3.13. Table 3.13 also lists the term weights (from the TF-IDF calculation) to provide an idea of how influential those pairs were in the weighting of the final correlation calculation.

There are some interesting findings in table 3.13. The similarity between *Clonazepam* and *Intracranial Hypotension* and between *Histidine* and *Anxiety* seems to be a result of side effects of both drugs. The pairing between *Acetaminophen* and both *Morphine* and *Opioid Analgesics* along with the pairing between *Ketamine* and *Anti-Inflammatory Agents, Non-Steroidal (NSAIDs)* is not surprising, as a pain-based mailing list is often discussing pain relief, and though these are different types of pain relief, in the context of pediatric pain they are quite similar, and they often arise together in conversations. The correlations to *Pain* at the bottom of the table are a demonstration of why TF-IDF weighting is important.

S0493		**S0654**	
Score	*Term*	*Score*	*Term*
9582	Pain	2968	Vitamin B 12
7111	Pharmaceutical Preparations	2790	Folic Acid
6288	Morphine	2606	Anticonvulsants
6173	Analgesics, Opioid	2515	physiology
5177	Child	2381	Acetaminophen
4593	Ketamine	2000	Monitoring, Physiologic
3790	Thinking	1722	Analgesics
3482	Constipation	1722	Kidney Diseases
3440	Analgesics	1654	Kidney
3368	Methotrimeprazine	1618	Long-Term Care
3262	Adult	1000	Reading
3138	Disease	1000	Kidney Failure, Chronic
3004	Neoplasm Metastasis	1000	Intracranial Hypotension
3000	Clonidine	1000	Blood Cell Count
3000	Biological Assay	827	Dietary Supplements
2991	Pediatrics	812	Serum
2856	Histidine	804	Pharmaceutical Preparations
2684	Helping Behavior	790	Anti-Inflammatory Agents, Non-Steroidal
2465	Typhlitis	753	Tears
2390	Pain Management	753	Therapeutics

Table 3.11: A sample of terms used by the PPML users S0493 and S0654

Co-occurrence with the term *Pain* is common on a pain mailing list but not overly informative, and the weight ensures that it will not have a significant influence on the correlation calculation.

Table 3.13 also demonstrates the potential risks of mapping with Metamap and co-occurrence correlations. The terms *Hip* and *Tears* should not have a relationship, but the phrase hip labral tear has arisen in a conversation on the mailing list. This phrase has incorrectly mapped to *Tears*, as in what is produced when you cry, which has created a relationship between terms that are not related. This sort of mis-mapping when using semantic mapping programs is a known risk, and must be considered when interpreting the similarity measures reported. The assumption is that the poor mappings are out-weighed by the accurate mappings, resulting in little influence of incorrect terms.

The correlation between *Swimming Pools* and *Blood Cell Count* is an example of how rare terms can result in spurious similarities. Swimming pools have only arisen twice in the

Term 1	Term 2	Sem. Cor
Pain	Pain	1.000
Analgesics	Analgesics	1.000
Pharmaceutical Preparations	Pharmaceutical Preparations	1.000
Anxiety	Anxiety	1.000
Sensation	Pain	0.951
Analgesics, Opioid	Analgesics	0.922
Urinary Bladder	Kidney	0.788
Cystitis	Kidney Diseases	0.744
Porphyria, Acute Intermittent	Hepatitis E	0.717
Infertility	Kidney Diseases	0.680
Cystitis	Kidney Failure, Chronic	0.658
Pain Management	Therapeutics	0.628
Emotions	Anxiety	0.627
Analgesics	Anti-Inflammatory Agents, Non-Steroidal	0.621
Thalamus	Dura Mater	0.610

Table 3.12: A sample of the semantic correlations between the PPML users S0493 and S0654

archives, and both times were about the same patient, who experienced pain relief while swimming. Since this patient also had blood counts mentioned in their messages, there is a co-occurrence between the two terms, despite their lack of similarity. This is one of the short-comings of co-occurrence correlations, and why semantic correlations are generally preferred where available, however the overall value of co-occurrence correlations in table 3.13 outweighs their risk.

This example demonstrates the power that co-occurrence correlations can lend to the user similarity process. By supplementing the semantic correlations with co-occurrence correlations users with similar interests that use different lexicons may still be able to be connected.

S0493 Term	Weight	S0694 Term	Weight	Corr
Mitochondrial Diseases	3.33	Anticonvulsants	2.82	0.167
Clonazepam	2.08	Intracranial Hypotension	2.87	0.167
Hip	2.18	Tears	2.51	0.182
Mucopolysaccharidosis III	2.70	Blood Cell Count	1.30	0.200
Domperidone	1.07	Anticonvulsants	2.82	0.167
Hypertension	1.02	Intracranial Hypotension	2.87	0.250
Hip	2.18	Blood Cell Count	1.30	0.222
Dyskinesias	1.97	Blood Cell Count	1.30	0.200
Transport Vesicles	1.80	Blood Cell Count	1.30	0.167
Swimming Pools	1.78	Blood Cell Count	1.30	0.200
Analgesics, Opioid	1.82	Acetaminophen	1.17	0.186
Analgesics	1.75	Acetaminophen	1.17	0.175
Leg	1.53	Blood Cell Count	1.30	0.182
Morphine	1.70	Acetaminophen	1.17	0.168
Histidine	2.23	Anxiety	0.89	0.155
Pharmaceutical Preparations	1.85	Analgesics	1.05	0.178
Adult	1.78	Analgesics	1.05	0.158
Morphine	1.70	Analgesics	1.05	0.172
Fungi	0.98	Pain	0.37	0.200
Thinking	1.33	Pain	0.37	0.158

Table 3.13: A sample of the co-occurrence correlations between the PPML users S0493 and S0694

3.4.2.3 Ranking User Similarities

As with the comparison of the BGM vs BICGM methods, we are ultimately interested in whether the two methods return a different set of most similar users for each member of the community. Table 3.14 presents the overlap between the top 5 most similar users returned by one method and the top 10 returned by the other method for the PPML, and table 3.15 for the SURGINET data. The table demonstrates far more difference between the two methods than between the BICGM and BGM methods, though they still have significant overlap. 50% of users would find their top 5 most similar users according to the combined method in the top 10 of the semantic list, and another 29% would find at least 4 of them.

I explored potential explanations for why some users experience large overlap and others experience a small overlap. For the PPML data figure 3.19 presents the relationship between overlap and number of messages and mappings per user to see if the more prolific users are more or less likely to experience overlap, along with a comparison to the user's mapping

		Combined Similarity					
		0	1	2	3	4	5
Overlap with Semantic top 10	n	0	8	21	72	140	249
	proportion	0.000	0.016	0.043	0.147	0.286	0.508
		Semantic Similarity					
		0	1	2	3	4	5
Overlap with combined top 10	n	0	6	35	89	155	205
	proportion	0.000	0.012	0.071	0.182	0.316	0.418

Table 3.14: Comparing the overlap of most similar users on the PPML returned by each function

		Combined Similarity					
		0	1	2	3	4	5
Overlap with Semantic top 10	n	0	0	1	10	47	136
	proportion	0.000	0.000	0.005	0.052	0.242	0.701
		Semantic Similarity					
		0	1	2	3	4	5
Overlap with Combined top 10	n	0	0	2	21	64	107
	proportion	0.000	0.000	0.010	0.108	0.330	0.552

Table 3.15: Comparing the overlap between the most similar users on SURGINET returned by each function

score (normalized at the message level) to the term *Pain*, to see if specific terms are causing overlap. As the figure demonstrates, none of the relationships produced any meaningful outcomes.

For the SURGINET data figure 3.20 presents the relationship between overlap and number of messages and mappings per user, and reveals a noticeable trend, in that more active users are more likely to experience overlap between their methods. This means that, for the SURGINET data, as a user's activity level increases then the differences between the semantic correlation and the combined correlation diminish. This pattern may not have been present in the PPML because of the increased activity levels of the SURGINET users, or because of the comparatively smaller user pool on SURGINET. This results does suggest that supplementing the semantic correlations with co-occurrence correlations is of most value to the users that are less active within the community, even though this result was not borne out in the PPML data.

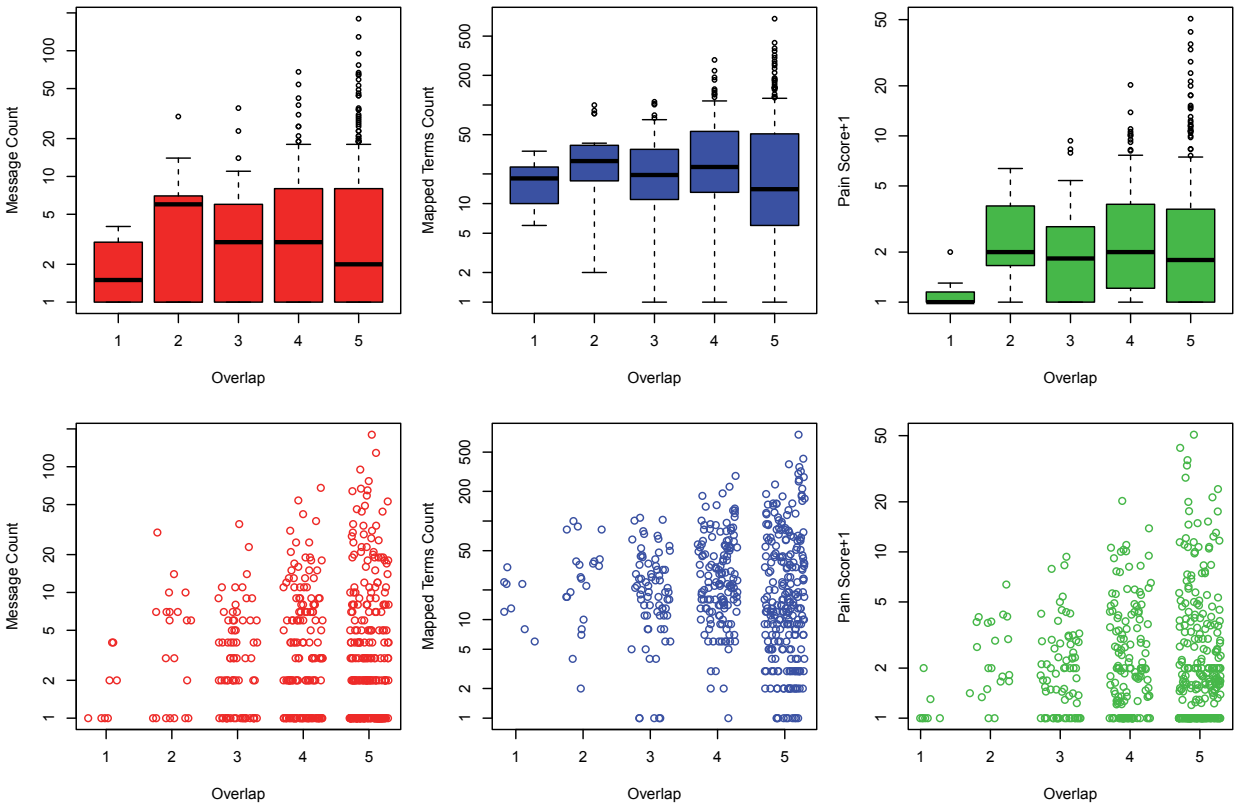


Figure 3.19: Exploring the potential causes of overlap (top 5 combined vs. top 10 semantic)

The effect of adding co-occurrence correlations to the semantic methods seems to be producing changes to the GVSM process. The examples above demonstrate the potential power of including these co-occurrence methods, particularly for the less active members of the community.

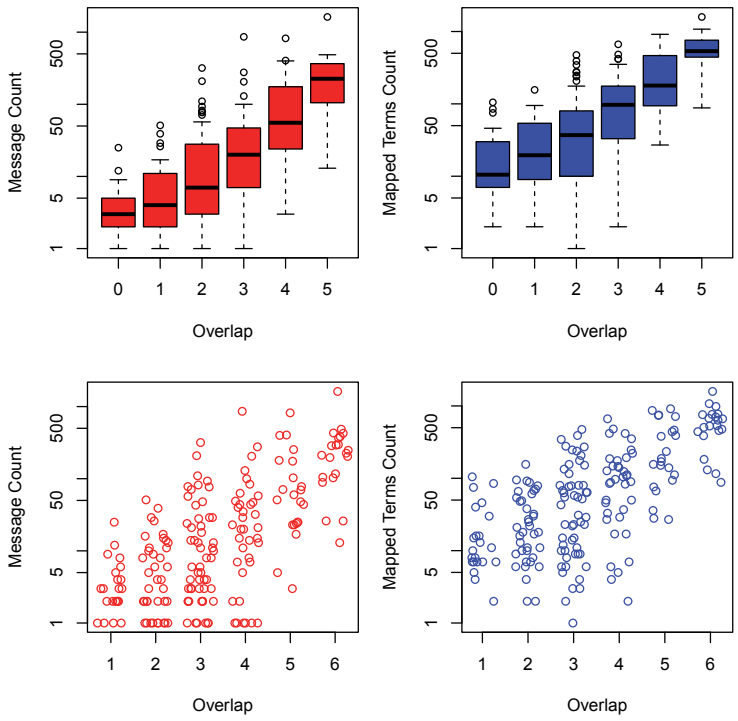


Figure 3.20: Studying the relationship between GVSM and BICGM overlap and network activity on SURGINET

3.4.3 GVSM vs. BICGM

The GVSM and BICGM approaches are measuring different relationships between two users. The BICGM method is asymmetric, so it measures how interested user A is in user B. If all the subjects that user A is interested in also interest user B then he would have a high BICGM similarity to user B. The converse is not necessarily true, however. If user B is much more prolific then she might not be as interested in user A, as he only covers part of user B's overall field of interest. Table 3.16 presents the nature of the relationships between users based on their BICGM values.

GVSM methods, on the other hand, measure how similar two users are. If a third party

		$BICGM[j,i]$	
		low	**high**
$BICGM[i,j]$	**low**	indifferent	CE→student
	high	student→CE	peers

Table 3.16: The theoretical relations between users based on their BICGM values (CE = Content Expert

were to look at the communications within the community and want to identify users that have similar knowledge bases, GVSM would provide that information.

We are interested in this section in investigating the nature of the relationship between BICGM and GVSM values. Figures 3.21 and 3.22 presents the pairwise differences between GVSM and BICGM values for the PPML and SURGINET data respectively.

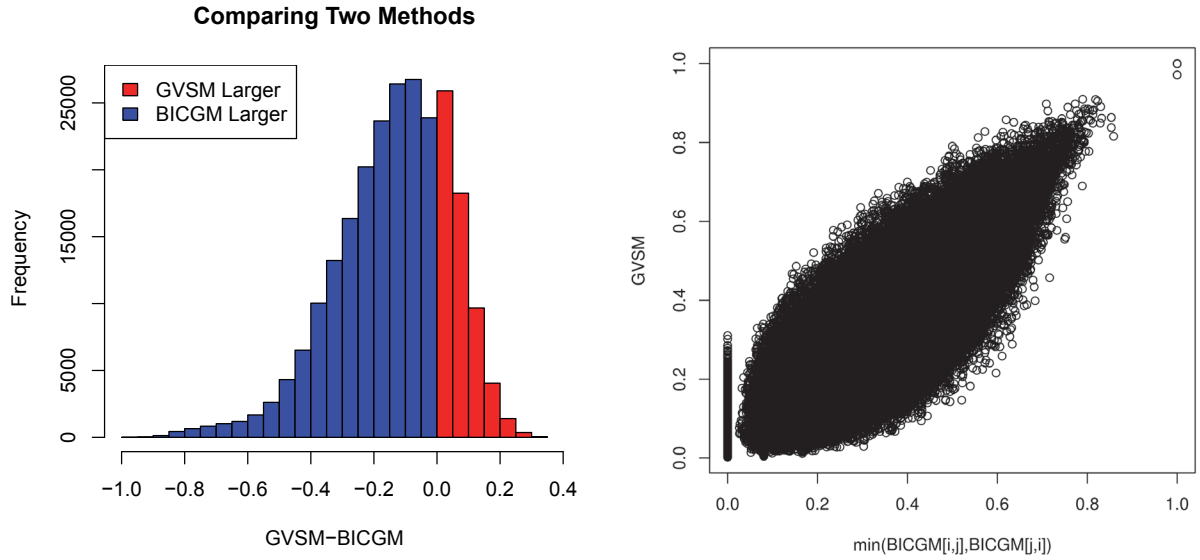


Figure 3.21: The pariwise differences between GVSM and BICGM values (left) and the linear relationship (right) for the PPML data

In general the BICGM values seem larger than the GVSM values, though this is mostly because the BICGM values were scaled larger to improve their distribution in the previous section. The figure also demonstrates that there is a difference between the methods.

The GVSM values are moderately correlated with the BICGM values. Since GVSM is symmetric and BICGM is asymmetric they cannot be compared directly, but the BICGM correlation matrix can be split into two symmetric matrices, one that records the larger of the two BICGM values for a pair, and one that records the smaller. For the PPML data the GVSM values have a correlation of 0.69 with the higher of the two BICGM values, and a correlation of 0.81 with the lower of the two BICGM values, while the SURGINET data has similar correlation values of 0.66 and 0.86 for the same pairs.

The figures also demonstrate the linear relationship between GVSM and the two BICGM values. When the smaller of the two BICGM values is low, the GVSM value ranges from low to moderate (left side of the figure). In the context of table 3.16 we're in the first

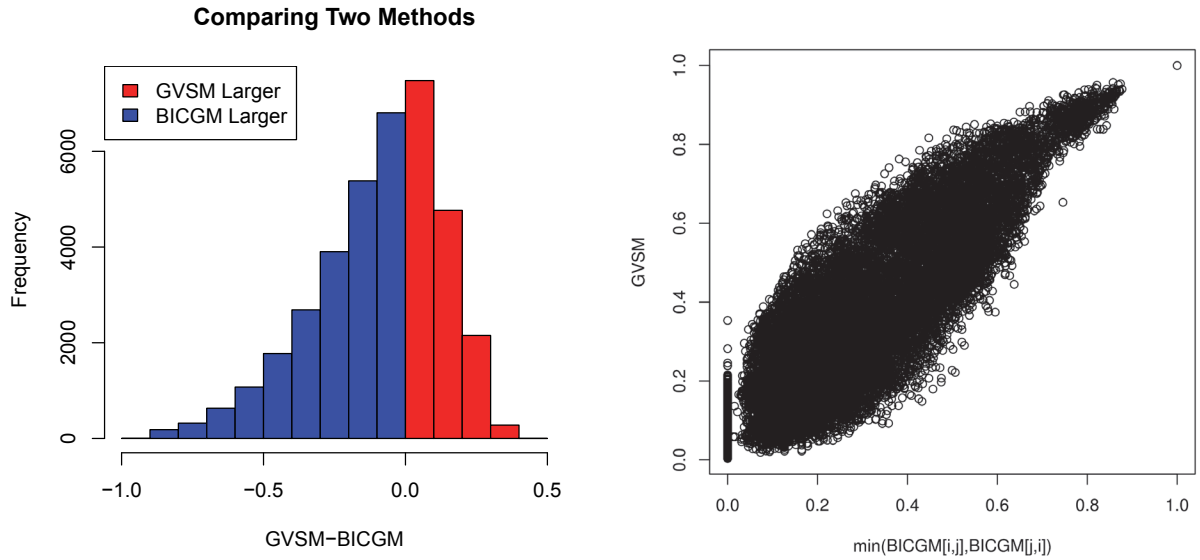


Figure 3.22: The pairwise differences between GVSM and BICGM values (left) and the linear relationship between GVSM and BICGM (right) for the SURGINET data

row, suggesting that the GVSM values are going to be low to moderate when one user is uninterested in the other. As the minimum value in the pair increases we move from the first to the second row of the table, and we see that the potential GVSM values both increase and narrow in range. At high BICGM minimum values we see high GVSM values. This is the bottom-right scenario in the table, where both users are interested in the work the other is doing.

Investigating the PPML examples from the previous sections can provide more insight into the nature of the relationship between the two methods. User S0602 had a BICGM similarity of 0.848 to user S0605, but user S0605 only had a value of 0.193 back. This is because S0605 has participated on a number of threads about a variety of topics, and is therefore less interested in the user that has only participated in a single thread about gum recession. The GVSM value for this pair is 0.178, indicating a low similarity between the two, which is expected, as there is a lot of content that S0605 has shared that S0602 is not interested in.

We investigated user S0872's BICGM value of 0.337 with a number of users. The reverse BICGM value, along with the GVSM value, are listed in table 3.17. The BICGM values in both directions are low, indicating that the pairs do not have much in common. Since there

is little interest in either direction, it is not surprising that the GVSM value is also low.

	S0504	S0602	S0702	S0735	S0783	S0814	S0857	S0888	S0937
$BICGM[i,j]$	0.337	0.337	0.337	0.337	0.337	0.337	0.337	0.337	0.337
$BICGM[j,i]$	0.297	0.357	0.373	0.303	0.331	0.245	0.237	0.360	0.278
$GVSM$	0.164	0.047	0.045	0.161	0.137	0.148	0.127	0.149	0.171

Table 3.17: The BICGM and GVSM similarities between PPML user S0872 and several other users

In only 2.6% of pairs does the GVSM exceed both BICGM values. In cases where the GVSM is higher, it is likely due to the co-occurrence similarity that is not captured by the BICGM model. Looking back at figure 3.21 there is a column of GVSM values at BICGM=0. Users whose MeSH subtrees did not overlap at all have a BICGM similarity of 0, but if their terms had a large co-occurrence similarity then there is potential for them to have a GVSM similarity. Consider users S0582 and S0736, who mapped to only 7 and 2 terms respectively (sample terms listed in table 3.18).

S0582 Terms	
treeLocation	term
D02.455.426.392.368.367.652	Ketamine
D26	Drugs
E01.370.520	Monitoring, Physiologic
E02.760.190	Critical Care
F02.463.902	Volition
H02.403.670	Pediatrics
N02.278.421.556.437	Hospitals, Pediatric
N02.421.585.190	Critical Care

S0736 Terms	
treeLocation	term
I02.903	Teaching
I03.946	Work

Table 3.18: The mappings for PPML users S0582 (left) and S0736 (right)

When you compare these terms to the terms there is no overlap between their MeSH trees, therefore they have BICGM similarities of 0. However, in the co-occurrence matrix the term Teaching had co-occurrence correlations > 0.1 with the terms Pediatrics, Drugs and Pediatric Hospitals, resulting in a GVSM similarity of 0.181.

3.4.3.1 Ranking User Similarities

As before, what is most interesting is whether the methods return the same sets of rankings. Table 3.19 presents the overlap in rankings between the top 5 from one method and the top

10 from another for the PPML data. Unlike the BGM vs BICGM comparison in table 3.8 and similar to the two GVSM methods in table 3.14 there are significant differences between the GVSM and BICGM methods, and the differences are even more pronounced. 25% of users have a top 5 most similar users that don't appear in their BICGM top 10, and only 3% find all 5 of them there. Similar patterns hold for the SURGINET data, presented in table 3.20

		GVSM					
		0	1	2	3	4	5
Overlap with BICGM top 10	n	123	153	96	71	30	17
	proportion	0.251	0.312	0.196	0.145	0.061	0.035
		BICGM					
		0	1	2	3	4	5
Overlap with GVSM top 10	n	139	139	99	72	41	0
	proportion	0.284	0.284	0.202	0.147	0.084	0.000

Table 3.19: Measuring the overlap between individual rankings of the GVSM and BICGM similarity measures on the PPML

		GVSM					
		0	1	2	3	4	5
Overlap with BICGM top 10	n	22	38	54	37	22	21
	proportion	0.113	0.196	0.278	0.191	0.113	0.108
		BICGM					
		0	1	2	3	4	5
Overlap with GVSM top 10	n	13	37	74	49	21	0
	proportion	0.067	0.191	0.381	0.253	0.108	0.000

Table 3.20: The overlap between individual rankings of the GVSM and BICGM similarity measures on SURGINET

What is more novel about the GVSM-BICGM overlap is its relation to number of messages and mapped terms, presented in figure 3.23 for the PPML data and figure 3.24 for the SURGINET data. We can imagine message count and number of mapped terms as a proxy for activity in the network. The users with multiple mapped terms are the more active in the community. When we considered the similarity between user S0602 and S0605 above we noted that S0602 would be interested in S0605 because all the subjects that S0602 had discussed had also been discussed by S0605, but the converse BGM or BICGM value would

be low, because S0605 is interested in several topics that S0602 has not participated in. The users that a particular user has a high BICGM similarity with are likely to have as many or more messages and mappings in the community. As your personal participation levels increase, the number of users more active than you decreases, therefore you are more likely to be paired with someone at a similar activity level, who in turn may have a converse BICGM value that is high. As we saw earlier in this section, when both BICGM values between a pair of users are high, the GVSM value is also likely to be high.

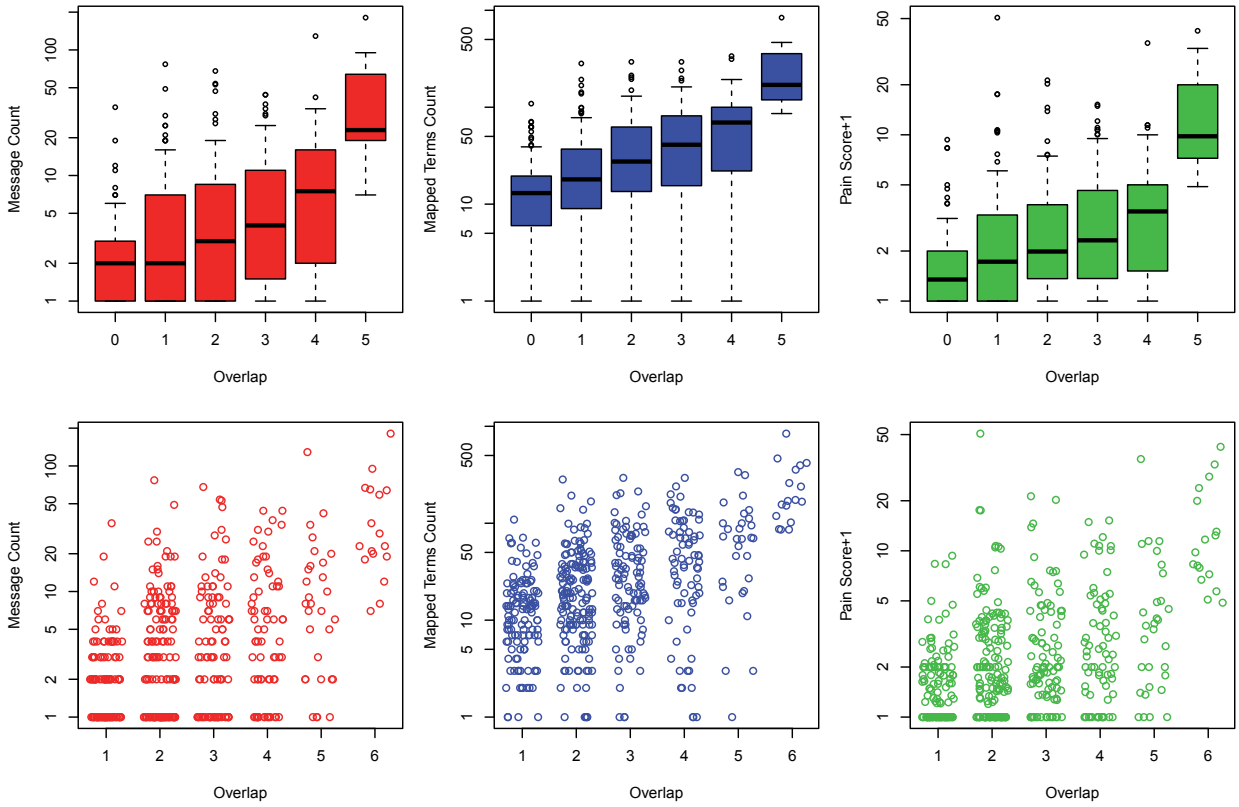


Figure 3.23: Comparing the relationship between community activity and overlap in BICGM and GVSM similarity on the PPML

The relationship outlined in both figures seems to be confirming the interpretation of the differences between BICGM, which measures how interested A may be in B's knowledge, and GVSM, which measures how similar two users may be.

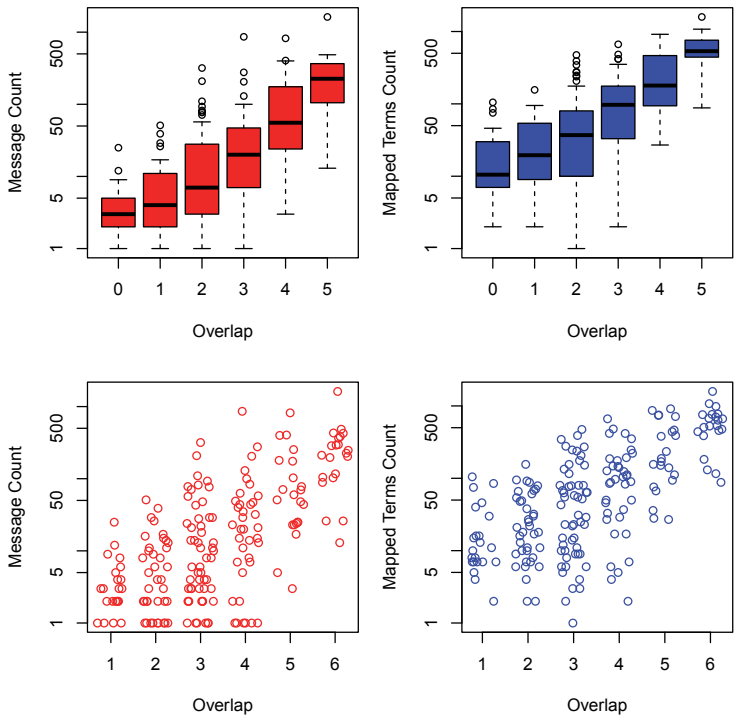


Figure 3.24: Comparing the relationship between community activity and overlap in BICGM and GVSM similarity on SURGINET

3.4.4 Summary

We have explored two different approaches to measuring user similarity within an online community, the BICGM and the GVSM. The two methods answer slightly different questions. The BGM and BICGM methods provide a user with the other community members that could be a resource to them: they are the users that have the most overlap in terms of communication locations. It is an asymmetric relationship, in that the relationship between A and B says nothing about the relationship between B and A. The GVSM is different, in that it symmetrically suggests which users are the most similar, in terms of the content they have communicated.

The evaluation of the BICGM at a micro level revealed some significant instances in which the addition of information content to the BICGM improved the term similarities, but at the user level there was little difference found between the two methods, suggesting that the information content did not affect the relationships enough to significantly alter the BGM method. For the GVSM the addition of co-occurrence correlation did significantly affect both individual term correlations and the user network as a whole, so moving forward

the GVSM with semantic and co-occurrence correlations should be used.

The comparison between the BICGM and GVSM methods are complex, as they are not measuring the exact same metric. We did discover a potential relationship, however, between GVSM and both pairs of a BICGM correlation. The evidence seems to suggest that if both users have a high BICGM value for the other then their GVSM value will be high, but if the BICGM values are different, or if both are low, then the GVSM value will also be low.

3.5 Combining Collaboration and Content Analysis

In figure 1.1 the analysis was split into two sections: the red section presented the analytic methods based on communication patterns, and the blue presented analytic methods based on knowledge content, but it is in their combination (the purple section) that real insight into the community at large can be made. The collaboration analysis provided the tools for finding pendants, but the content analysis will provide the means for investigating the potential causes for messages going unanswered. The semantic summaries can be used to summarize the connection-based clusters to identify if the clusters that were detected based on communication patterns may be attributable to specific knowledge areas.

The two methods for detecting user clusters (first based on communication patterns, second based on communication content) can be compared to see if there is any overlap, and any insights into the clustering can provide more insight into how the community functions. Finally, the BICGM correlations can be used to create a directed network, which can then be used to identify content experts based on centrality measures.

3.5.1 Understanding Pendants

Comparing the semantic content in the pendants compared to the other conversations may provide suggestions for why the questions were unanswered: The pendants could be covering a subject area that is not common on the list, the number of mapped terms could either be too high or too low, suggesting that the question was either lacking in medical details, or was too complex for the community to respond. Some mailing lists are going to have pendants for no discernible reason, but it is vital to ensure that messages are not being ignored based on systemic problems with the community.

3.5.2 Semantic Summaries of Communication Clusters

The communication clusters from section 3.2.3 are based on communication ties within the community, but what we are particularly interested in is the content of these clusters. The content of the messages within these clusters may provide insight into what is causing the clustering pattern to occur. Some of the divisions are going to be caused by network effects, high density clusters that are not based on content but are instead based on communication patterns. Beyond the clusters defined by network density, however, there is additional insight to be gleaned from the clustering. Even if the network exhibits a strong core-periphery structure there may be sub-clusters within the periphery defined by content, or there may be content guiding clustering within the core.

Section 3.3.9.4 outlined how to summarize the content of clusters based on the threads within them. We will apply these methods first to compare the user clusters and 2-mode clusters based on content, and then we will look at the content produced by the core versus the periphery to see if there is a difference in the terms being used by the power users compared to the rest.

3.5.3 Comparing Clustering Methods

There are three different methods for calculating user clusters: the 1-mode connection clusters, the 2-mode clusters (generalized blockmodeling) and the content clusters. These three clustering methods provide three different ways to find clusters in the network, so comparing their components may provide additional insight into the structure of the community. The clusters are formed in fundamentally different ways, so there may be no meaningful overlap in their structures, but the overlap may also help to better understand the clusters and reveal additional insights about the community.

3.5.4 Detecting Content Experts with BICGM Correlations

The Balanced Information Content Genealogy Model (BICGM) produces an asymmetric correlation matrix. The interpretation of the BICGM values are different than the GVSM results: The GVSM results are a measure of how similar two users are, while the BICGM results are a measure of how interested user i is in user j (with a separate value for the converse relationship).

This asymmetric relationship can be well represented in a directed network, using the directed network methods from section 3.2.2.3. We can create an *expertise* network from the BICGM correlations by creating an edge for every BICGM correlation above a minimum threshold d. The resulting adjacency matrix can be considered an expertise matrix R, in that an edge between i and j means that user i is interested in the content that user j has communicated. The choice of the threshold is subjective and dependent on the nature of the BICGM values for the network.

Within these directed networks the centrality indicators take on very specific interpretations. In-degree in the expertise network is the number of users that consider target user an expert. Proximity prestige extends the idea of in-degree beyond one step, as the premise is that, if user A considers B and expert, and B considers C an expert then A should also consider C an expert. This transitive property may not necessarily true, but is a concept worth investigating within the community. The final directed centrality to be investigated is rank-centrality, which is the concept that users are experts if they are considered experts by other experts. Whether we use the raw eigenvector decomposition or the authority measure from hub-authority analysis makes little difference, and either should provide a strong indication of who the content experts are within the community.

3.6 Conclusion

The objective of this thesis is to improve online KT practices through better understanding of the existing knowledge sharing dynamics within the community. We have isolated two major components of the LINKS [1] model that must be evaluated and monitored within a community in order to ensure a healthy KT environment: the culture of collaboration and the knowledge content. For the culture of collaboration we have presented network-based analytic methods for detecting and preventing pendants, for identifying knowledge seekers and community leaders, and for identifying the core of the community.

To understand and control the knowledge context of the community the content of the messages being shared must be analyzed. Using a semantic mapping program such as Metamap [5] we can represent a message, a user or a thread by their formal medical terms instead of the unstructured text. At a corpus level summarizing these mappings provides the knowledge maps for the community, a high-level summary of the knowledge being shared by the users, which can be used to better understand what is being discussed, as well as

identifying potential growth areas within the community.

Moving beyond the corpus level to the user/thread level, the term-based representations of the users and threads can be used to calculate similarities between both. There are a myriad of approaches to calculating these similarities in the literature, but most fail to fully incorporate the relationships between the medical terms used in the representation. Using MeSH as a sample medical lexicon, we have developed two novel approaches to calculating user and thread similarities: an asymmetric similarity using the BICGM and a symmetric similarity using a GVSM with a term correlation calculated using semantic and co-occurrence correlation. We evaluated these new methods through both individual investigations and network-level comparisons. At the individual level we found that both methods can improve the calculation of the correlation between users, but at the network level the BICGM was not found to be overly different from the BGM. Comparing the GVSM to the BICGM reveal some interesting correlations between the two, where users with high GVSM values tended to have high BICGM values, but differences in either BICGM value caused the GVSM values to drop.

The incorporation of the knowledge context analysis into the collaboration analysis is the final component of the project. Identifying the components of the culture of collaboration within the community is important, but the content analysis allows us to move beyond identification to investigate the nature of these components. The term mappings can provide insight into the nature of pendants and of the clusters and cores identified through SNA. Comparing and contrasting the different approaches to user clustering may provide insight into the nature of the community, and the BICGM can be converted into a directed network, through which content experts can be detected through directed centrality measures.

The next chapter will apply the methods developed in this chapter to our two sample datasets, the PPML and SURGINET, to evaluate their utility in the real world.

Chapter 4

Results

In this chapter we are going to apply the methods outlined in chapter 3 to two mailing lists: the Pediatric Pain Mailing List (PPML), a mailing list around the subject of pediatric pain, and a general surgeon mailing list (SURGINET). We will start by exploring the semantic mappings of both lists, studying the knowledge map that each mapping provides. From there we will study the network patterns of both lists to get insight into the culture of collaboration within the community. After that we will investigate the user and thread similarity methods. We will explore four different ways to calculate the similarities between users, and will also investigate the user and thread clustering methods on both mailing lists.

4.1 Data

There are two datasets that will be used for this project. The PPML is a community of 460 clinicians from around the world who meet online to discuss issues pertaining to pediatric pain. The archives of the PPML from 2009-02-02 to 2013-02-03 were extracted, during which time 2505 messages were shared on 783 threads.

The SURGINET dataset is a community of 865 clinicians from around the world that use the forum to discuss general surgical issues. The community is much more active, sharing over 17,000 messages on 2,111 threads by 231 users during the period 2012-01-01 to 2013-04-05.

The SURGINET community is more active that the PPML, but they are also more likely to have non-medically relevant conversations. On the PPML all messages are medically relevant: there are a mix of conference announcements, job advertisements and medical content. SURGINET does not appear to have many job advertisements or conference announcements, but it has a number of strictly non-medically relevant conversations: many humorous posts, or political discussions, or arranging meetups with other community members. This is not necessarily a problem, and one could make the argument that it may strengthen the community, but for the purposes of identifying medical knowledge within the community these

threads are "noise", and can bias the results of the semantic mapping and SNA significantly.

Because of the size of the SURGINET sample manually culling the non-medical threads was not feasible. 300 random threads were reviewed and noted as medically relevant or irrelevant. Figure 4.1 presents boxplots comparing the overall mapping score for each thread to relevance. The boxplots suggest that there might be a cutoff that can at least partially predict relevance.

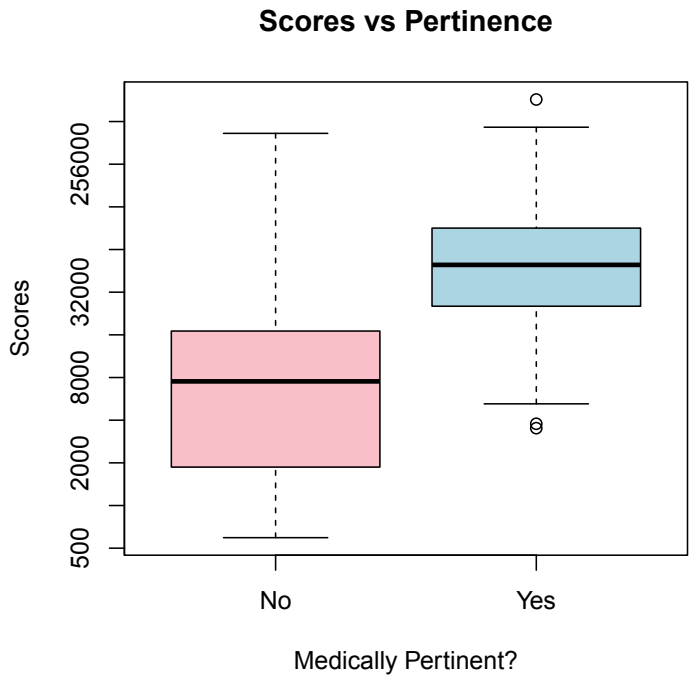


Figure 4.1: Comparing relevance to overall thread score

To determine the optimal cutoff value 1001 bootstrapping samples were taken from the list. For each sample an ROC was fit, and the optimal cut-point (determined by the phi-coefficient, or the highest sensitivity+specificity value) was found. Once the optimal cutpoint was fit it was tested on the observations that were not included in the fitting sample, and the accuracy of the classifier was found. Figure 4.2 presents the accuracy of classifiers by their cut-value.

The figure reveals three potential cutoffs: the average cutoff value, 16944, the average of the cutoffs that returned an accuracy $> 80\%$, 16180, or the most commonly reported cutoff, 15182. Looking at the figure, the most common cutoff, 15182, and the one next to it, 15185, report the highest accuracies, and are the most conservative, i.e., they are the least likely to

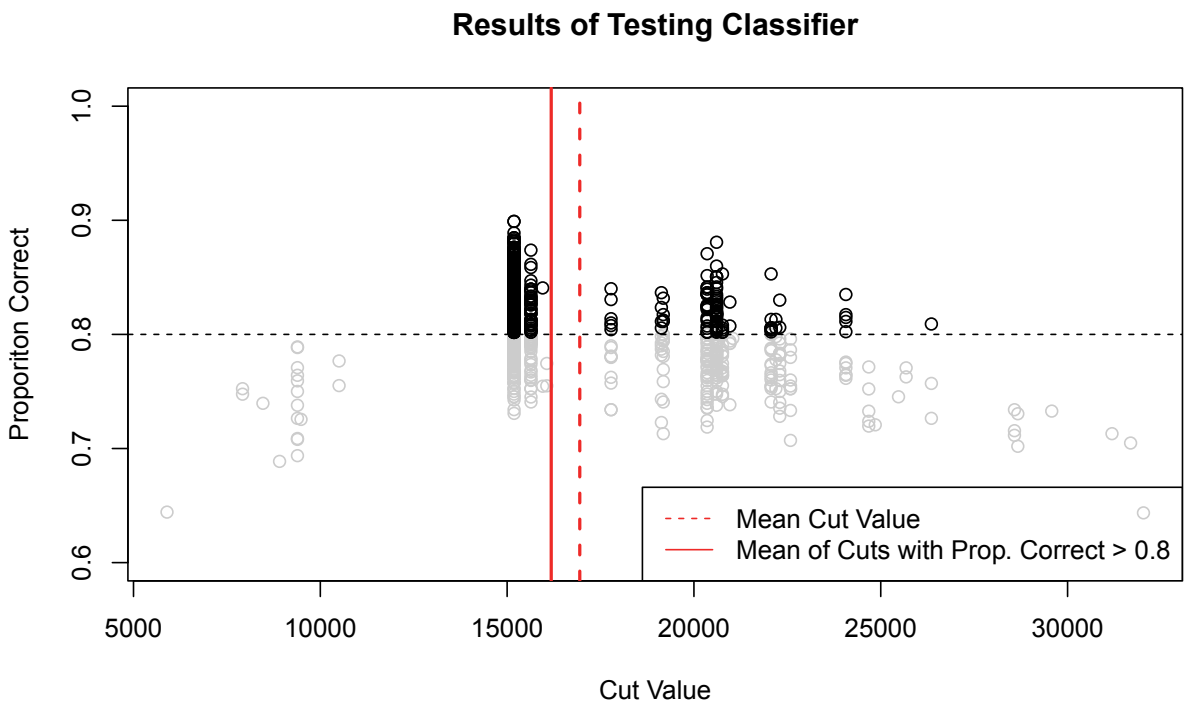


Figure 4.2: The results of bootstrapping to determine the optimal cut value

remove pertinent conversations, therefore I decided to drop all threads that had a cumulative thread score < 15182. In the sample of 300 threads there were 115 non-relevant threads and a cutoff of 15182 correctly classified 98 of them (specificity=85.2%). Of the 171 relevant threads, the cutoff correctly identified 137 of them (sensitivity=80.1%). A manual review of the content of these threads revealed that most of the threads incorrectly classified as irrelevant (i.e. dropped despite being medically relevant) were because of processing errors or incorrect separation from their source thread, and were a technical problem rather than an algorithmic one. The other misclassifications, irrelevant threads being reported as relevant, found long, non-medical conversations. There were some subject-line clues that could help in further cleaning, which helped improve the final dataset.

Finally, there was a small problem with thread-reuse, i.e., separate threads being grouped together by common subject lines. This was do largely to ambiguous subject lines (ex: "What would you do?", "Need Help", "Article Request") that were re-used multiple times. These threads were also removed to prevent further confusion in the analysis, and future work on the mailing list should focus on useful and descriptive subject lines rather than

uninformative, generic ones.

The result was a set of 13404 messages on 948 threads with 50597 total semantic mappings.

4.2 Knowledge Maps

There were 50597 terms mapped to the SURGINET data, compared to 27924 terms mapped to the PPML data. Figure 4.3 presents the number of mappings per thread and per message for each dataset.

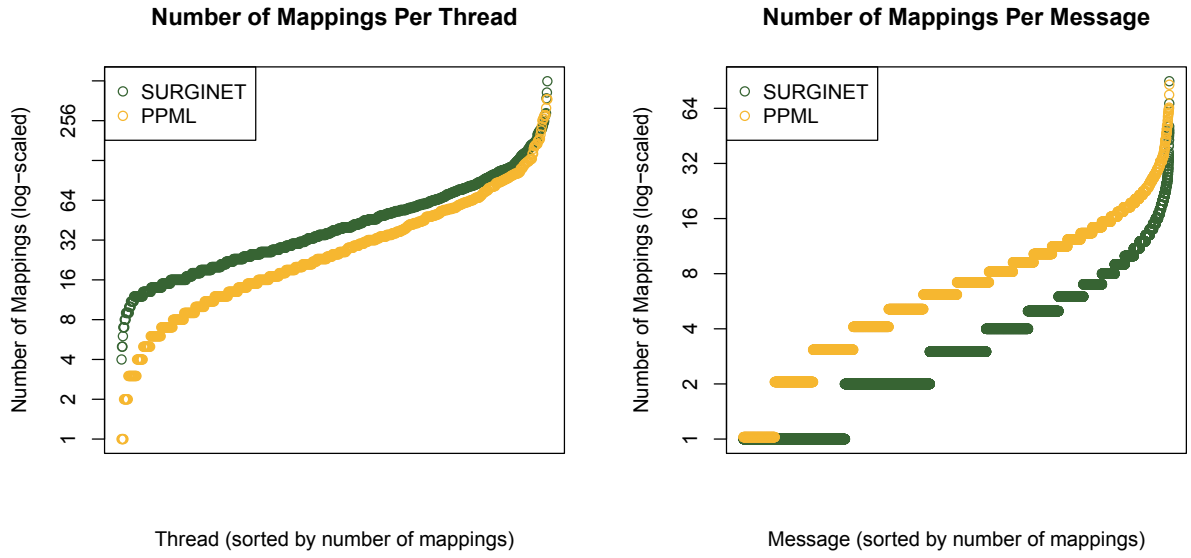


Figure 4.3: The number of MeSH terms per thread (left) and number of MeSH terms per message (right), sorted and log-scaled

4.2.1 Comparing Knowledge Maps

If the semantic mappings returned for each mailing list are thought of as the representation of the knowledge map that each list occupies, then studying the map that each mailing list occupies can provide further insight into the content within each community. Figure 4.4 compares the mappings of each mailing list at the root mesh level, where there are 15 classes of terms. The term labels are presented in table 4.1.

SURGINET has significant mappings to *Anatomy* (A) and *Analytical, Diagnostic and Therapeutic Techniques and Equipment* (E), while the PPML maps a lot to *Chemicals and*

Root	Root Name
A	Anatomy
B	Organisms
C	Diseases
D	Chemical and Drugs
E	Analytical, Diagnostic and Therapeutic Techniques and Equipment
F	Psychiatry and Psychology
G	Phenomena and Processes
H	Disciplines and Occupations
I	Anthropology, Education, Sociology and Social Phenomena
J	Technology, Industry, Agriculture
K	Humanities
L	Information Science
M	Named Groups
N	Health Care
V	Publication Characteristics
Z	Geographicals

Table 4.1: The MeSH Tree Roots

Drugs (D), and both mailing lists map significantly to *Diseases* (C). We will investigate each of these subtrees in more detail.

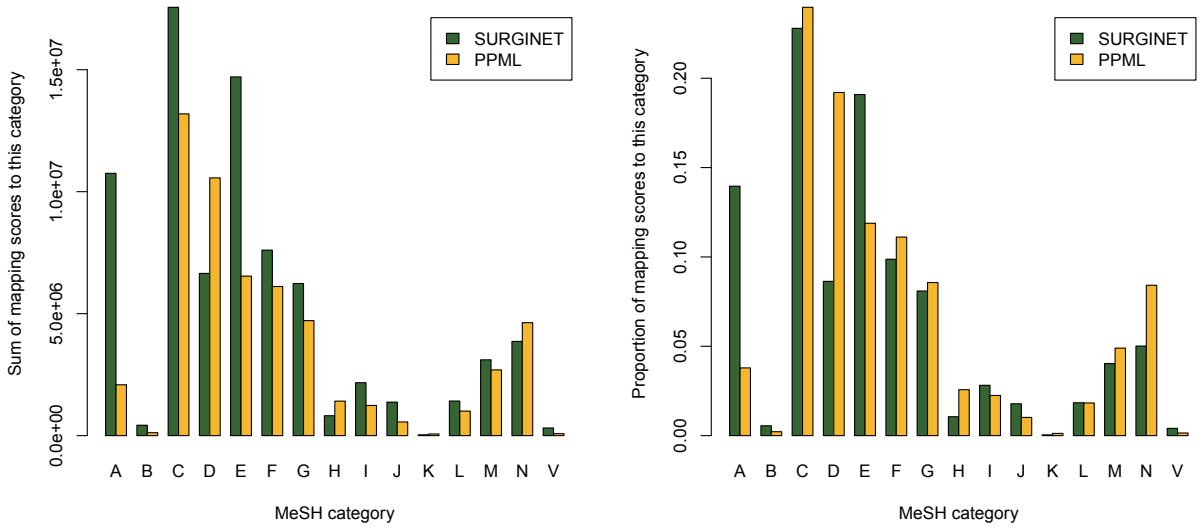


Figure 4.4: Comparing the overall (left) and proportional (right) mappings for each mailing list at the MeSH root level

4.2.1.1 Anatomy (A)

The terms most mapped to in the Anatomy section are presented in figure 4.5, and show that the categories A01: *Body Regions* and A03: *Digestive System* are the most commonly mapped terms on SURGINET. Figure 4.6 presents the breakdown of each of these these categories.

Body Regions seems to have mapped to four major terms: *Breast, Amputation Stumps, Back* and *Abdomen*. This is in contrast to *Digestive System*, which has no single dominant term, but many mappings to different components of the gastro-intestinal tract (A03.556). These terms represent areas that are often the target of surgical procedures, and thus their presence in SURGINET is not surprising. Pediatric pain patients rarely have the same targeted discussions about specific body parts: the issues that raise the most discussion on the PPML are pain problems that are difficult to isolate, and therefore anatomy issues rarely arise.

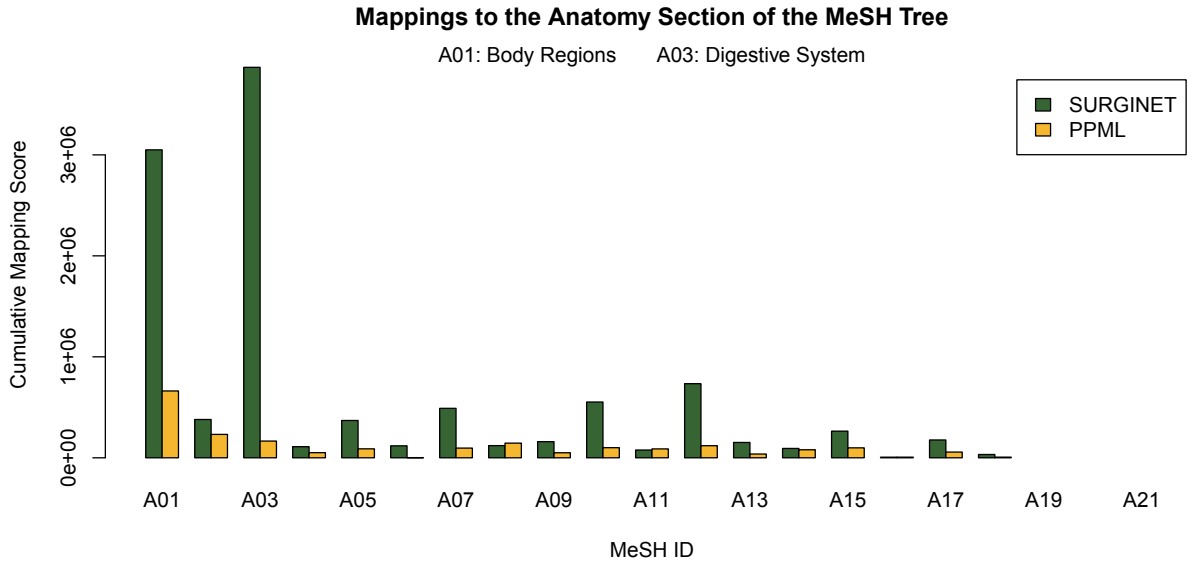


Figure 4.5: The section mappings for the Anatomy section for both SURGINET and the PPML

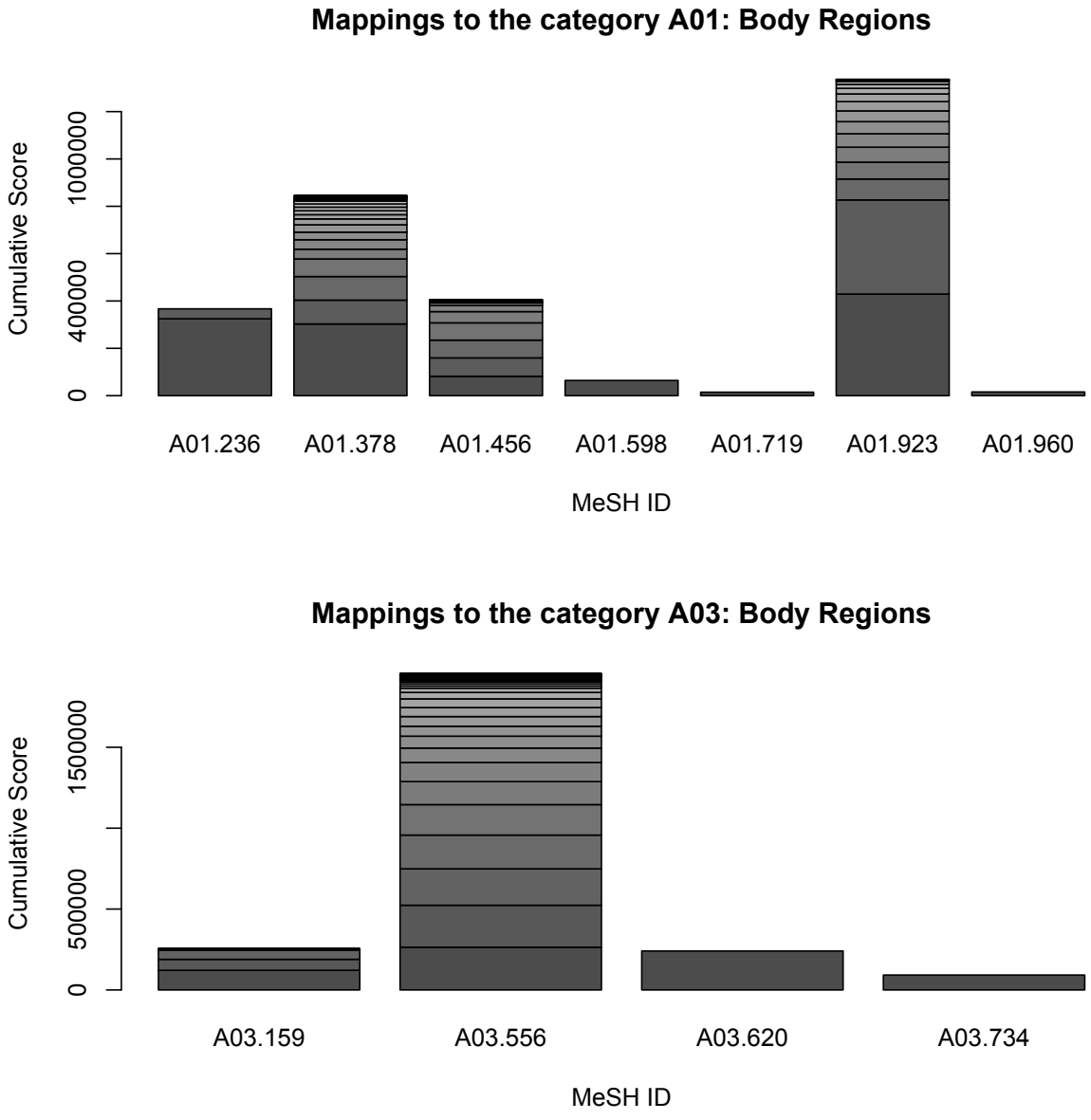


Figure 4.6: The components of the two most popular subgroups of the Anatomy tree for SURGINET (breaks within each bar represent individual terms)

4.2.1.2 Diseases (C)

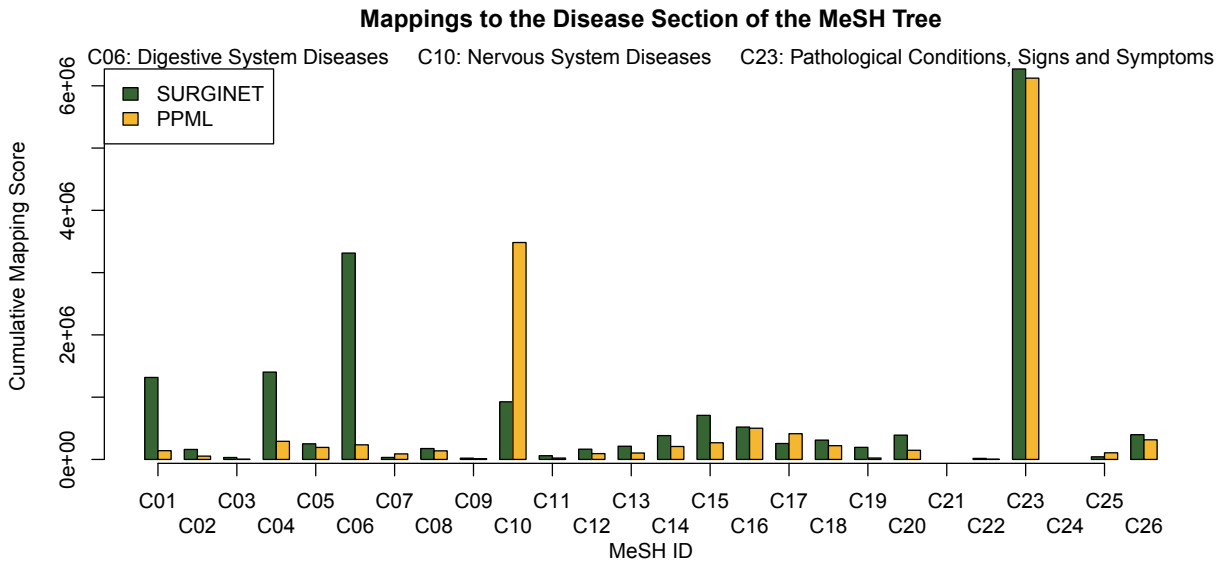


Figure 4.7: The section mappings for the Disease section for both SURGINET and the PPML

The mappings for the subsections of Disease (C) are presented in figure 4.7, and reveal that both lists map significantly to C23: *Pathological Conditions, Signs and Symptoms*. Beyond that very common subgroup the PPML discusses C10: *Nervous System Diseases* more, while SURGINET discusses C06: *Digestive System Diseases* more.

C06 is driven largely by C06.405: *Gastrointestinal Diseases*, and in particular by *Appendicitis*, a very common surgical procedure.

C10.597.617 is one of the codes for *Pain*, so the prevalence of terms mapped to C10 is driven completely by the mappings to the term Pain in the PPML.

Figure 4.8 presents the components of mappings to C23. For the PPML the mappings are again driven by the term *Pain* (C23.888.646), but SURGINET is more spread out, with no single sign or symptom driving the mappings. In the process of discussing surgical procedures a number of *Signs and Symptoms* (C23.888) are discussed, so it not surprising that the surgeons discuss a number of terms from this list.

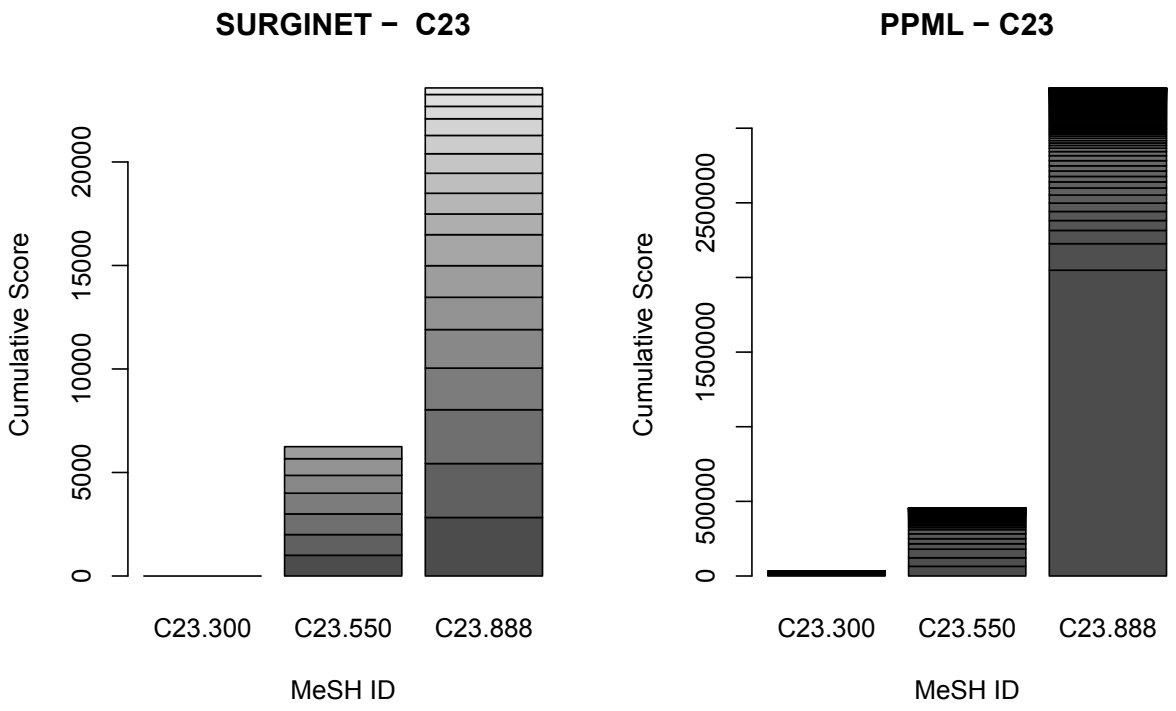


Figure 4.8: The components of mappings to C23: *Pathological Conditions, Signs and Symptoms* for both SURGINET and the PPML (breaks within each bar represent individual terms)

4.2.1.3 Chemicals and Drugs (D)

Figure 4.9 presents the mappings to the *Chemicals and Drugs* subgroups for both SURGINET and the PPML. Two terms in particular drive the PPML's dominance of these mappings, D03: *Heterocyclic Compounds* and D27: *Chemical Actions and Uses*. Figure 4.10 presents the proportional contributions to each component of D03 and figure 4.11 presents the contributions to the components of D27..

D03 is driven by four terms for specific pain relievers: *Morphine, Fentanyl, Clonidine* and *Codeine*, while D27 is driven by analgesics: *Analgesics, Opiod Analgesics and NSAIDs*. In a mailing list focused on pain relief the prevalence of a myriad of pain relievers is not surprising. Particularly within the realm of pediatric pain there is not always a consensus solution to every pain problem, and thus main discussions can arise.

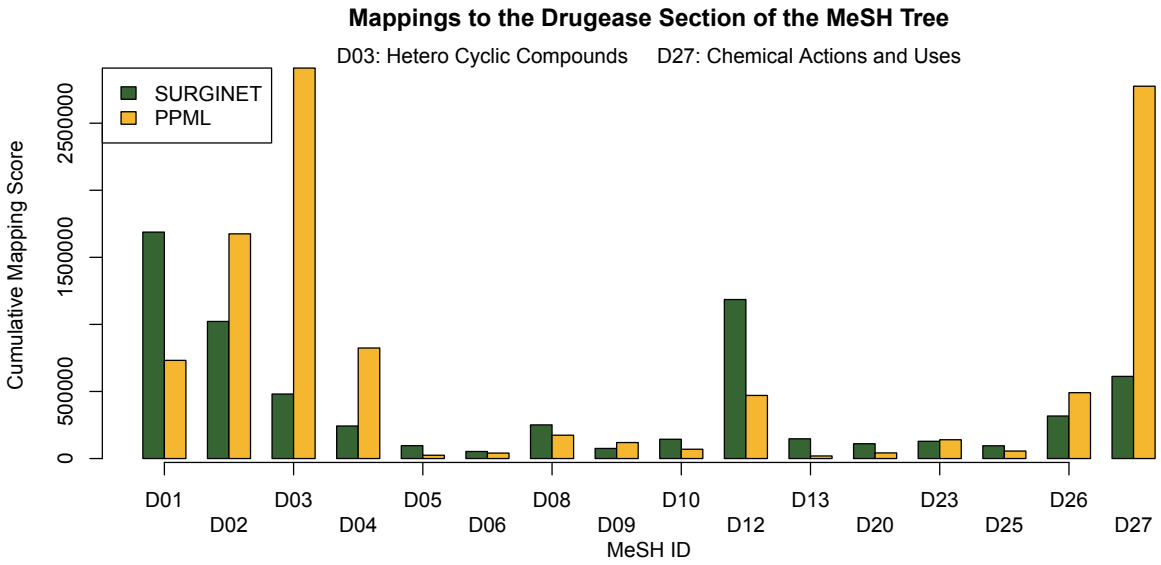


Figure 4.9: The section mappings for the Chemicals and Drugs section for both SURGINET and the PPML

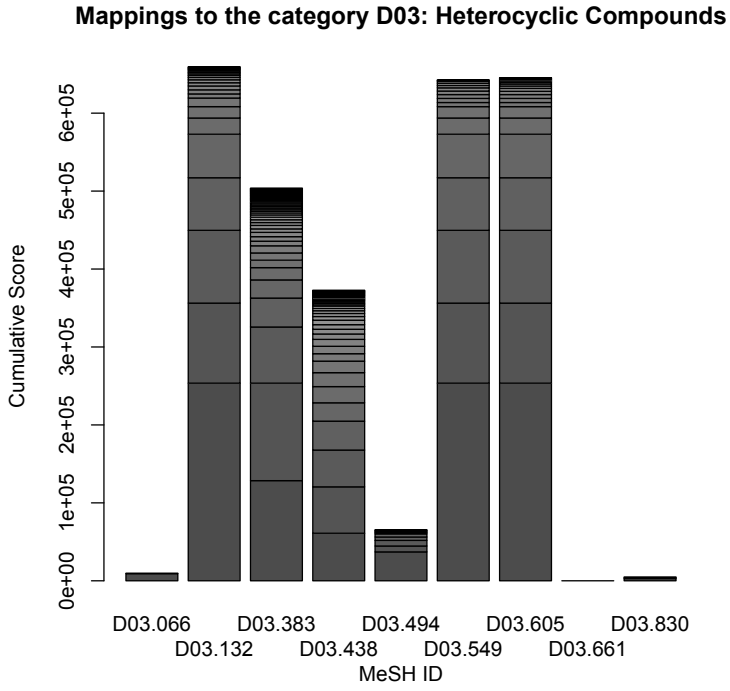


Figure 4.10: The components of D03 in the PPML (breaks within each bar represent individual terms)

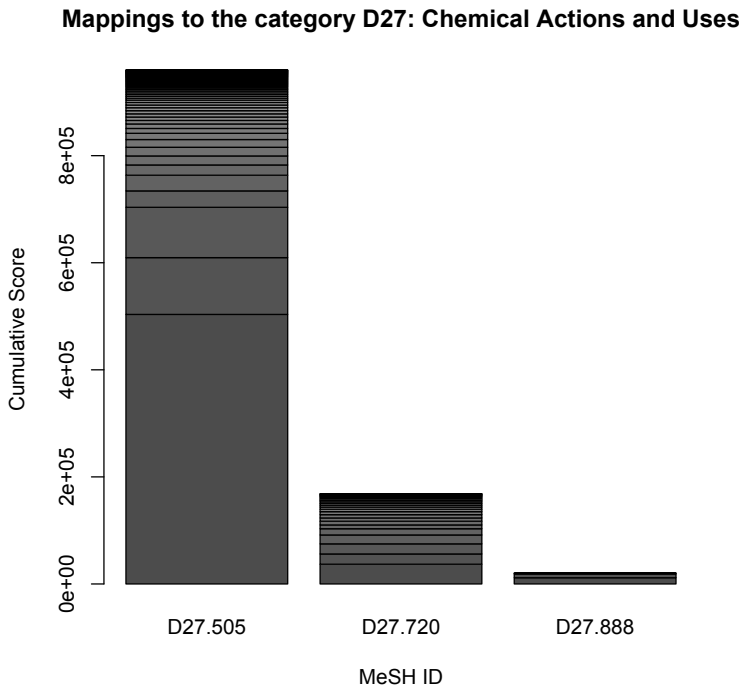


Figure 4.11: The components of the two subgroups D27 in the PPML (breaks within each bar represent individual terms)

4.2.1.4 Analytical, Diagnostic and Therapeutic Techniques and Equipment (E)

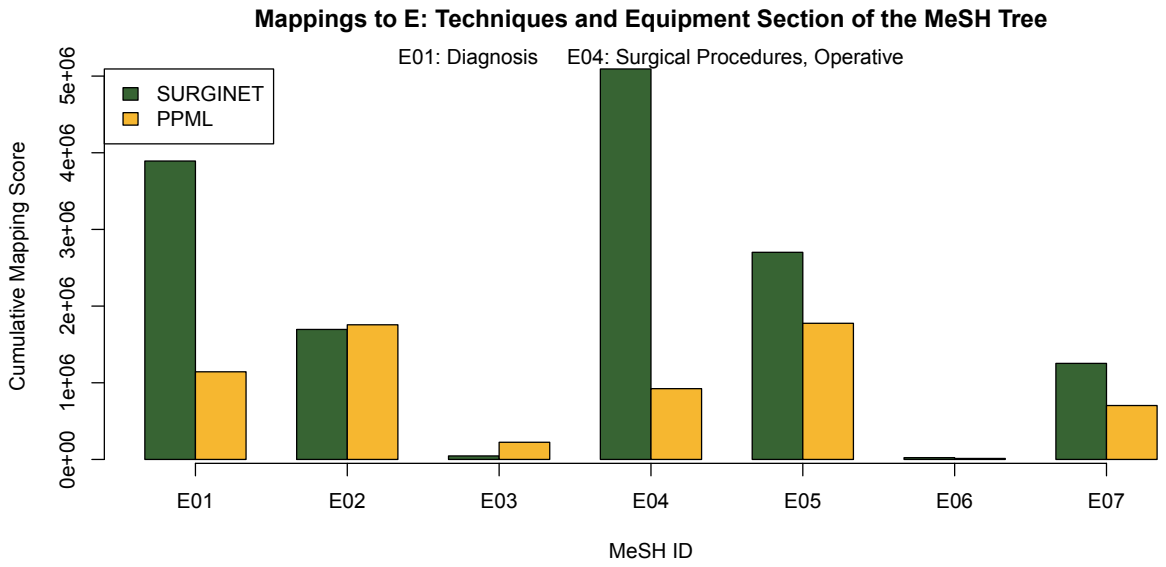


Figure 4.12: The section mappings for the Analytical, Diagnostic and Therapeutic Techniques and Equipment section for both lists

Figure 4.12 presents the components of the mappings to *Analytical, Diagnostic and Therapeutic Techniques and Equipment* (E) section of the MeSH tree. SURGINET has far more terms in this section of the tree, driven by E01: *Diagnosis* and E04: *Operative Surgical Procedures*. Neither of these results are surprising at all, as the presence of more surgical procedures on a surgical mailing list compared to a pain mailing list is expected. E01 is driven by the terms *Biopsy, Sentinel Lymph Node Biopsy* and *Laparoscopy*, while E04 has a number of procedural terms present: *Mastectomy, Laparoscopy, Laparotomy, Drainage, Appendectomy, etc....* Figure 4.13 presents the components of each term.

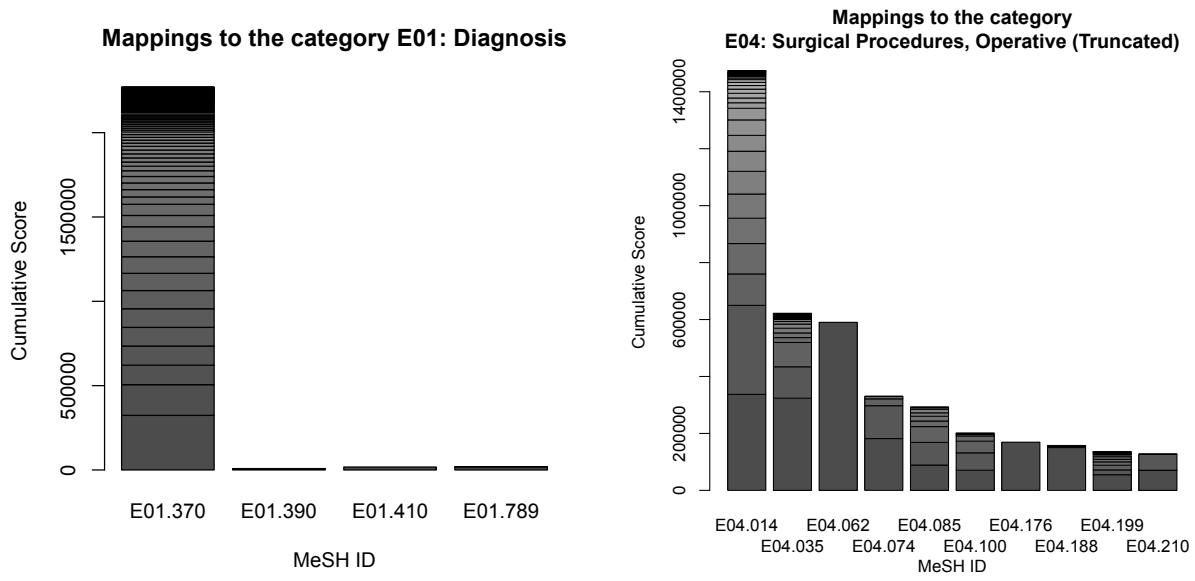


Figure 4.13: Exploring the components of terms E01 and E04 in SURGINET (breaks within each bar represent individual terms)

4.2.2 Individual Mappings

Beyond comparing knowledge maps, comparing individual mappings can provide some valuable insight into the structure of the communities. SURGINET mapped to 2711 individual MeSH terms, while the PPML mapped to 2485 terms. Of those terms there are 1199 that appear in both lists, and the correlation between the mapping scores of the terms in both lists is 0.46, which represents a moderate amount of agreement. Figure 4.14 plots the mapping scores of terms that appear in both lists against one another.

In both lists we see the individual terms that we would have expected from the previous section. The PPML maps to *Pain, Child* and *Pediatrics* more, along with *Opiod Analgesics* representing one of the many pain relievers that are present in the mailing list. For SURGINET the elevated level of *Patients* and *Thinking* is odd, but the presence of *Drainage* and *Appendicitis* is what would have been expected from the previous analysis.

What is also interesting is the terms that did not appear in one list and were quite prevalent in the other. Table 4.2 presents the highest scoring terms in one list that were not present in the other.

The highest scoring SURGINET terms that are not present in the PPML are procedure specific: *Bile, Fistula, Mastectomy, Labarotomy, Breast. Breast* is interesting here, as breast

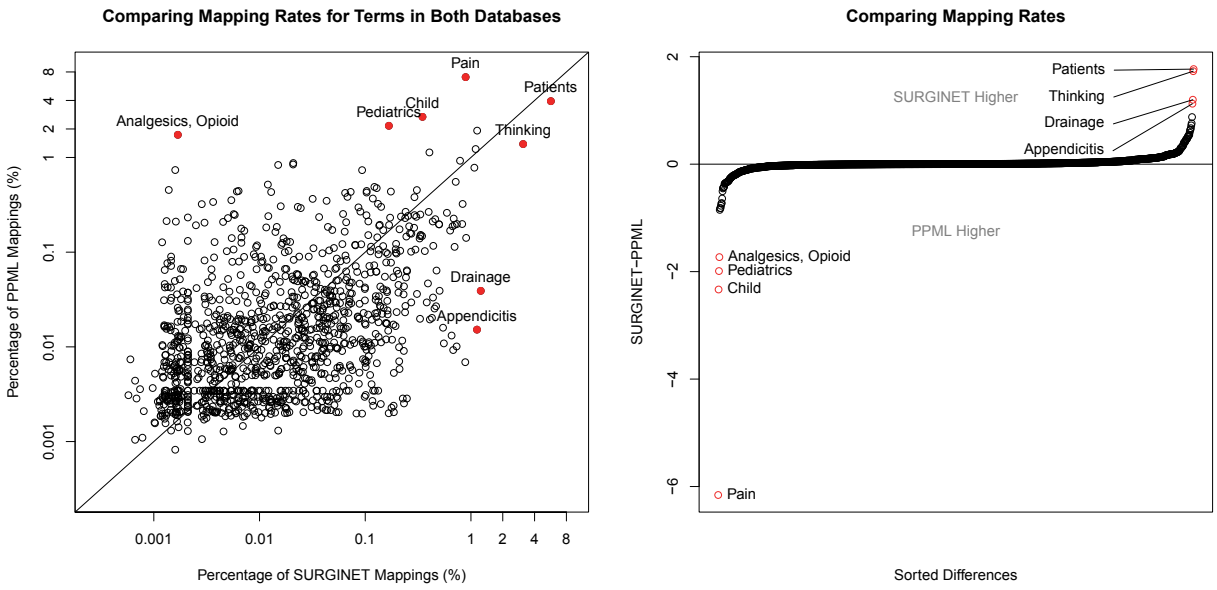


Figure 4.14: The mapping scores for terms that appeared in both lists

feeding is obviously a term in the PPML (0.027% of scores), but the ability of Metamap to differentiate *Breast* from *Breast Feeding* left the term *Breast* out of the PPML.

PPML terms that are not in SURGINET are about specific pain relievers: *Methadone, Ketamine, Clonidine, Pain Measurement/Management, Chronic Pain.* Oddly, *General Surgery* appears in the PPML, but not in the surgery mailing list. This is an example of implied language, as the surgeons do not use the term general surgery: they would never say "contact General Surgery" because they are general surgery, so the specific term never arises despite it being the context of the entire community.

SURGINET	
MeSH Term	**Proportion of Mappings**
Mastectomy	0.315
Bile	0.331
Laparotomy	0.355
Fistula	0.389
Duodenum	0.395
Carbidopa	0.432
Breast	0.681
Cholecystectomy	0.707

PPML	
MeSH Term	**Proportion of Mappings**
General Surgery	0.373
Clonidine	0.430
Organization and Administration	0.494
Methadone	0.597
Chronic Pain	0.612
Pain Measurement	0.657
Weights and Measures	0.768
Ketamine	0.787
Methods	1.168
Pain Management	1.769

Table 4.2: The most prevalent terms from each list that were not present in the other

4.2.3 Conclusion

The user of knowledge maps has provided detailed insight into the structure of both communities. The PPML is, not surprisingly, focused on pain, but within the context of pain they discuss medications and chemical solutions to manage pain more than any other subgroup, and spend less time on pain-specific conditions, and neurological aspects of pain. SURGINET seems to focus on the details of surgical procedures and surgical locations. An investigation of the specific surgical areas that are present and absent from the list of popular terms may provide some insight to the community at large about what they are discussing and ignoring, but that content is beyond the scope of this thesis. The significant overlap in $C23$ demonstrates what would most likely be true of all medical mailing lists, which is the significant overlap of diagnostic questions and criteria, along with signs and symptoms of disease.

4.3 Identifying Collaboration Patterns

The first step in understanding the KT patterns in the community is understanding the collaboration patterns at a network level. This section will investigate the issue of unresolved knowledge seeking behaviour (pendants), and will investigate the general activity rates of the two lists to get a sense of how the community members are participating within the community.

There are 1121 threads in our database from the PPML and 3596 messages, resulting in an average number of messages per thread of 3.2 (median of 2, range of 1-37). The PPML has 493 users, which means that the average messages per user is 7.3 (median of 3, range of 1-180). Figure 4.15 presents the distribution of number of messages per thread and per user, and we can see that the number of messages is much more skewed per user than per thread, demonstrating a significant difference between the most and least active users in the community.

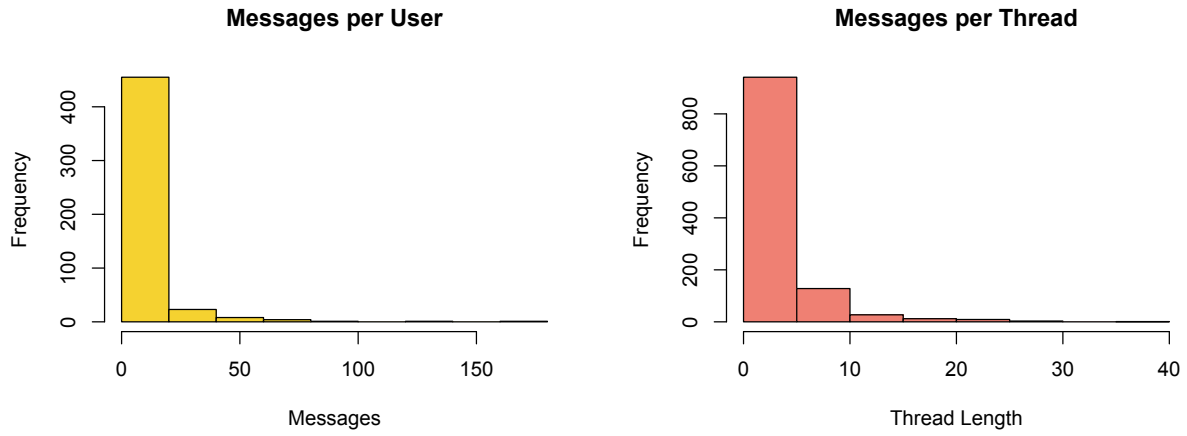


Figure 4.15: The number of messages per user (left) and per thread (right) on the PPML

For the SURGINET data threads had an average of 14 messages (median of 11, range of 2-77), while users averaged 68 messages each (median of 12, range of 1-1632). Figure 4.16 present the distribution of the messages per thread and per user.

The SURGINET community is more active that the PPML, but shows a similar pattern of having a couple of threads and a couple of users that are significantly more active than the rest.

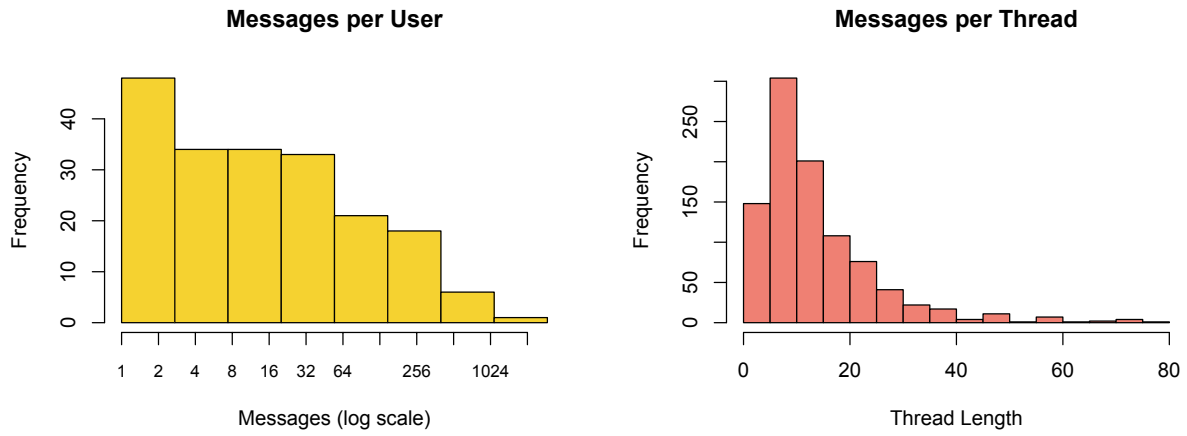


Figure 4.16: The number of messages per user (left) and per thread (right) on SURGINET

4.3.1 Isolates

There are 525 messages on the PPML that did not receive a response in the database. Through a combination of automated subject-line checking and manual review I stratified the isolates into three categories: There were 373 administrative messages, 74 error messages (mostly incorrect out-of-office messages, along with some mis-classification) and 78 true messages that did not receive a reply, i.e, pendants. When we factor out the other isolates that leaves 674 threads in the community, of which 11.5% did not receive a reply.

What we are particularly concerned about is that pendants occur due to exclusion by the community. To investigate this potential problem we looked at pendant rates relative to initial posting rates. Of the 453 community members, 149 had their first message on the mailing list be a new thread (as opposed to replying to an existing thread). Of those 149 new threads, 19 were pendants, which means that the pendant rate amongst new users was 12.7%,, roughly the same as the overall pendant rate of 11.5%. This suggests that the community is not discriminating against new users, and that their initial threads are just as likely to be replied to as existing users' new threads.

An investigation of the terms used in the pendants compared to the first message in non-pendant threads did not reveal any evidence that the pendants contained significantly different semantic terms than the non-pendants. The most common terms used in pendant threads were pain relievers (Morphine, Opioid Analgesics) along with general terms (Pain, Patients, Hospitals, Pain Management), with no prevalent terms outside what is normally

seen in the mailing list.

The number of mapped terms per message was similar between pendants and non-pendants with median total scores of 8417 and 8436 respectively (p-value = 0.2208), so there is no evidence to suggest that pendants are shorter messages than threads that received replies. Overall there does not seem to be any discernible cause of pendant messages. This can be seen as both an advantage, in that there is no systemic bias causing messages to go unanswered, and a disadvantage, in that without a discernible cause it is more difficult to find a solution. The empirical methods for detecting pendants based on response times can be used to identify pendants in process, but the lack of explanation may warrant further investigation.

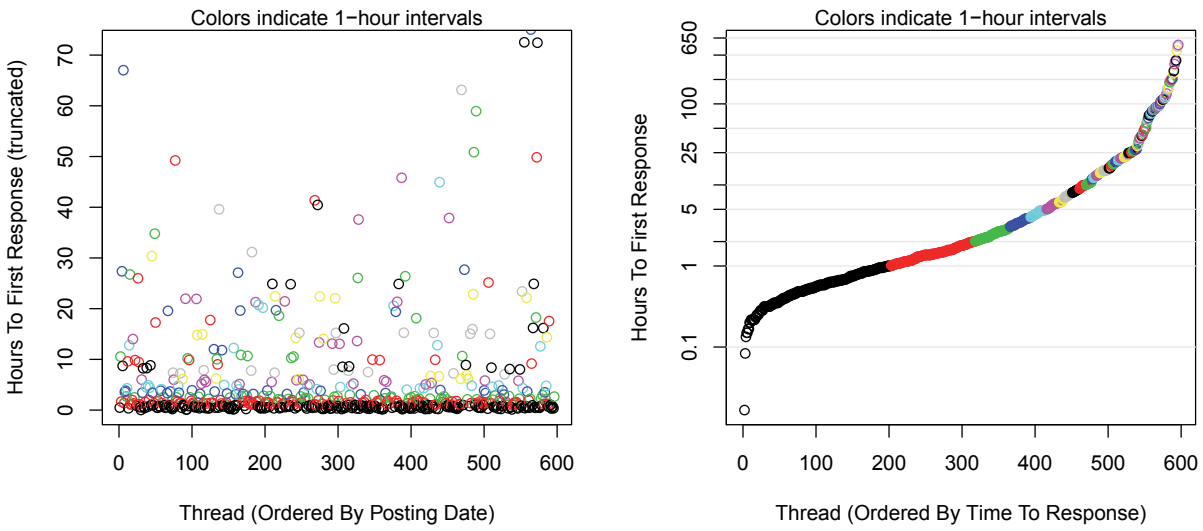
SURGINET does not suffer from the same issue of isolates that the PPML does. Of the nearly 18,000 messages within the community only 500 were isolates, and the majority of them were technical errors in parsing the archives rather than actual pendants. Due to the cleaning processes outlined above all the isolates were removed from the dataset, but even if they were retained the active nature of SURGINET precluded any pendants arising within the community.

4.3.2 Response Times

For the PPML figure 4.17 presents the time from initial post to time to first response for all 596 threads that received at least 1 response. The response times for the mailing list seem quick, with 34% of threads receiving their first reply within an hour, and 88% receiving their first reply within the first day. In fact, of the 596 threads that received a reply, only 47 had to wait more than 2 days, and only 26 had to wait more than 4.

Because of its higher activity level SURGINET has much faster first-response times than the PPML. Figure 4.18 presents the time to first reply for all the threads in the dataset. 66% of threads receive their first reply within an hour and 98.5% of the threads receive their first reply within 24 hours.

Figure 4.19 presents the message times for each of the threads that received a response on the PPML. 10% of threads last about an hour, and over 55% of threads last a day or less. Some threads stretch on longer, with 25% of threads lasting 75 hours (about 3 days) or more. The SURGINET thread lengths are remarkably similar to those on the PPML, with the majority completing within 2 days, and a few stretching on significantly longer. Figure



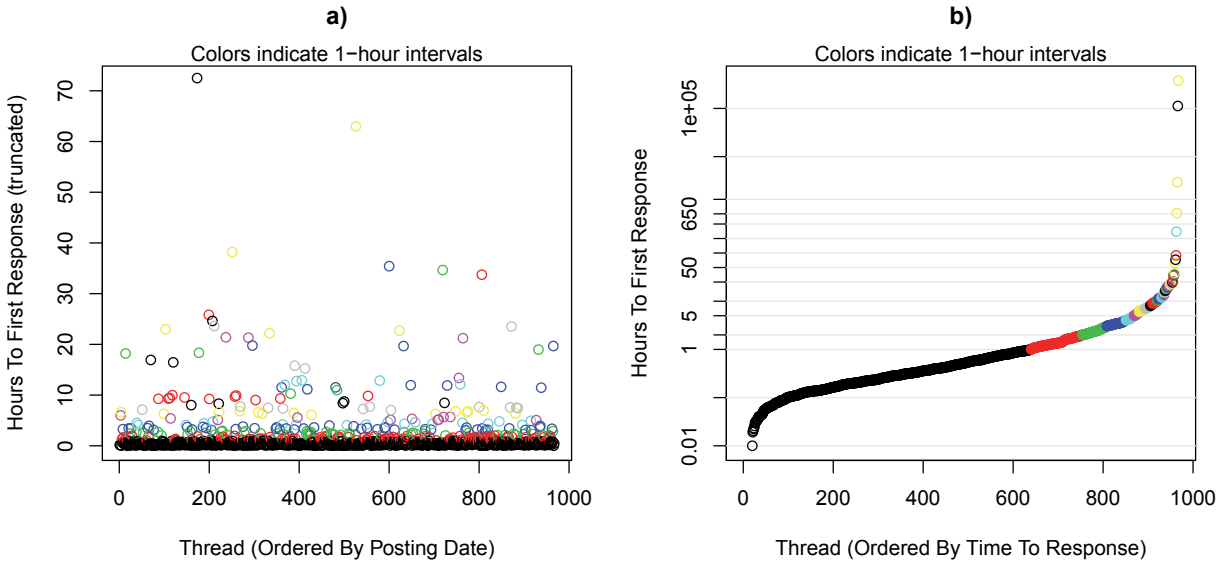
Quantile	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Time To First	0.32	0.42	0.52	0.63	0.75	0.90	1.05	1.30	1.46	1.78

Quantile	55%	60%	65%	70%	75%	80%	85%	90%	95%
Time To First	2.19	2.75	3.82	5.08	7.76	10.87	18.03	26.91	91.84

Figure 4.17: Presenting the time to first reply on the PPML. The colors are there to differentiate hourly intervals, i.e, the large black cluster at the start of panel b) represent the threads with < 1 hour for response, red for $1-2$ hours, etc...

4.20 presents the SURGINET thread durations.

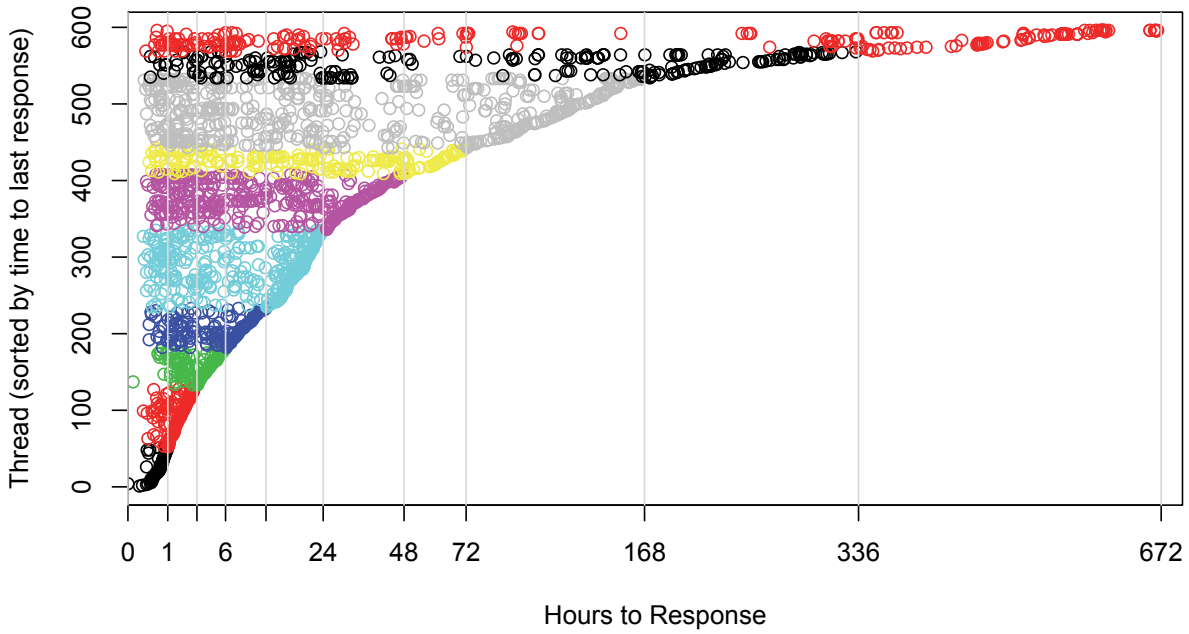
The similarity in attention spans is interesting, as the SURGINET threads have the same durations as the PPML despite having far more messages on average. This pattern suggests a global thread duration, which may be helpful in facilitating KT. If after three days or so a question does not seem resolved it may need additional facilitation work done by the list administrators to ensure that it maintains its space within the community.



Quantile	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Time To First	0.06	0.10	0.13	0.16	0.20	0.23	0.29	0.34	0.41	0.50

Quantile	55%	60%	65%	70%	75%	80%	85%	90%	95%
Time To First	0.61	0.78	0.94	1.24	1.68	2.25	3.26	5.10	9.88

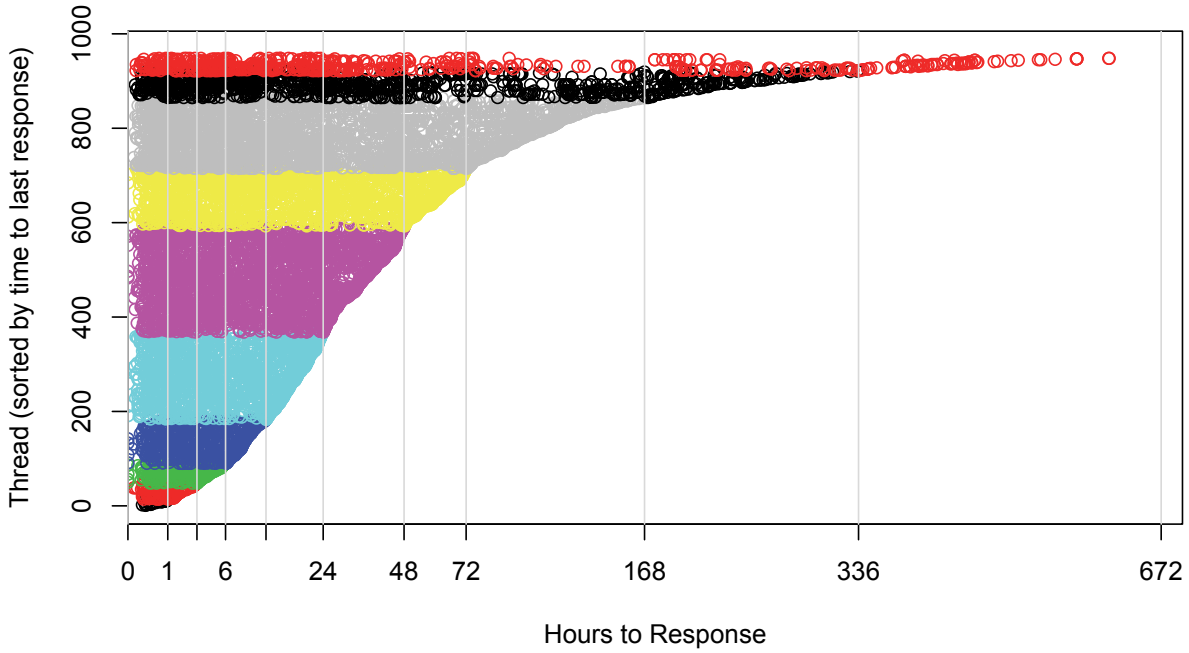
Figure 4.18: The time to first reply for all the threads on SURGINET. The colors are there to differentiate hourly intervals, i.e, the large black cluster at the start of panel b) represent the threads with < 1 hour for response, red for 1 - 2 hours, etc...



Quantile	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Time to Last	0.74	1.08	1.69	2.52	3.81	5.91	8.55	12.78	16.36	20.07

Quantile	55%	60%	65%	70%	75%	80%	85%	90%	95%
Time to Last	23.14	29.07	39.40	55.45	74.72	107.82	137.28	172.55	312.57

Figure 4.19: The responses for each of the threads on the PPML, sorted by time to last response. The colors are only to distinguish time intervals



Quantile	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Time To Last	3.05	6.39	9.00	12.64	15.84	19.09	22.55	24.92	28.64	35.42

Quantile	55%	60%	65%	70%	75%	80%	85%	90%	95%
Time To Last	42.02	47.01	54.09	65.59	74.53	95.14	121.02	174.25	311.77

Figure 4.20: The responses for each of the threads in SURGINET, sorted by time to last response. The colors are only to distinguish time intervals

4.3.3 Conclusion

The collaboration pattern analysis provided detailed insight into the issue of pendants within the PPML (the same methods were applied to SURGINET, but due to their activity levels no pendants were found), but found no evidence of any systemic bias related to either the user or the content of the messages themselves. We have identified potential screening methods (2 days of inactivity) for preventing pendants from occurring. In the response analysis both mailing lists seemed to have fast initial responses and similar, three-day durations, suggesting that the average active lifetime of a thread may be consistent across mailing lists.

4.4 Community Leaders: Individuals and Groups

In order to properly understand the KT nature of a community it is important to identify the community leaders. Centrality metrics will provide the means to identify individuals, and clustering methods (blockmodels and core-periphery analysis) will attempt to find dominant groups and other potential subgroups within the community.

4.4.1 Identifying Leaders: Centrality Measures

Figure 4.21 presents the centrality distributions for the PPML data. With the exception of closeness they are all significantly right-skewed, indicating that there are a couple of users that dominate the centrality measures based on their much higher posting rates. The exception is closeness, which demonstrates that the majority of users can be considered somewhat close to the rest of the community. The strong relationships between the measures are a known attribute of centrality measures; they are highly correlated to one another in most scenarios. The users that are central in terms of degree tend to also be central in terms of betweenness, closeness and coreness.

Table 4.3 presents the top users in each of the 5 centrality categories. As you can see the top users are repeated in each category. These users represent the "power users" in the community: the highly active members that are central to the community because of their activity levels. The centrality measures do not provide much more insight into the community other than the fact that active users are central, but the overall distribution of the centralities nevertheless provides valuable insight into the community as a whole.

The SURGINET centrality measures are presented in figure 4.22 and table 4.4. They

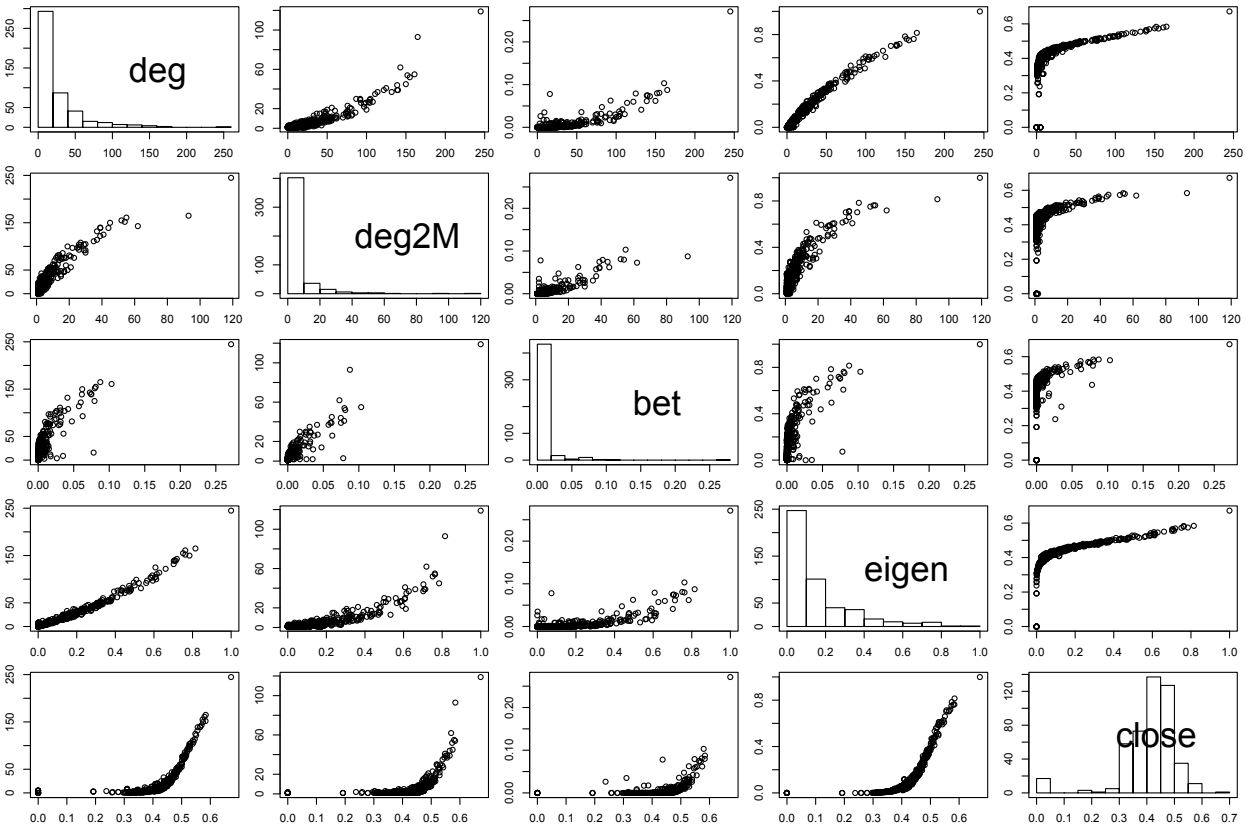


Figure 4.21: The distributions and relationships between PPML centrality measures. Note that the betweenness measures are scaled up by a factor of 1000 to ease presentation, and their true values are minuscule.

present results that are similar to those on the PPML, in terms of correlation between metrics, but there are some differences. The degree and coreness measures are less skewed on SURGINET, indicating a larger group of power users. The other issue of note is the disparity between 2-Mode degree and the number of messages reported in figure 4.16. This is because the network structure of the community records multiple messages to a single thread as a single contribution. This means that users are often contributing multiple messages to the same thread, which was not the case in the PPML. SURGINET users seem to engage in conversations more than their PPML counterparts, who compose longer, more concrete responses to questions rather than engaging in dialogue.

senderID	deg	deg2M	bet	eigen	close
S2509	245	119	0.272	1.000	0.673
S2340	165	93	0.088	0.815	0.584
S2119	143	62	0.073	0.719	0.570
S2100	161	55	0.103	0.762	0.580
S2105	155	52	0.081	0.751	0.573
S2523	152	54	0.080	0.763	0.584
S2111	150	45	0.062	0.784	0.575
S2236	140	39	0.075	0.708	0.565
S2085	139	39	0.062	0.713	0.556
S2122	139	44	0.075	0.703	0.565
S2155	125	41	0.079	0.608	0.548
S2153	122	39	0.060	0.661	0.551
S2176	132	37	0.041	0.702	0.556
S2271	114	37	0.057	0.628	0.546
S2130	111	35	0.031	0.615	0.540
S2198	108	27	0.042	0.592	0.542
S2239	105	30	0.032	0.608	0.531
S2201	104	26	0.037	0.583	0.529
S2225	104	29	0.026	0.606	0.539
S2091	103	27	0.026	0.586	0.527
S2078	99	26	0.030	0.509	0.528
S2410	97	25	0.018	0.565	0.524
S2520	97	22	0.015	0.595	0.529
S2095	100	19	0.024	0.612	0.529
S2566	93	26	0.063	0.496	0.528

Table 4.3: The "top" members of the PPML in terms of centrality indicators

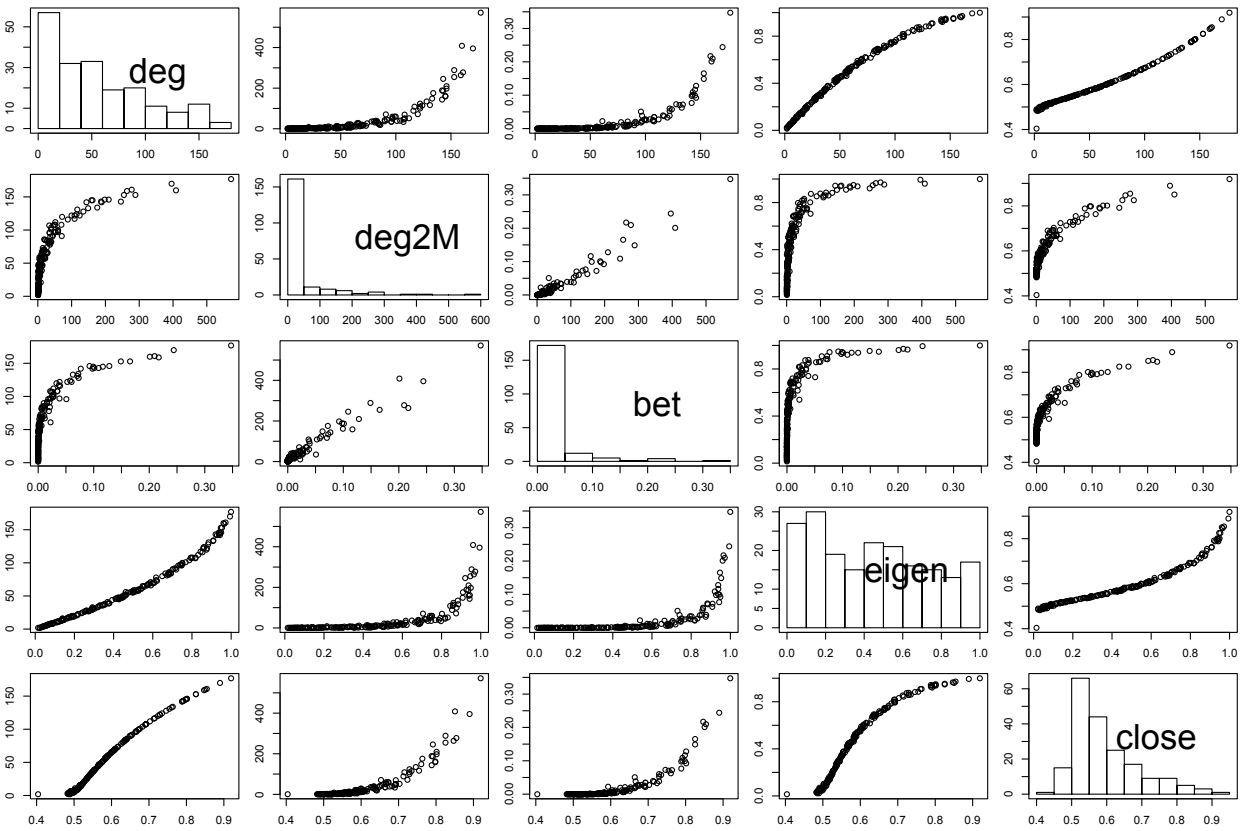


Figure 4.22: The centrality distributions for the members of SURGINET

senderID	deg	deg2M	bet	eigen	close
S0951	177	572	0.348	1.000	0.919
S0952	160	409	0.201	0.961	0.851
S0971	170	396	0.244	0.994	0.890
S0959	153	289	0.149	0.954	0.826
S0989	161	278	0.210	0.972	0.855
S0998	159	264	0.217	0.965	0.847
S0968	153	255	0.166	0.947	0.826
S0970	143	246	0.109	0.921	0.792
S0977	146	210	0.128	0.938	0.802
S0950	146	198	0.093	0.949	0.802
S0992	142	189	0.098	0.928	0.789
S0966	143	187	0.101	0.937	0.792
S0972	134	176	0.072	0.910	0.764
S0956	145	162	0.099	0.945	0.798
S0980	145	159	0.117	0.939	0.798
S0953	133	149	0.063	0.908	0.761
S0957	142	144	0.077	0.943	0.789
S0975	128	135	0.073	0.883	0.746
S0954	122	124	0.060	0.861	0.729
S0993	131	117	0.071	0.885	0.755
S1020	124	115	0.058	0.861	0.735
S0955	117	109	0.038	0.844	0.716
S1022	122	109	0.053	0.858	0.729
S1009	122	100	0.039	0.877	0.729
S0979	116	89	0.040	0.854	0.713

Table 4.4: Centrality measures for the most active 25 users on SURGINET

4.4.2 Knowledge Translation Activity

We identified three archetypes that community members may fill within the community: Knowledge seekers are those that initiate knowledge-based conversations, facilitators are those that encourage conversation, and content experts are the community leaders in terms of expertise in a specific field. Knowledge seekers are the simplest of the three roles to determine, as they are those that initiate the conversations. Counting the number of threads initiated relative to the number of threads participated in can tell us who the most active knowledge seekers are within the community. Figure 4.23 presents the number of threads initiated relative to the number of threads participated for the PPML data, and table 4.5 presents the number and proportion of threads initiated in the 3rd and 4th columns.

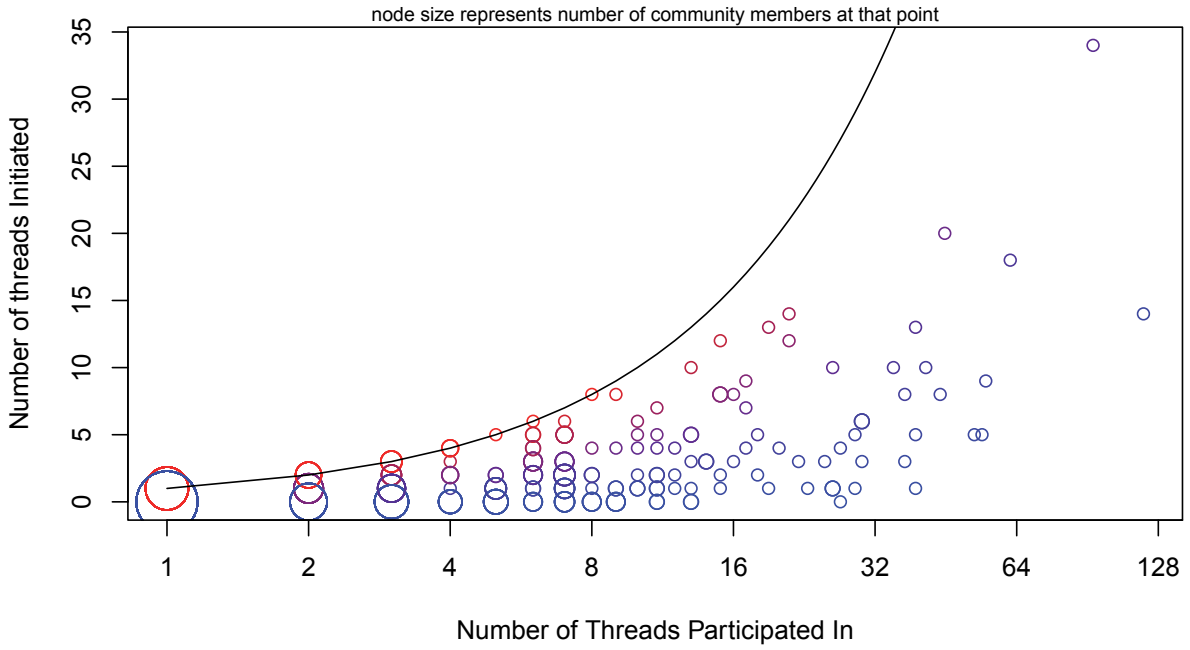


Figure 4.23: The initiation rate relative to the number of threads participated in on the PPML. The line through the plot represents the maximum possible initiation rate, and the colour of each node represents that proportion from blue (low) to red (high).

As the table and figure demonstrate there are some active knowledge seekers within the community. From the table there are several users that have initiated multiple threads without contributing to others, demonstrating a clear pattern of knowledge seeking. At the

ID	Threads	Threads Initiated	Prop. Initiated	First Reply	Avg Reply Position
S2318	8	8	1.000	0	
S2299	5	5	1.000	0	
S2378	9	8	0.889	0	10.667
S2290	15	12	0.800	1	4.000
S2395	13	10	0.769	0	4.714
S2090	19	13	0.684	4	6.300
S2304	21	14	0.667	2	4.700
S2111	45	20	0.444	5	7.892
S2153	39	13	0.333	8	4.034
S2119	62	18	0.290	12	4.259
S2130	35	10	0.286	3	5.727
S2200	18	5	0.278	5	5.188
S2155	41	10	0.244	9	4.200
S2122	44	8	0.182	9	6.980
S2100	55	9	0.164	19	4.629
S2236	39	5	0.128	10	4.811
S2509	119	14	0.118	37	4.862
S2523	54	5	0.093	13	6.552
S2085	39	1	0.026	9	6.500
S2198	27	0	0.000	9	4.971
S2242	13	0	0.000	4	8.167
S2348	13	0	0.000	2	4.385

Table 4.5: A sample of the initiation and reply patterns for some users from the PPML.

opposite end of the table are users that participate a lot without initiating, taking on a more passive role as a content expert. These users contribute to the community when knowledge is needed, but do not initiate conversations themselves.

For the SURGINET users, figure 4.24 presents the initiation rate relative to the overall participation for the SURGINET members. There is a similar pattern to the PPML data, with the exception of a single power user with a huge number of initiations.

Table 4.6 presents a sample of the initiation rates for users. There are some users who have only initiated threads, i.e., only come to the community with questions, while at the other end there are users that have largely responded to threads without initiating many conversations.

There does not seem to be the same density of users that are only knowledge seekers within SURGINET, as there are fewer users that have initiated a high number of threads relative to their overall number, which may indicate that the community is more mature and

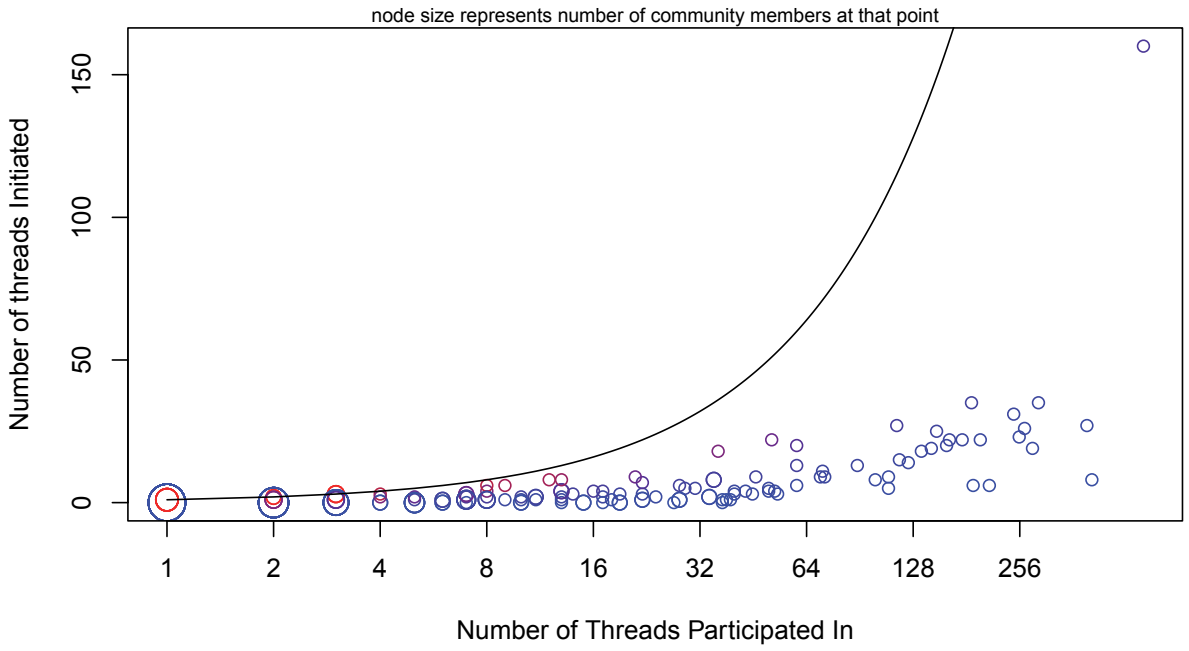


Figure 4.24: The initiation rate relative to the number of threads participated in on SURGINET. The line represents the theoretical maximum.

populated with content experts rather than junior members.

The first reply count and average reply count were investigated to try and identify the facilitators, those that reply early to initiate conversation, and the content experts, those that reply late to end conversations, but no definitive pattern was found. Figures 4.25 and 4.26 presents the distribution of average response position in the conversation per user for the PPML and SURGINET data respectively.

For the PPML there is a pretty clear regression to the mean, with most users varying from the mean due to random error. The reply position is an interesting metric for SURGINET because the community has such fast response times. Given that the community operates in several timezones and surgeons often perform procedures that last for several hours, response position may be less about facilitation and more about availability. The figure presents the average response positions, and demonstrates less regression to the mean than was present in the PPML. Even at the high activity levels there are some users that consistently reply early and others that reply late, but due to the fast response times and the member availabilities

ID	Threads	Threads Initiated	Prop. Initiated	First Replies	Avg Reply Position
S0995	3	3	1.000	0	7.000
S1014	3	3	1.000	1	17.429
S1043	2	2	1.000	0	18.250
S0964	3	3	1.000	2	8.556
S1109	2	2	1.000	0	20.308
S1084	8	6	0.750	0	3.000
S0962	4	3	0.750	0	7.429
S1045	12	8	0.667	0	10.882
S0973	36	18	0.500	3	16.615
S0965	51	22	0.431	1	16.838
S0958	21	9	0.429	0	2.000
S1032	60	20	0.333	5	6.483
S1041	22	7	0.318	30	11.606
S1027	13	4	0.308	1	10.222
S0998	264	26	0.098	3	12.500
S0968	255	23	0.090	32	8.589
S0989	278	19	0.068	0	5.000
S0971	396	27	0.068	3	15.036
S1022	109	5	0.046	79	12.967
S0992	189	6	0.032	130	9.316
S0977	210	6	0.029	0	11.694
S0952	409	8	0.020	2	12.975

Table 4.6: A sample of the initiation and reply patterns on SURGINET

this difference is difficult to attribute to facilitation.

Knowledge seekers can be easily identified in a mailing list in the way they initiate conversations, but facilitators and content experts are more difficult to identify with simple network methods. Content experts will be investigated in the BICGM directed network in section 4.6 below.

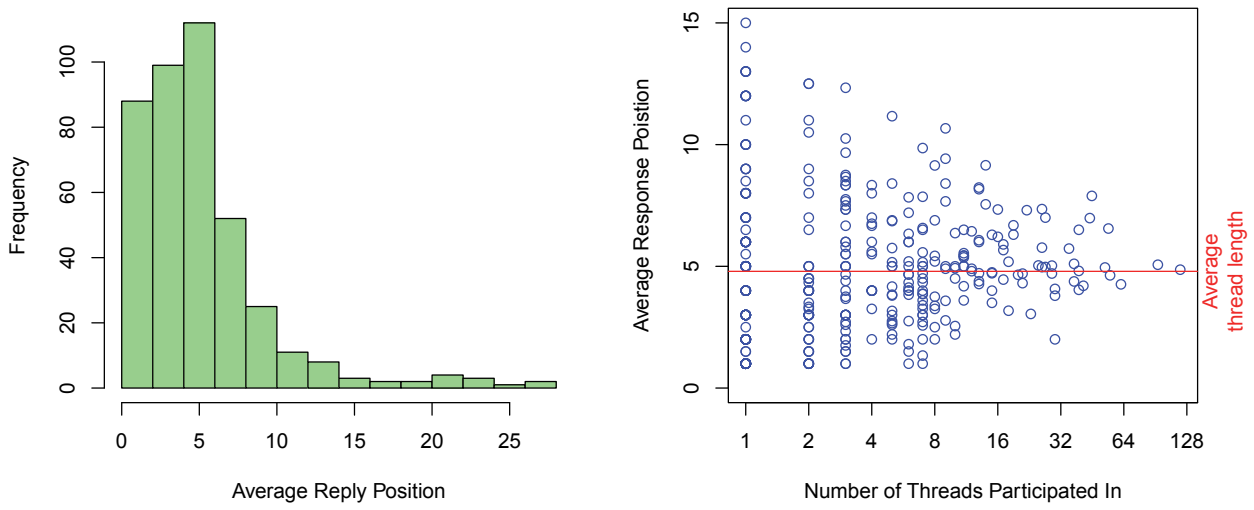


Figure 4.25: The distribution of the average response position per user (left) and the individual values (right) on the PPML.

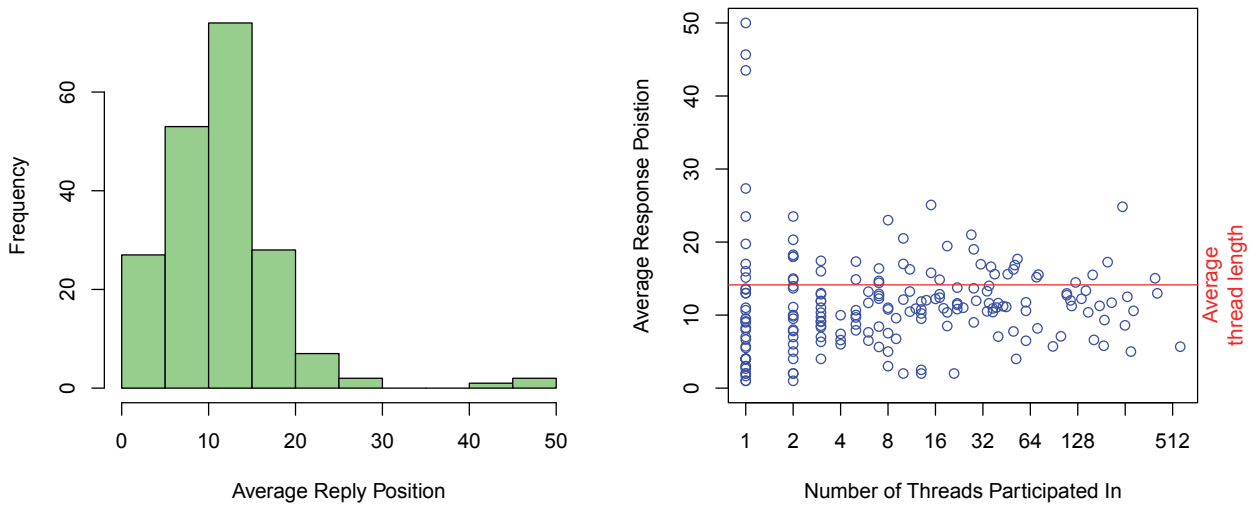


Figure 4.26: The average response position per user (left) and individual values versus participation level (right) on SURGINET.

4.4.3 Identifying Leadership Groups: Connection Clusters

Clustering and core-periphery analysis are both methods designed to find subgroups of users within the community. The objective is to identify potential groups of power users within the community based on their shared communication ties.

PPML Connection Clustering The clustering (using Ward's Distance) of the shared threads between PPML users is presented in figure 4.27. Looking at the cluster it appears that there is a large, inactive group and a small active group. The active group (far right) can be split into 3 subgroups, and the inactive group may have a subsection that is slightly more active. The DS_1 or DS_2 splits seem to be the most optimal.

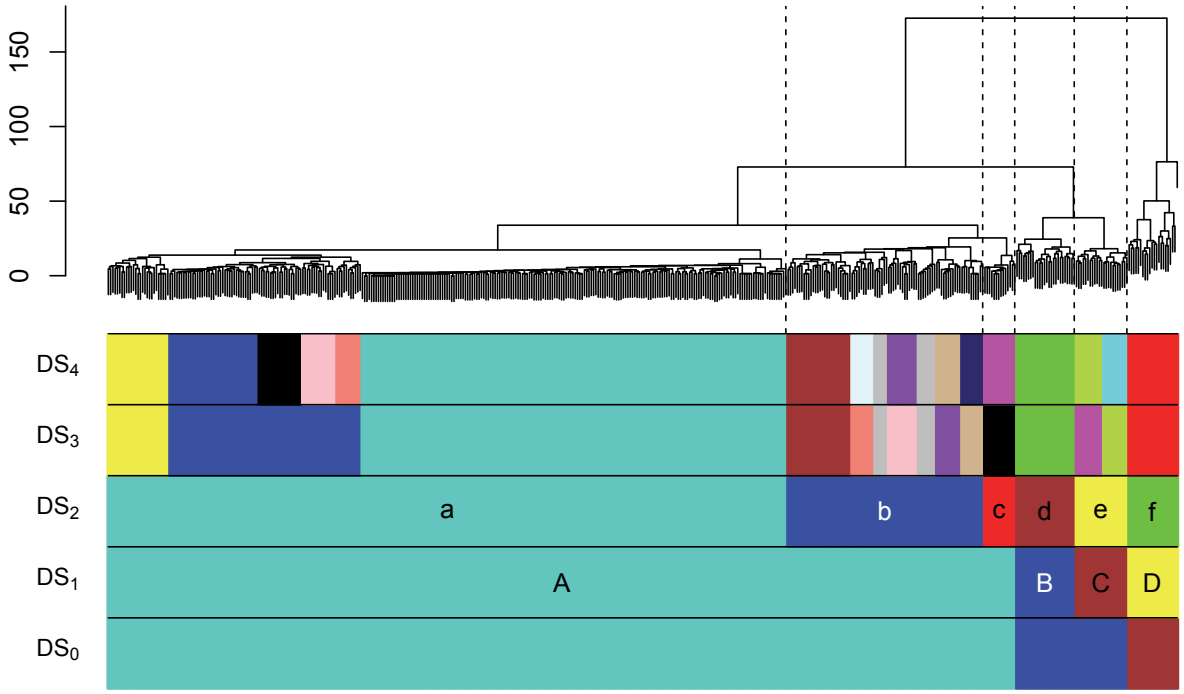


Figure 4.27: Results of clustering the PPML 1-mode user co-occurrence matrix using Ward's method

Table 4.7 presents the densities of each of the clusters in terms of average number of threads shared between the users. From the densities it appears that the smallest cluster (D/f depending on the clustering method) is a cluster of the most active users, with not only the highest in-cluster average, but with more shared threads with other clusters than their

DS_1

	A (397)	B (26)	C (23)	D (22)
A (397)	0.02	0.10	0.06	0.24
B (26)	0.10	0.75	0.30	1.80
C (23)	0.06	0.30	0.56	1.30
D (22)	0.24	1.80	1.30	5.06

DS_2

	a (297)	b (86)	c (14)	d (26)	e (23)	f (22)
a (297)	0.01	0.01	0.01	0.05	0.04	0.13
b (86)	0.01	0.10	0.03	0.26	0.12	0.60
c (14)	0.01	0.03	1.00	0.16	0.20	0.33
d (26)	0.05	0.26	0.16	0.75	0.30	1.80
e (23)	0.04	0.12	0.20	0.30	0.56	1.30
f (22)	0.13	0.60	0.33	1.80	1.30	5.06

Table 4.7: The average number of shared threads in each PPML cluster from figure 4.27

own in-cluster averages. This is in concordance with other findings that the PPML does not have any significant segmentation, but is rather dominated by a group of super users that respond to a majority of threads.

The next step in studying the content clusters is to investigate the terms being used within them using term presence. If you look at the connection-based clustering of the PPML in figure 4.27 there is some significant overlap between the DS_1 and DS_2 clusters. The large cluster A has been split into three sub-clusters a, b, c while the other three clusters are the same in both divisions. Table 4.8 presents the term presence and most common terms in cluster A and how those values break down in the smaller clusters, while table 4.9 presents the same values for the other three clusters in both methods.

The split of cluster A into sub-clusters a, b, c are not overly informative. The larger cluster a again seems to be a catch-all cluster much like A is, with low term values and no general direction. Cluster b may be of interest, as the terms within it demonstrate some interest in pain management, including discussion of care facilities, Complex Regional Pain Syndromes (CRPS), chronic pain and helping, but that may be a stretch. Cluster c is small (only 14 users) and demonstrates one of the risks of SNA clustering. The 14 users within the cluster all participated in a single conversation, a thread entitled "Withdrawal tools and Weaning Protocols", but beyond that specific thread there is no real pattern in the messages. With SNA clustering a single, large thread amongst users without a lot of other communication

(a)

A	a	b	c
1.74	1.50	3.03	3.18
1.76	1.52	3.11	3.25
1.84	1.53	3.12	3.30
1.91	1.54	3.14	3.64
1.93	1.61	3.14	3.76
1.94	1.67	3.26	3.98
2.05	1.71	3.28	4.37
2.08	1.80	3.34	4.53
2.10	1.80	3.37	5.08
2.10	1.85	3.62	12.70

(b)

A	a	b	c
Medication	Clinical Trials as Topic	Thinking	Fever
Chronic Pain	Pain	Helping Behavior	Analgesics, Opioid
Hospitals	Pain Management	CRPS	Ventilators, Mechanical
Work	Work	Work	Cells
Thinking	Hospitals	Child	Proline
Pain Management	Thinking	Pediatrics	Happiness
Pediatrics	Pediatrics	Ambulatory Care	Anemia, Sickle Cell
Child	Helping Behavior	Pain Management	Interleukins
Helping Behavior	Child	Methods	Gene Library
Patients	Patients	Chronic Pain	Weaning

Table 4.8: The most common terms in the PPML user connection clusters A and a, b, c

(no user in that cluster has > 11 messages) can result in users clustering around a specific thread, and can falsely find a meaningful cluster. There is not a cluster of users related to "weaning" on the PPML, but rather a single discussion of "weaning" along with a number of other unrelated messages.

The other three clusters present somewhat separated contents. Cluster B/d seems to be based around pain management, with 7 terms related specifically to pain management drugs. Cluster C/e is related to injections (issues of phlebotomy, needle stick injuries, veins, TIPS) and some drug issues. Cluster D/f is harder to define, with issues related to pain measurement (including faces scales, weights and measures and publishing) along with general pain issues. These three clusters are the three most active clusters of users (see table 4.7) suggesting that stability of the content of the connection clusters is dependent on a reasonable level of activity.

(a)

B/d	C/e	D/f
7.49	6.85	11.82
7.81	6.98	12.00
8.08	7.17	12.01
8.28	7.18	12.15
8.69	7.22	12.19
8.96	7.33	12.21
9.07	8.13	12.52
9.08	8.41	12.78
9.40	8.89	12.79
9.53	11.59	12.97

(b)

B/d	C/e	D/f
Bandages	Drug Tolerance	Aptitude
Hydromorphone	Phlebotomy	Sleep
Fentanyl	Methods	NSAIDs
Infant	Veins	Face
Analgesics, Opioid	Hospitals, Pediatric	Weights and Measures
Epidermolysis Bullosa	Buffers	TIPS
Clonidine	Needlestick Injuries	Single Person
Morphine	Lidocaine	Blood
Ketamine	Weights and Measures	Publishing
Methadone	TIPS	Pain Measurement

Table 4.9: The most common terms in the PPML connection clusters B, C, D and d, e, f

SURGINET Connection Clustering For the SURGINET data Ward's distance was also used to cluster the 1-mode network data, and the clusters are presented in figure 4.28.

Once again there seems to be a large, sparse group of users, along with 2-4 smaller, tighter clusters (depending on the clustering depth chosen). Table 4.10 presents the densities of the clusters, and results in the same, strong core that was present in the PPML, with much higher density disparities than before.

The next step is to investigate the content of the SURGINET clusters. Table 4.11 presents the content of each of the four clusters in DS_3 along with the average TF-IDF scores of those terms. Cluster d is interesting, in that several of the terms seem to be related specifically to hernias of the stomach and gastrointestinal tract, while three (Walking, Dreams and Carbidopa, a drug) seem to be completely unrelated. Further investigation reveals that the members of this cluster were all participants in a couple of threads about hernia surgeries

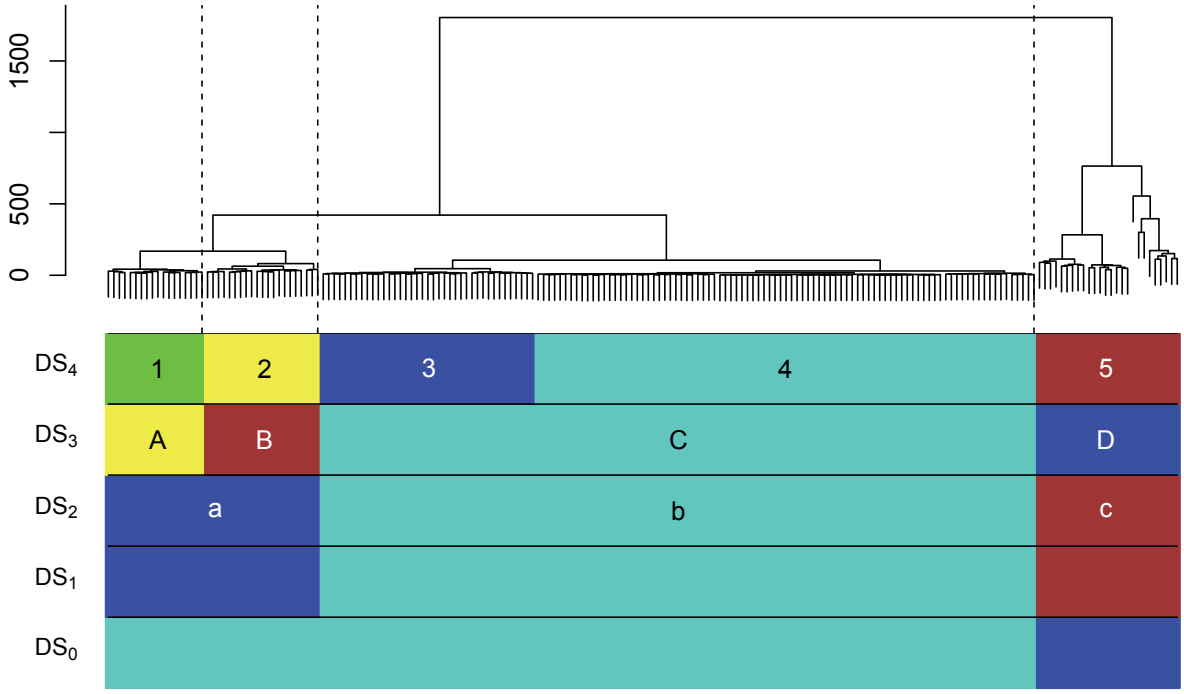


Figure 4.28: Results of clustering the 1-mode user co-occurrence matrix for the SURGINET data

(a common subject on the mailing list), and a single thread about Carbidopa and its side effects, which include trouble sleeping.

The other interesting cluster is *b*, which has several terms unrelated to medicine (Unemployment, Poverty, Running, Reading and Homosexuality), reflecting the sometimes non-medical nature of the SURGINET conversations. The users in this cluster seemed to participate disproportionately in these long, non medically relevant conversations.

DS_3

	a (18)	b (21)	c (130)	d (26)
a (18)	0.97	2.08	0.20	7.26
b (21)	2.08	3.53	0.39	13.60
c (130)	0.20	0.39	0.04	1.36
d (26)	7.26	13.60	1.36	49.52

DS_2

	A (39)	B (130)	C (26)
A (39)	2.26	0.31	10.68
B (130)	0.31	0.04	1.36
C (26)	10.68	1.36	49.52

Table 4.10: The average number of shared threads in each SURGINET cluster from figure 4.28

(a)

a	b	c	d
4.75	6.15	1.30	11.02
4.76	6.16	1.30	11.04
4.81	6.17	1.33	11.09
4.87	6.19	1.34	11.12
4.92	6.33	1.40	11.18
4.94	6.39	1.45	11.26
4.97	6.65	1.47	11.29
5.01	6.78	1.52	11.43
5.08	6.79	1.55	11.51
5.27	7.26	1.55	11.55

(b)

a	b	c	d
Thinking	Internet	Emotions	Duodenum
Wound Healing	Neoplasm Metastasis	Histidine	Omentum
Fluorides	Unemployment	Work	Needles
Drainage	Poverty	Sutures	Ileus
Work	Running	Breast	Hernia, Inguinal
Fasting	Appendix	Patients	Fistula
Hospitals	Laparotomy	Helping Behavior	Carbidopa
Comprehension	Reading	Cholecystectomy	Walking
Eating	Homosexuality	Abdomen	Dreams
Paper	Appendicitis	Thinking	Herniorrhaphy

Table 4.11: The most popular terms in each of the four SURGINET connection clusters. (a) has the highest scores, and (b) has the corresponding terms

4.4.3.1 Coreness

Cluster analysis on the shared threads on both mailing lists revealed a single strong, connected core followed by a set of weakly connected periphery groups. Core-periphery analysis looks for a single, highly connected core group of users and a large set of secondary users that contribute much less to the community.

The coreness measure can be thought of as a centrality measure, so it is presented as eigenvector centrality in table 4.3 and figure 4.21 for the PPML, and table 4.4 and figure 4.22 for SURGINET. These coreness measures can be used to re-arrange the shared threads network as in figure 4.29 for the PPML data and figure 4.30 for the SURGINET data.

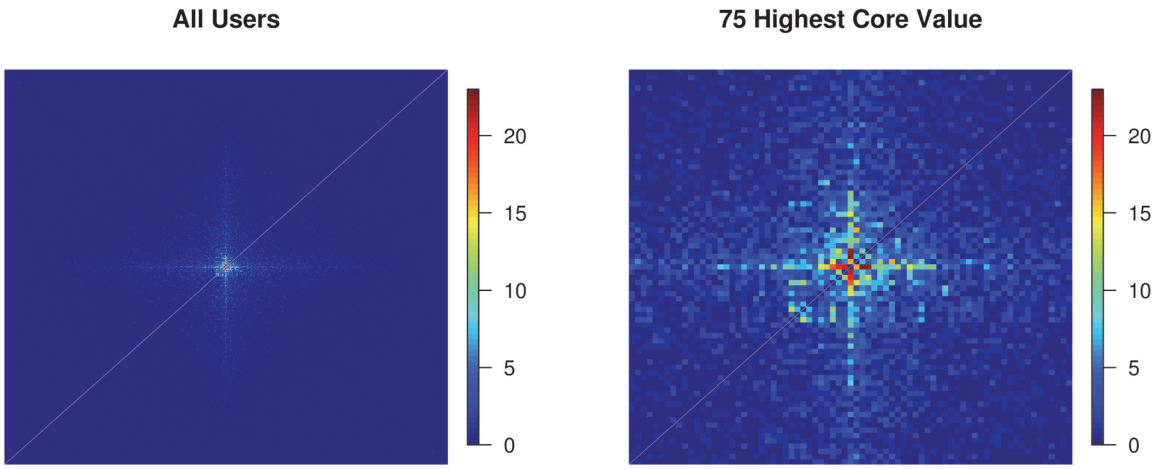


Figure 4.29: A plot of the shared threads in the PPML 1-mode network. This heatmap uses the shared threads re-arranged by coreness (with the most core members in the center). The left plot is all users, and the right is zoomed in on the center

The coreness plots demonstrate that the majority of users have not had interactions with one another on threads. The interaction seems to occur between a tight group of users at the center of the community. One interesting finding from the PPML is that the users with the highest coreness measures branch out father into the community, giving the center of the heatmap its "star" shape rather than the more circular form visible in the SURGINET data.

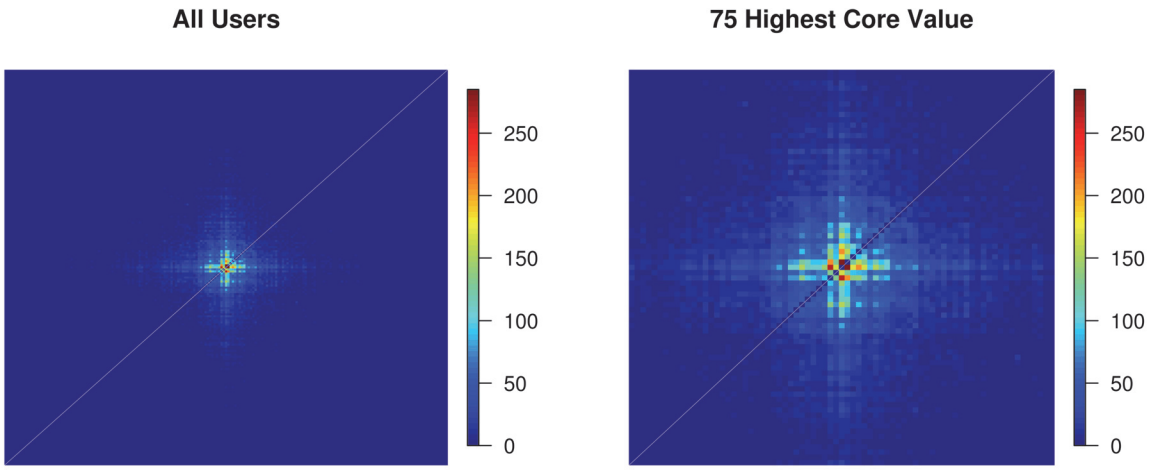


Figure 4.30: Shared threads between users of SURGINET. The zoomed-in version is of the 75 most active users

The actual assignment of a "core" group of users is difficult, as the eigenvector centrality did not produce a clear gap in the users. Figure 4.31 and table 4.12 present four potential cutpoints for the PPML data that are made subjectively by looking at where the divide is within the data. The procedure is done for the SURGINET data in figure 4.31 and table 4.13, again somewhat subjectively, and even more difficult to differentiate.

Cutoff	n	messages	% Messages	threads	% Threads
0.745	6	418	0.150	290	0.420
0.695	11	639	0.229	373	0.540
0.650	12	678	0.243	387	0.560
0.550	24	1026	0.368	460	0.666

Table 4.12: The resulting contribution of the 4 potential cores in the PPML

At its smallest definition the core of the PPML is made up of 6 users, who account for 15% of the messages within the community and have communicated on 42% of the threads. As the core grows the contributions increase proportionally. What is interesting is that even at the largest core only 37% of the messages and 67% of the threads are represented, so though there does not appear to be any communication outside the core in figure 4.29 in reality the majority of the individual messages come from outside this small group.

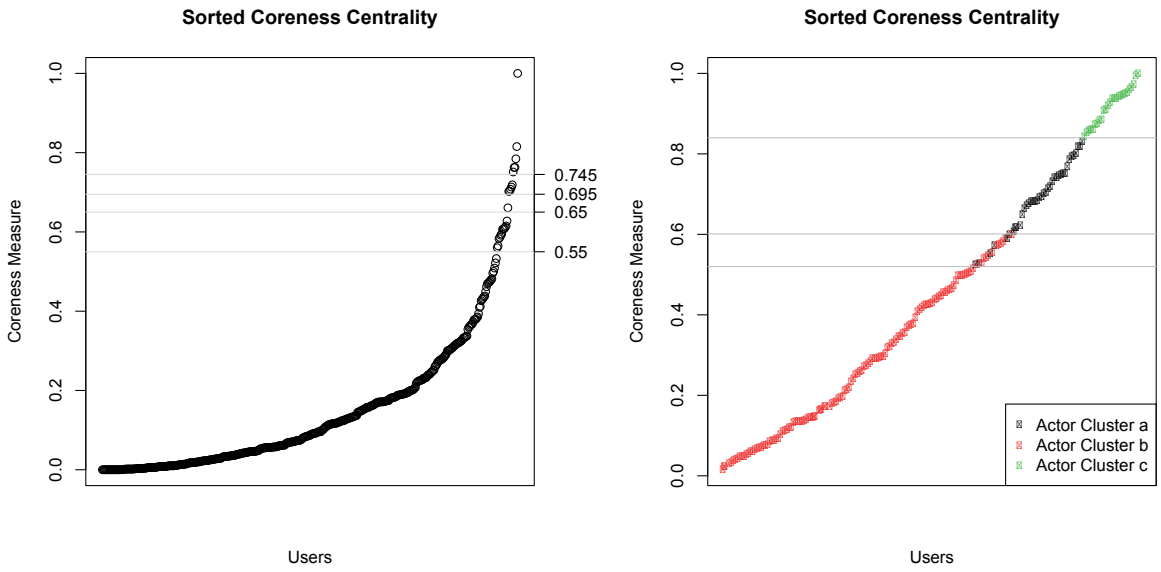


Figure 4.31: The PPML coreness values (left) and the SURGINET coreness values (right), sorted, with cutpoints made somewhat subjectively

cutoff	n	messages	propMessages	threads	propThreads
0.840	26	5253	0.715	939	0.991
0.601	59	6578	0.895	947	0.999
0.520	77	6883	0.936	947	0.999

Table 4.13: The resulting contribution of the 3 potential cores of SURGINET

For the SURGINET users the three cutpoints in the figure are based on the DS_2 clusters from before. The cores of the SURGINET data are much larger, the tightest of which having $\frac{26}{195} = 13\%$ of the community overall. These "core" users represent a significant portion of the communication on the community, and between them have communicated on almost every thread, but their size relative to the overall community is much larger than it was for the PPML.

4.4.3.2 Generalized Blockmodelling

The idea of generalized blockmodeling is that we are not only interested in which users are clustering, but what they are clustering around. The results of the generalized blockmodeling for the PPML are in table 4.14. The idea behind generalized blockmodeling is to permute the user-thread matrix such that blocks of users are either all present (1-block) or all absent (0-block). For the PPML data the generalized blockmodeling found 6 user clusters (with 1

cluster being user S2509, the most active in the community), and 7 thread clusters. The table presents both the density (a) and total counts (b) for the clustering, along with a comparison of the overlap with the DS_2 clustering.

The results of the 2-Mode clustering for the PPML data provide additional support to the 1-Mode clustering. Looking at table (c), 17 of the 22 users from the "super-user" cluster from the 1-mode clustering are in the 2-Mode cluster U3, which suggests that that cluster is again a set of super-users. The most active user, however, was split out into her own cluster. Comparing the communication patterns of U3 to U5, we see that they are both very active in thread cluster T7, but the U3 cluster is much more active in cluster T5, while the U5 cluster is far more active in thread cluster T3. What the algorithm seems to have done is to parse out the high-density cluster by recognizing that the most active user has a slightly different communication pattern than the rest of the core users.

The thread clusters seem to be driven somewhat by the patterns of the users. $T1$ has no contribution from either $U3$ or $U5$ which represents the core of the community. This suggests that the threads in cluster $T1$ are periphery threads. These are short conversations (472 total messages in 236 so exactly two messages per thread) that did not engage any of the central users to the community.

Cluster $T3$ is notable in that all 93 threads have a contribution from the most active user, so the driving force behind that cluster is obvious. As well the last cluster, $T7$ has full contribution from $U5$ and significant contribution from $U3$. It also has 372 messages on 26 threads, or an average of just over 14 messages per thread, which suggests that it is the cluster of very active threads.

There are 6 user clusters and 7 thread clusters, or 42 total cells. Because of some 0-cells the total number is less, but the analysis still requires content summaries of 34 cells, which is a difficult task. Tables 4.15, 4.16 and 4.17 present summaries of the highest scoring terms for all 34 cells.

The summaries of the two-mode clusters do not provide much insight into the community. This was somewhat expected. The 2-mode clusters split the threads across multiple user clusters, which means partitioning the primary knowledge object within the community across multiple user clusters. The network structure that arose out of the two mode data is of interest, with the identification of the periphery threads and the isolation of the most active user, but the contents of the cells is not of any particular interest.

(a) Cluster Densities

	T1(n=236)	T2(n=228)	T3(n=93)	T4(n=55)	T5(n=46)	T6(n=7)	T7(n=26)
U1(n=326)	0.0032	0.0022	0.0033	0.0026	0.0041	0.0000	0.0065
U2(n=48)	0.0087	0.0120	0.0240	0.0180	0.0250	0.0000	0.0580
U3(n=17)	0.0000	0.0770	0.0640	0.0340	0.2000	0.2000	0.3100
U4(n=21)	0.0150	0.0150	0.0230	0.1000	0.0340	0.2200	0.0970
U5(n=1)	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000
U6(n=55)	0.0041	0.0063	0.0059	0.0060	0.0150	0.1600	0.0210

(b) Cluster Participation Counts

	T1	T2	T3	T4	T5	T6	T7	Total
U1(n=326)	247	165	101	47	62	0	55	677
U2(n=48)	98	126	105	47	55	0	72	503
U3(n=17)	0	299	101	32	157	24	136	749
U4(n=21)	74	71	44	121	33	32	53	428
U5(n=1)	0	0	93	0	0	0	26	119
U6(n=55)	53	79	30	18	37	63	30	310
Total	472	740	474	265	344	119	372	2786

(c) Overlap with DS_2 clusters

	U1	U2	U3	U4	U5	U6
a	286	5	0	0	0	6
b	40	22	0	1	0	23
c	0	0	0	0	0	14
d	0	15	0	8	0	3
e	0	6	0	8	0	9
f	0	0	17	4	1	0

Table 4.14: Results of the 2-Mode clustering. (a) presents the densities of each cluster, but (b) is more informative, as densities are skewed lower by cluster size. Part (c) presents the overlap between the 2-Mode clustering of the users and the Hybrid Clustering of the actor network (DS_2 from table 4.7)

Looking just at the columns of the clustering, i.e., the clustering of the threads, however, could provide some insight. Table 4.18 presents a summary of the thread clusters.

The scores in part (a) of the table confirm what had been expected before, with thread clusters $T1$ and $T2$ being catch-all clusters without any legitimate pattern. $T3$ was the cluster that the most active user contributed to every message, and the content of the cluster seems to be based strongly around pain management medications. Clusters $T4$ and $T5$ are the other two thread clusters that were contributed to by the core users, but there does not seem to be any strong pattern to them. Finally, cluster $T7$ is the one that seems to be centred around long threads, and the contents are quite general, with terms that are very pertinent to almost any pediatric pain scenario. One interesting finding is that the term *Safety* is very prominent, which has not been the case in any previous summary of users or threads.

	T1	T2	T3	T4	T5	T6	T7
U1	2.20	1.95	2.42	3.18	2.38		2.98
	2.28	2.00	2.43	3.27	2.50		3.14
	2.64	2.44	2.63	3.84	2.81		3.18
	2.78	2.44	2.69	4.21	2.95		3.55
	2.92	2.71	3.00	4.29	2.96		4.23
	3.26	3.10	3.20	4.44	3.21		4.31
U2	1.81	2.09	2.42	2.94	2.52		2.79
	2.01	2.14	2.50	2.94	2.61		2.81
	2.10	2.31	2.64	2.97	2.63		2.90
	2.18	2.36	3.14	3.39	2.75		2.94
	2.63	2.39	3.25	3.64	2.94		3.27
	2.68	3.31	3.33	3.65	3.15		4.03
U3		1.86	2.37	2.31	2.23	5.62	2.32
		1.91	2.46	2.35	2.27	6.02	2.50
		2.13	2.79	2.51	2.36	6.38	2.62
		2.55	2.94	2.54	2.42	7.05	2.97
		2.67	3.24	2.66	2.65	7.71	2.98
		2.96	3.72	2.78	2.94	8.00	3.93
U4	2.49	2.44	2.79	2.13	3.28	3.81	2.74
	2.51	2.74	3.07	2.23	3.33	4.51	2.80
	2.53	2.77	3.29	2.59	3.40	4.52	2.95
	3.01	2.84	3.33	2.94	3.45	4.55	3.21
	3.16	2.88	3.98	2.95	3.63	5.08	3.51
	3.55	3.00	4.20	2.99	5.22	7.41	4.25
U5			2.91				4.37
			3.08				4.62
			3.33				4.91
			3.40				5.12
			3.61				5.33
			3.74				5.83
U6	2.29	2.09	2.64	3.14	2.98	3.44	3.49
	2.39	2.15	2.78	3.32	2.98	3.66	3.80
	2.56	2.23	3.45	3.64	3.02	4.12	3.81
	3.01	2.29	3.80	3.71	3.07	4.27	4.02
	3.26	2.42	3.87	4.56	3.11	4.82	4.07
	3.26	2.50	4.46	4.89	3.66	5.51	4.15

Table 4.15: The highest TF-IDF values for each user-thread cell in the PPML 2-Mode clustering

	T1	T2	T3	T4
U1	Adolescent	Pain Management	Thinking	Methods
	Work	Child	Patients	Child
	Helping Behavior	Pediatrics	Erythromelalgia	Attitude
	Child	Helping Behavior	Analgesics, Opioid	Helping Behavior
	Pain	Patients	Morphine	Videotape Recording
	Pediatrics	Pain	Helping Behavior	Needles
U2	Analgesics, Opioid	Methadone	Morphine	Pain
	Patients	Thinking	Methadone	Videotape Recording
	Methods	Pain	Patients	erythritol anhydride
	Midazolam	Hospitals	Pain	Nasopharynx
	Pain	Analgesics, Opioid	Ketamine	Helping Behavior
	Pediatrics	Patients	Thinking	Codeine
U3		Child	Pain Management	Child
		Pain Management	Pain	Pediatrics
		Helping Behavior	Thinking	Face
		Pain	Analgesics, Opioid	Pain Measurement
		Analgesics, Opioid	Patients	Pain
		Patients	gabapentin	Codeine
U4	Child	Analgesics, Opioid	Pain	Infant
	Infant, Newborn	Patients	Analgesics, Opioid	Internet
	Research	Org. and Admin.	Morphine	Child
	Pain	Pain	Pain Management	Helping Behavior
	Internet	Pediatrics	Constipation	Pain
	Pediatrics	Medication	Patients	Pediatrics
U5			Pain Management	
			Lidocaine	
			Clonidine	
			Thinking	
			Therapeutics	
			Analgesics, Opioid	
U6	Internet	Pediatrics	EMLA	headline
	Relate	Gene Library	Solutions	Hospitals, Pediatric
	Syringes	Pain Measurement	Chest Tubes	Garbage
	Analgesics, Opioid	Patients	Hearing	Health Facilities
	Pain Management	Medication	Clonidine	Videotape Recording
	Volition	Pain	Pain Management	Stress, Psychological

Table 4.16: The most common terms in each of the PPML 2-Mode clusters (for thread clusters 1-4)

	T5	T6	T7
U1	Child		Pediatrics
	Pain Management		Pain Management
	Cells		CRPSs
	Patients		Spinal Puncture
	Ondansetron		Methods
	Hospitals		Epidermolysis Bullosa
U2	Pruritus		Patients
	Patients		Pediatrics
	Child		Codeine
	physiology		CRPSs
	Pain		Chronic Pain
	Cold Temperature		Child
U3	Pain	EMLA	CRPSs
	Helping Behavior	Needlestick Injuries	Child
	Work	Pain Measurement	Pain Management
	Electromagnetic Radiation	TIPS	Patients
	TIPS	Sleep	Spinal Puncture
	Patients	Weights and Measures	Methods
U4	Pain	Pediatrics	Pain
	Behavior, Addictive	Teaching	Epidermolysis Bullosa
	Clinical Coding	Education	NSAIDs
	Skin Aging	Pain Measurement	Pediatrics
	Pain Management	Sleep	Child
	Infant	Peeling skin syndrome, acral type	Acetaminophen
U5			Recovery Room
			Analgesics, Opioid
			CRPSs
			Defibrillators
			Child
			Acetaminophen
U6	Needlestick Injuries	Needlestick Injuries	Patients
	Electromagnetic Radiation	TIPS	Codeine
	Sleep	Weights and Measures	Spinal Puncture
	Monitoring, Physiologic	Peeling skin syndrome, acral type	Methods
	Naloxone	Pediatrics	Pediatrics
	Respiratory Rate	Weaning	Pain Management

Table 4.17: The most common terms in each of the PPML 2-Mode clusters (for thread clusters 5-7)

(a)

T1	T2	T3	T4	T5	T6	T7
1.60	2.19	3.59	3.09	4.20	8.80	6.79
1.62	2.20	3.63	3.14	4.30	9.15	6.94
1.67	2.26	3.66	3.16	4.30	9.30	6.96
1.67	2.27	3.76	3.23	4.45	9.35	7.08
1.72	2.33	3.82	3.26	4.55	9.41	7.11
1.72	2.36	3.90	3.59	4.57	10.49	7.18
1.83	2.41	3.93	3.65	4.70	10.51	7.20
1.86	2.70	3.97	3.75	4.79	10.58	7.27
1.92	2.73	4.26	3.86	4.82	10.67	7.58
2.30	2.75	4.41	4.10	5.03	14.52	7.93

(b)

T1	T2	T3	T4
Pain	Therapeutics	Morphine	Clinical Trials
Adolescent	Child	Clinical Trials as Topic	Work
Methods	Pain Management	Ketamine	Attention
Internet	Work	Pain Management	Palliative Care
Pain Management	Medications	Clonidine	Internet
Work	Helping Behavior	Analgesics	Child
Chronic Pain	Thinking	Medications	Infant
Helping Behavior	Analgesics, Opioid	Therapeutics	Helping Behavior
Child	Patients	Analgesics, Opioid	Hospitals, Pediatric
Pediatrics	Hospitals	Thinking	Pediatrics

T5	T6	T7
Hospitals	Hearing	Pharmaceutical Preparations
Child	Lead	Pain Management
Patients	Research	Learning
Infant	Advanced Practice Nursing	Organization and Administration
Pain Management	Ventilators, Mechanical	Publishing
Hospitals, Pediatric	Facial Expression	Thinking
Adult	Health Facilities	Exploratory Behavior
Work	Vital Signs	Hospitals, Pediatric
Education	Hospitals, Pediatric	Lead
Aptitude	Congresses as Topic	Safety

Table 4.18: The most common terms in the thread clusters (b) and their average TF-IDF scores (a) from the PPML 2-mode clustering

SURGINET Generalized Blockmodelling The SURGINET 2-Mode clusters are presented in table 4.19. The 2-mode clustering for the SURGINET found 3 user and 4 thread clusters. The results confirm the previous suggestion of a larger, tight core, with the core users identified from before being split into a small,active group of 7 users and a slightly less active group of 19 users. The thread clusters provide very little insight, with no discernible patterns in which clusters were participating on which threads. When you compare the user-threads from the 2-mode clustering to the dendrogram in figure 4.28 you can see where the split of cluster c most likely occurred, but beyond that the 2-mode clustering does not provide much additional information about the community. Content summaries of the $3 \times 4 = 12$ clusters were pursued, but nothing of interest was obtained, so the summaries are not presented here.

(a)

	T1(n=375)	T2(n=323)	T3(n=175)	T4(n=75)
U1(n=169)	0.0089	0.0110	0.0160	0.0370
U2(n=19)	0.0930	0.1300	0.2100	0.4300
U3(n=7)	0.1700	0.4100	0.6100	0.6500

(b)

	T1(n=375)	T2(n=323)	T3(n=175)	T4(n=75)	Total
U1(n=169)	561	590	481	465	2097
U2(n=19)	665	793	713	619	2790
U3(n=7)	441	931	751	340	2463
Total	1667	2314	1945	1424	7350

(c)

	U1	U2	U3
a	39	0	0
b	130	0	0
c	0	19	7

Table 4.19: Results of the 2-Mode clustering for SURGINET. (a) is the cluster densities, (b) is the cluster counts and (c) compares the user clusters from the 1-mode and 2-mode clusters

4.4.4 Summary

The objective of this section was to evaluate methods for identifying community leaders. The centrality measures provide several simple metrics for evaluating who the most active members in the community are. The clustering, both the 1-mode and 2-mode methods,

were not overly successful, as they failed to find significant clustering beyond the strong core structure that core-periphery analysis identified in both communities. The PPML community has a smaller core than the SURGINET community, but there is also more activity by the periphery. Whether the failure of the clustering methods is due to the methods themselves or due to the lack of existing clusters in either mailing list is not clear.

4.5 Knowledge Based Subgroups and Similarities

The knowledge maps demonstrated how a knowledge-based representation of the messages within an online community can provide insight into the knowledge context of the community. In this section we will use knowledge-based representations of the users and threads to try and identify potential subgroups of knowledge within the community. The Generalized Vector Space Model (GVSM) using the combined semantic and co-occurrence correlation seems to be the most appropriate method for calculating similarity within the PPML. For both the users and the threads in the community similarities can be calculated, which will then be used to try and identify knowledge-based clusters in the data. The correlation matrices will be constructed and then clustered using hierarchical agglomerative clustering. Figure 3.11, part (b) presents the methods. We will investigate a number of parameters at each step, including what link method is most appropriate for the clustering method (single, average, complete or Ward), exploring the components of the dynamic hybrid cut algorithm, specifically cluster size, gap and max core distance, evaluations of the potential clusters using image matrices and silhouette coefficients, and attempting to label the clusters using proportional term contributions.

4.5.1 Content-based Thread Clustering

We investigated the hierarchical clustering using four different distance metrics for each dataset. The results for the PPML are presented in figure 4.32 as dendrograms (the results for SURGINET are similar and were omitted). Each of the methods is accompanied by an agglomerative coefficient that measures the quality of the clustering. From the figure it is clear that single-link clustering is not an option, and average link also did poorly, but both Ward's distance and complete link seem to provide decent clusterings of the data. The decision was made to use Ward's distance as the metric, because its coefficient is better, but investigations into complete link methods were also pursued, and no significant improvements

were found.

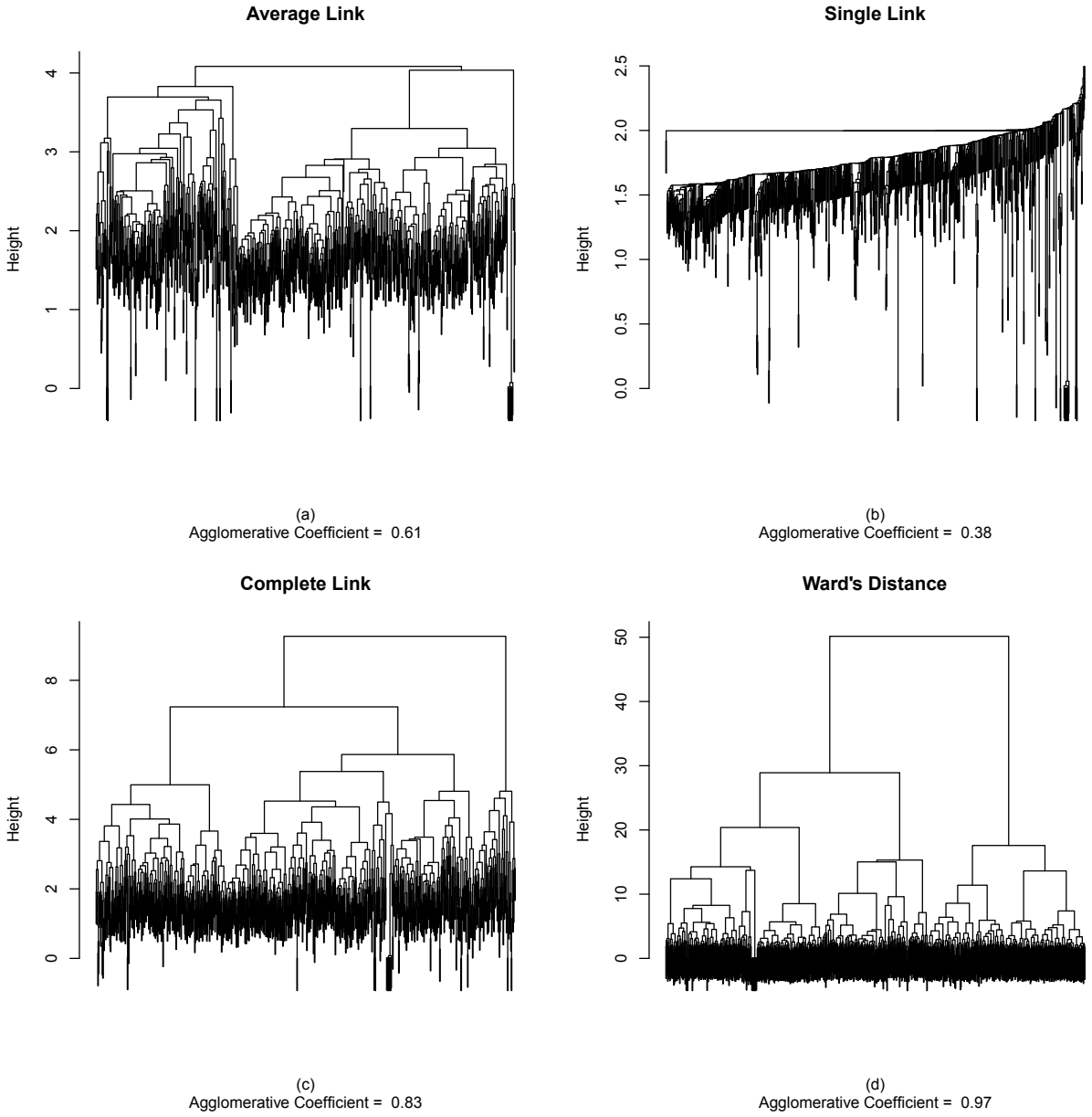


Figure 4.32: Four different approaches to clustering the PPML threads

4.5.1.1 Content Clustering on the PPML Threads

Figure 4.33 presents the results of the cutting of the dendrogram using both static cuts along with 4 different hybrid cuts. As the deepSplit values increase (see table 3.2 for the splitting parameters), the cores are required to be tighter and the allowable merge distance decreases,

resulting in smaller, tighter clusters. This is evident by the size and number of clusters across the 5 levels in figure 4.33.

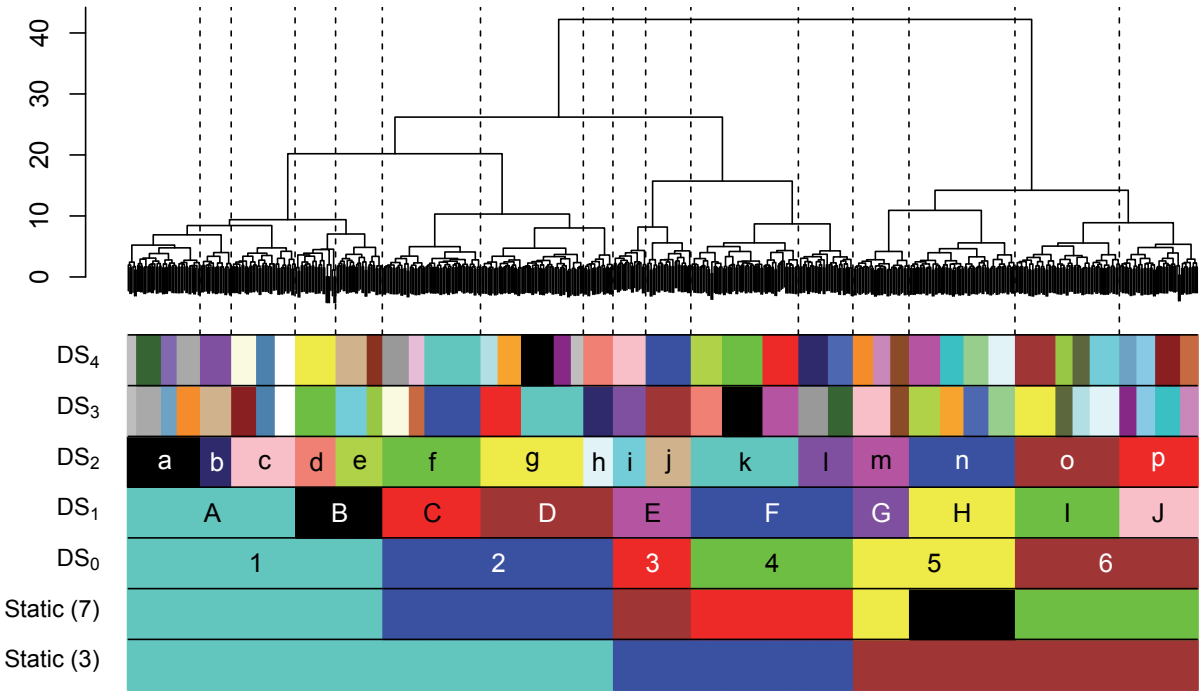


Figure 4.33: The partitioning of the PPML threads along with 7 different cuts of the data

Of the hybrid cuts, DS_0, DS_1 and DS_2 seem to be the best options, as the higher deep splits reduced the data into too many clusters. The DS_0 clustering is almost identical to the static cut with 7 groups, except that 1 group (the first one) was split into two sub-clusters. We can see again when we move from DS_0 up to DS_1 how some of the larger clusters are segmented into 2 or more subclusters, while other clusters (clusters $1, 2$ and 5) were reduced in size, and the same pattern moving from DS_1 to DS_2.

The densities of the clusters from DS_0, DS_1 and DS_2 are given in figure 4.34. For all three figures there seems to be a bipartite partitioning of the data into a high-density set of clusters (bottom-right corner) and a low density corner with a high-density diagonal. None of the splits provide evidence of strong clustering, as for most clusters there is a stronger relationship to the densest cluster ($5/G/m$) than within cluster. This suggest that the core-periphery structure detected in section 3.2.3.2 is a more appropriate structure than here. If we ignore the high-density "core" structure then the within-cluster densities (the diagonals

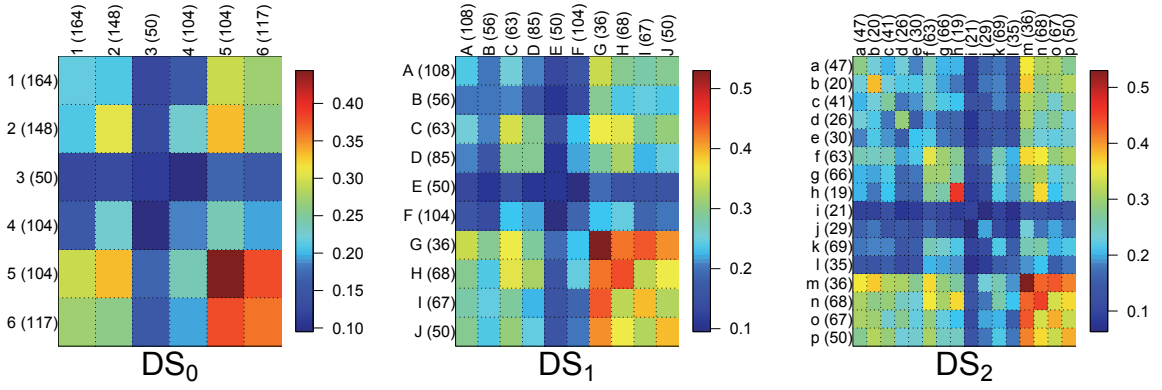


Figure 4.34: The similarity densities of the PPML clusters created using DS_0, DS_1 and DS_2

topleft to bottomright) are a good finding as, for example, we see that cluster C in DS_1 (and its equivalent e in DS_2) are a high density cluster amongst low-density colleagues. Further investigation into the content of these clusters will hopefully reveal similarities in their semantic terms.

The silhouette coefficients for the clusters (see figure 4.35) confirmed our findings, with all clusters having negative silhouette coefficients, demonstrating that most elements are closer to a different cluster than their own. This again demonstrates the core-periphery structure of the threads, with a cluster of large threads dominating the community.

Even though we have detected a core-periphery structure rather than a real clustering, the presence of a comparatively strong diagonal (and thus some within-cluster similarities) suggests that the threads in clusters are more related to each other than to their non-core neighbours. An investigation of the contents of the threads may reveal some suggestion as to why these clusters formed. Table 4.20 summarizes the DS_0 split and table 4.21 does for the DS_1 (DS_2 is omitted as no new interesting patterns were found).

For the DS_0 split all 6 clusters have reasonable term presence values (part (a) of the table). Looking at the actual terms reveals some patterns: cluster 2 seems to be grouped around pain medications, cluster 4 around a similar but different class of pain medications, perhaps related to the management of migranes. Cluster 5 (the densest cluster) seems to center around health care facilities (hospitals, ICUs and facilities) along with some general pain medication issues.

When looking at the DS_1 split we are most interested in where the splitting function divided existing clusters. Cluster 2 was split into clusters C and D. Cluster D seems to have

maintained and infact strengthened its focus on medications, while cluster C seems more interested in management (emotions, theapeutics, trials and organization) while still have some medication presence. Cluster 5 was split into G and H and again seems to have parsed the medication threads from the management threads.

We will ignore the indepth investigation into the DS_2 split. The resulting clusters are very small, which produces some unexpected and largely meaningless results, with the only meaningful clusters having been the same as those produced by the DS_1 split.

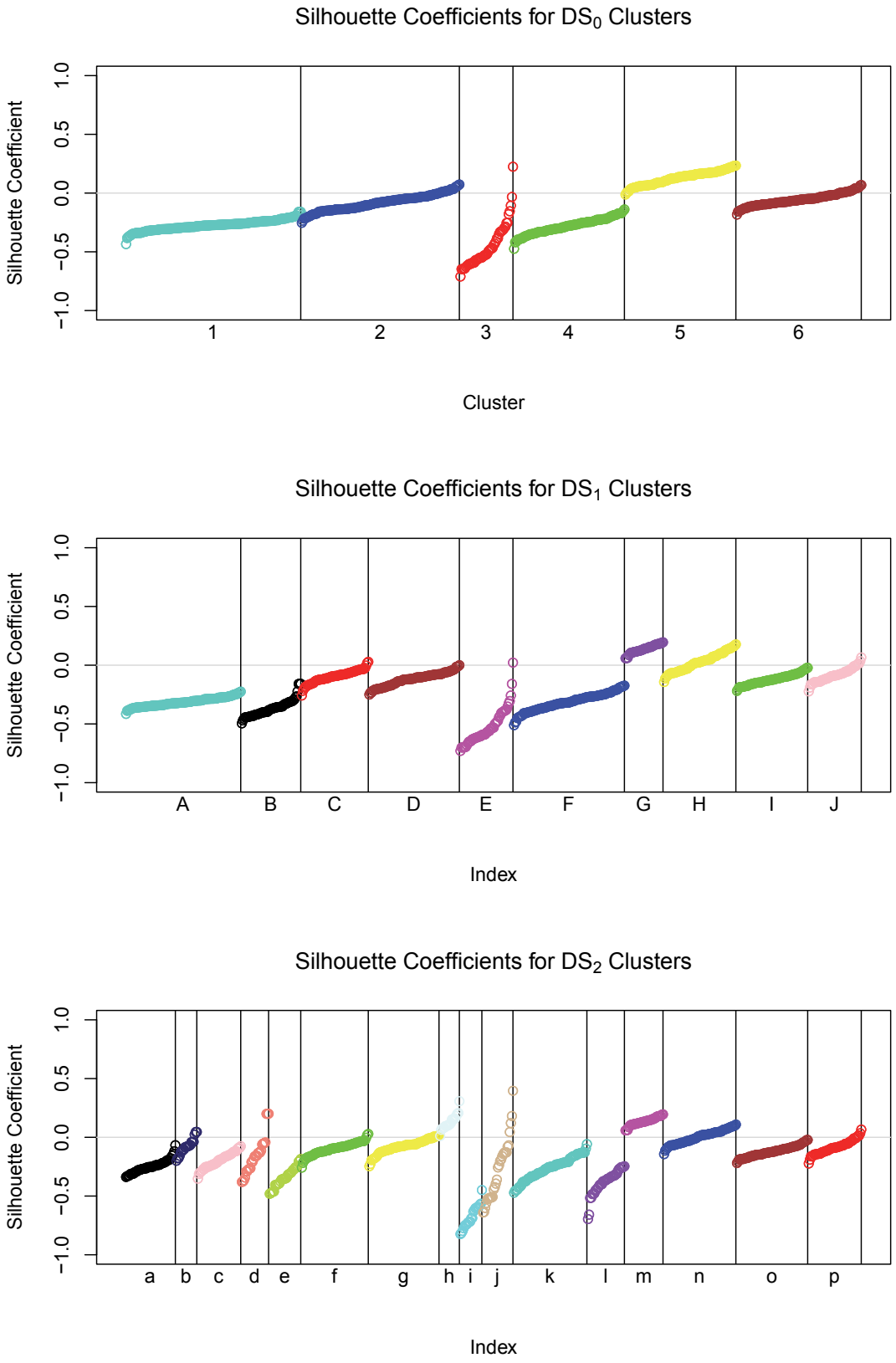


Figure 4.35: The Silhouette coefficients for the three potential thread clusterings of the PPML data

(a)

1	2	3	4	5	6
1.86	4.71	1.77	2.13	4.54	3.08
1.93	4.78	1.90	2.17	4.56	3.22
2.02	4.83	2.14	2.20	4.94	3.27
2.03	4.83	2.30	2.24	4.96	3.32
2.15	5.07	2.96	2.43	4.99	3.55
2.52	5.17	3.29	2.46	5.06	3.72

(b)

1	2	3
Child	Clinical Trials as Topic	Delivery, Obstetric
Pain Clinics	Methadone	Pediatrics
Hospitals	Anxiety	Computers
Helping Behavior	Therapeutics	Copying Processes
Research	Ketamine	Personality Disorders
Methods	Thinking	Disclosure
Volition	Pharmaceutical Preparations	Internet
Internet	Clonidine	Congresses as Topic
Pain Measurement	Analgesics, Opioid	Human Body
Pediatrics	Morphine	Postal Service

4	5	6
Neuralgia	Thinking	Helping Behavior
Diagnosis	Aptitude	Pain Management
Lidocaine	Work	Internet
Migraine Disorders	Morphine	Research
Wounds and Injuries	Health Facilities	Work
Patients	Intensive Care Units	Hospitals, Pediatric
Abdominal Pain	Pain Management	Education
Analgesics, Opioid	Pharmaceutical Preparations	Pediatrics
Headache	Hospitals	Infant
Acetaminophen	Analgesics, Opioid	Child

Table 4.20: The TF-IDF scores for the highest scoring terms in each PPML cluster (a) and the terms themselves (b) for the DS_0 thread cut

(a)

A	B	C	D	E	F	G	H	I	J
2.30	1.87	5.45	4.20	1.77	2.13	6.85	4.66	3.68	4.64
2.30	2.52	5.53	4.23	1.90	2.17	6.90	4.94	3.70	4.64
2.36	2.61	5.57	4.33	2.14	2.20	7.07	5.00	3.76	4.87
2.41	2.65	5.59	4.90	2.30	2.24	7.37	5.34	3.76	5.01
2.47	2.73	5.69	5.06	2.96	2.43	7.57	6.41	3.94	5.44
2.57	3.08	6.32	6.21	3.29	2.46	7.61	6.49	4.23	5.66

(b)

A	B	C
Pain Clinics	Patients	Adolescent
Helping Behavior	Pulse	Emotions
Research	Carbon Dioxide	Therapeutics
Health Facilities	Oximetry	Clinical Trials as Topic
Methods	Pain Measurement	Organization and Administration
Pain Measurement	Capnography	Thinking
Weights and Measures	Monitoring, Physiologic	Adult
Pediatrics	Volition	Ketamine
TIPS	Pediatrics	Pharmaceutical Preparations
Internet	Analgesia, Patient-Controlled	Morphine

D	E	F	G
Patients	Delivery, Obstetric	Neuralgia	Aptitude
Neoplasm Metastasis	Pediatrics	Diagnosis	Org. and Admin.
Ketamine	Computers	Lidocaine	Teaching
Methadone	Copying Processes	Migraine Disorders	Self Report
Medication	Personality Disorders	Wounds and Injuries	Hospitals, Pediatric
Thinking	Disclosure	Patients	Pain Management
Morphine	Internet	Abdominal Pain	Chronic Pain
Hydromorphone	Congresses as Topic	Analgesics, Opioid	Weights and Measures
Analgesics, Opioid	Human Body	Headache	Pain Measurement
Clonidine	Postal Service	Acetaminophen	Education

H	I	J
Intensive Care Units	Child	Needles
Ketamine	Interdisciplinary Studies	Child
Running	Education	Pharmacies
Hydromorphone	Research	Floors and Floorcoverings
Methadone	Ambulatory Care Facilities	Methods
Hospitals	Pain Clinics	Equipment and Supplies
Pharmaceutical Preparations	Rehabilitation	Needlestick Injuries
Fentanyl	Pediatrics	Sucrose
Morphine	Chronic Pain	TIPS
Analgesics, Opioid	Internet	Infant

Table 4.21: The TF-IDF scores for the highest scoring terms in each PPML cluster (a) and the terms themselves (b) for the DS_1 thread cut

4.5.1.2 Content Clustering on the SURGINET Threads

As with the PPML, the threads will be clustered using a hierarchical agglomerative approach using Ward's method, and then split using the dynamic method presented in section 3.3.9.2. The results of the clustering and a couple of potential cut points are presented in figure 4.36.

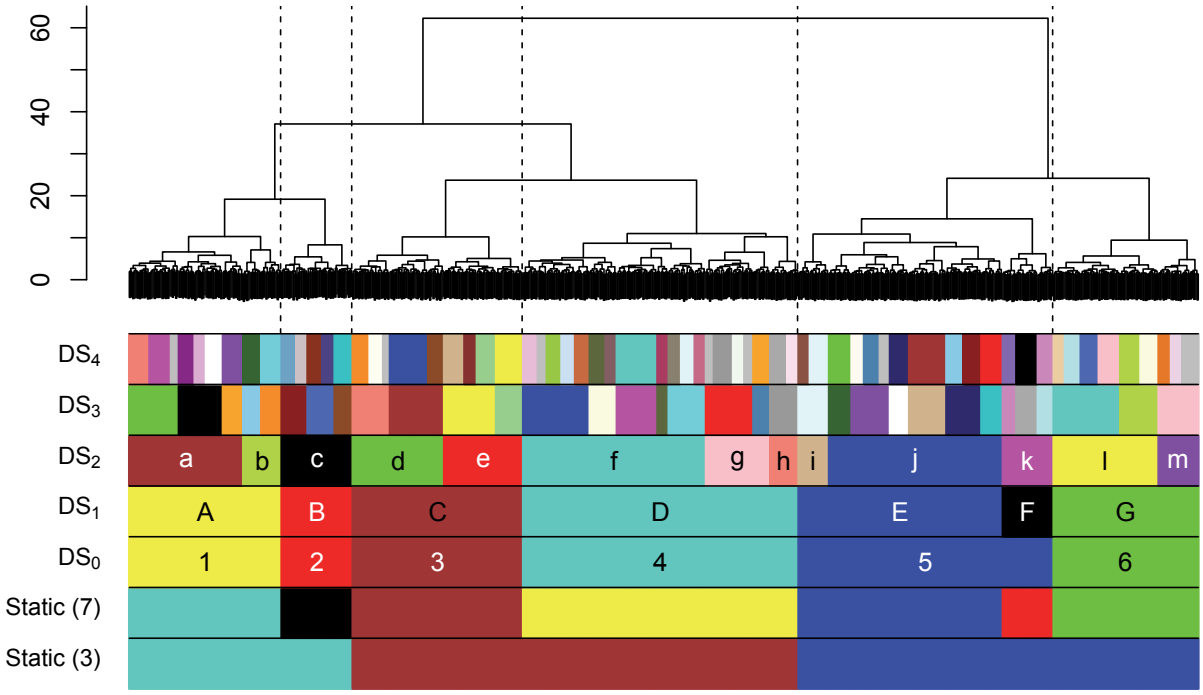


Figure 4.36: The results of the thread clustering on SURGINET, and several different potential cut points.

The clustering presents a 6-cluster partitioning of the data, with a potential split in one of the clusters. The image matrix of the DS_0 split is presented in figure 4.37, and presents a network that has a surprisingly strong core structure, with two thread clusters representing a significant portion of the communication. This is a similar pattern to what was found within the PPML data, but the SURGINET data identified a potential partitioning of the core threads into two separate subgroups.

Figure 4.37 also has the silhouette coefficients for the clustering, and confirms our finding that there are two tight thread clusters within the core that seem to dominate the network.

The contents of these clusters are presented in table 4.22. The focus of the table is clusters 5 and 6, which represent the core of the threads in the community. They seem to

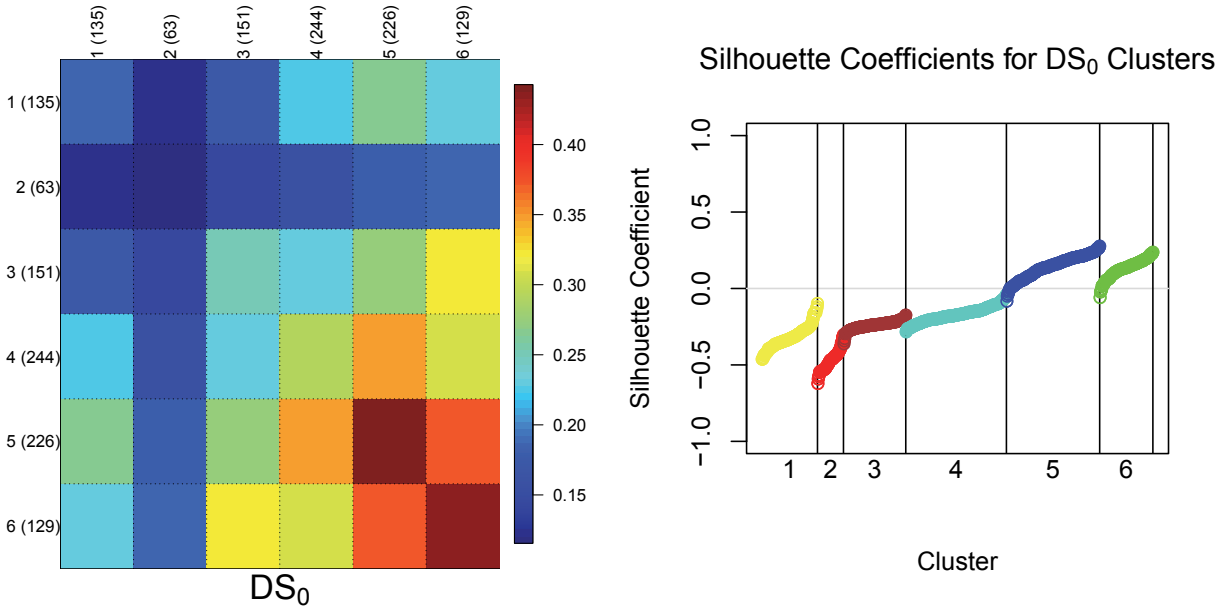


Figure 4.37: The image matrix (left) and silhouette coefficients (right) for the DS_0 cluster of the threads within SURGINET

be grouped around two different issues, with the first interested in laproscopic surgeries and the areas they may be applied to. The second cluster, in contrast, seems to be much more general, and with mappings to emotions, sensation, love and happiness may be focus on more general issues related to surgery than specific surgical procedures.

Both clustering methods identified a core group of threads that dominate the traffic, but within the core the SURGINET data was partitioned into to separate groups, while within the PPML data the periphery threads were slightly more active. This conforms to some of the connection-based user clustering, which suggested that the PPML core of users was smaller that SURGINET, but also that its periphery was more active.

(a)

1	2	3	4	5	6
2.05	1.40	2.28	2.66	4.56	4.46
2.06	1.44	2.33	2.67	4.59	4.51
2.11	1.46	2.50	2.73	4.64	4.82
2.19	1.47	2.74	2.87	4.70	5.12
2.21	1.48	2.82	2.89	5.15	5.40
2.74	1.57	3.03	3.19	5.34	5.41

(b)

1	2	3	4
Disease	Malaria	Running	Pain
Neoplasms	Helping Behavior	Wine	Drainage
Thyroid Gland	Printing	Hospitals	Disease
Sentinel Lymph Node Biopsy	Internet	Paper	Hernia
Patients	Cholesterol	Medication	Helping Behavior
Hemorrhage	Pediatrics	Printing	Clinical Trials as Topic
Medication	Paper	Work	Biopsy
Breast	Medication	Happiness	Diagnosis
Biopsy	Histidine	Internet	Hemorrhage
Mastectomy	Fractures, Bone	Reading	Abdomen

5	6
Appendectomy	Emotions
Cholecystectomy	Hospitals
Bile	Comprehension
Colon	Sensation
Sutures	Love
Laparoscopy	Learning
Pain	Happiness
Intestines	Histidine
Abdomen	Speech
Drainage	Reading

Table 4.22: The highest MeSH scores (a) and the MeSH terms (b) for the 6 thread clusters in SURGINET

4.5.2 Content-based User Clustering

As with the thread clustering we investigated four different distance metrics, and once again Ward's distance was the most appropriate.

4.5.2.1 Content Clustering on the PPML Users

The splitting dendrogram is available in figure 4.38, done using both static cuts and the dynamic tree cut method.

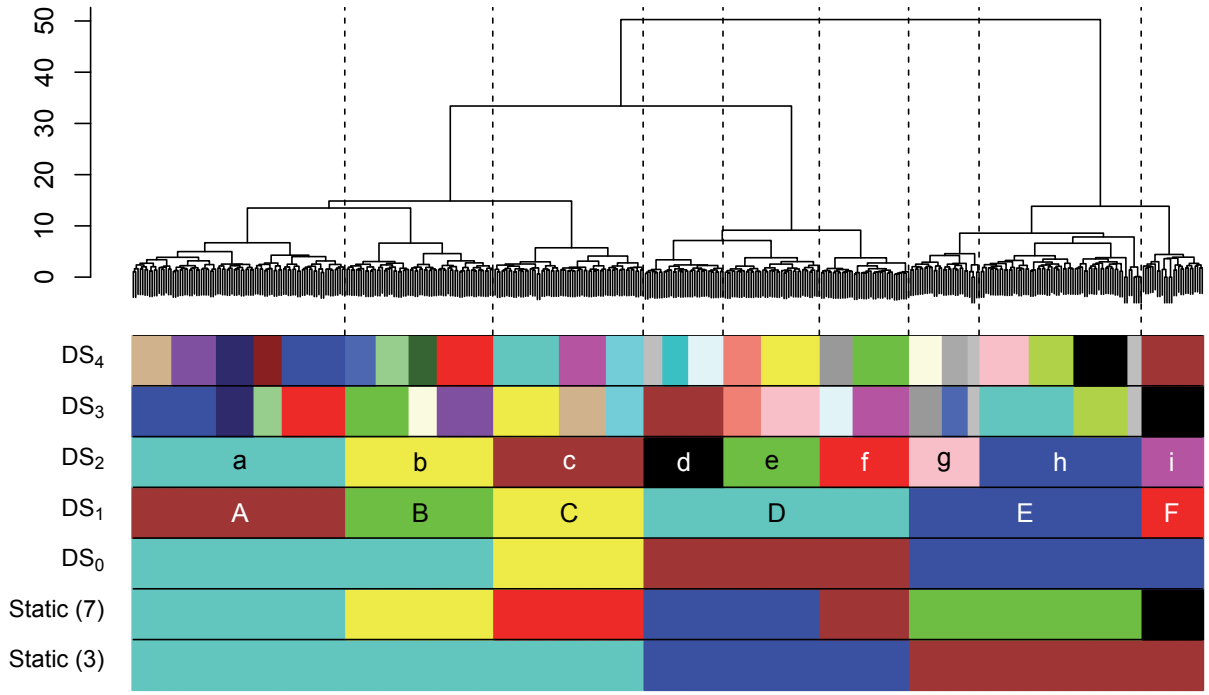


Figure 4.38: The dendrogram for clustering the PPML user semantic measures, along with several potential cut lines.

The clusters seem more stable in the user than the thread clustering, with less variation from $DS_0 - DS_2$. The static method only differs from the DS_1 method in the splitting of a single cluster, D. DS_2 has more clusters, splitting the D and E clusters from DS_1 into subgroups of 3 and 2 clusters respectively. The image matrices in figure 4.39 present the user similarity densities for both sets of clusters.

As with the thread clustering, and in concordance with the core-periphery results, the clustering seems to again have detected a core cluster, and the other clusters have a stronger

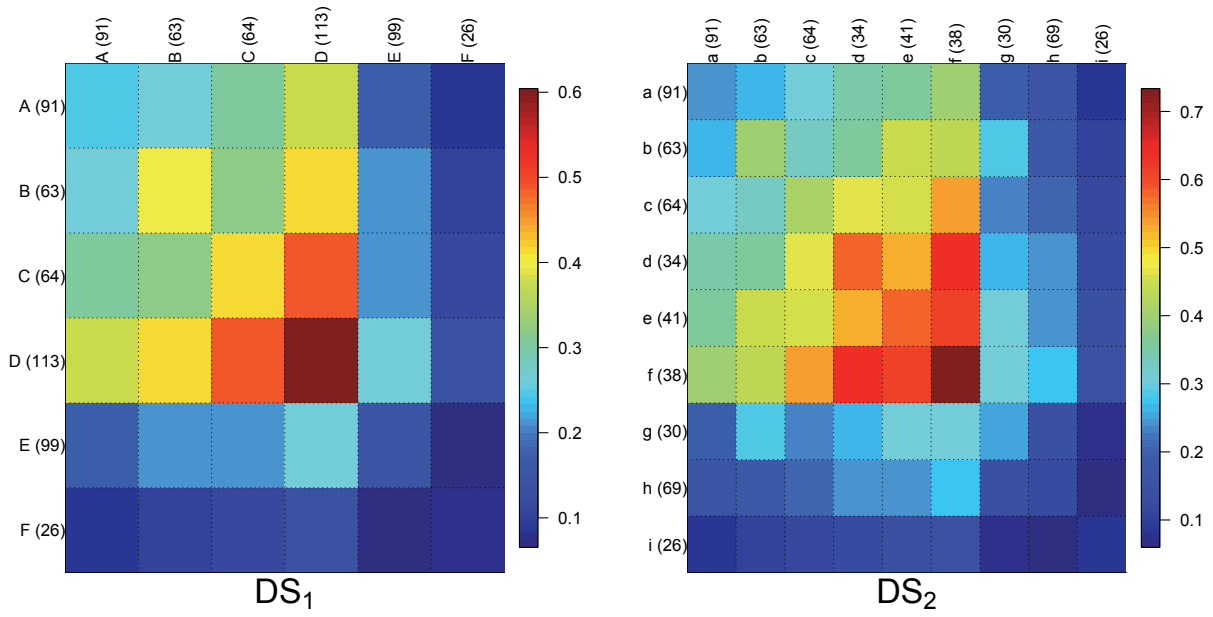


Figure 4.39: The similarity densities from the two hybrid clusterings of the PPML data

relationship to the core than their own cluster. This has again caused the silhouette coefficients to largely be negative, with the exception of the core cluster (see figure 4.40).

When we look beyond the core-cluster, however, we can see that the clusters do have some inter-cluster density, suggesting some pattern. Comparing the cluster D from DS_1 to d, e, f from cluster DS_2 we can see that they are all strongly related, but perhaps more inter-related within the smaller clusters than between the small clusters. This split took a tight-nit group and found 3 potential sub-groups within it. The second split, E into g, h seems to have parsed out a smaller sub-group from the larger population.

The overlap between the connection and semantic clusters for the PPML data is presented in table 4.23 (the semantic user clustering itself is presented in figure 4.38 and tables 4.24 and 4.25).

There is clearly no significant overlap between the two clustering methods. Comparing the DS_1 semantic cluster, we can see that the largest SNA cluster (a, the one that was thought of as "the rest" of the users) is spread relatively evenly between the 6 semantic clusters. Likewise the largest SNA cluster (f, the one with the most central users) is part of the largest and least specific semantic cluster.

There is a theoretical difference in how the two clusters are formed. The SNA clusters are traffic clusters, so they identify users that communicate with each other (or around each

	DS_1 Semantic Cluster					
	A	B	C	D	E	F
a-sna	77	41	40	19	88	21
b-sna	10	20	18	26	9	2
c-sna	2	2	2	3	2	3
d-sna	1	0	1	24	0	0
e-sna	1	0	3	19	0	0
f-sna	0	0	0	22	0	0

	DS_2 Semantic Cluster								
	a	b	c	d	e	f	g	h	i
a-sna	77	41	40	5	13	1	25	63	21
b-sna	10	20	18	12	14	0	5	4	2
c-sna	2	2	2	0	3	0	0	2	3
d-sna	1	0	1	11	2	11	0	0	0
e-sna	1	0	3	5	8	6	0	0	0
f-sna	0	0	0	1	1	20	0	0	0

Table 4.23: Comparing the DS_1 connection clusters (rows) to the DS_1 and DS_2 semantic clusters for the PPML data

other, as no communication is directed). The semantic clusters, in contrast, are content clusters, so they identify things that people have said in common.

In the traffic clusters (SNA clusters) you are rewarded for many messages, regardless of what they are, because that increases your presence in the community. For the content clusters (the semantic clusters) multiple messages around a variety of topics makes you less classifiable because you do not fit into a specific group. The users that are unclustered in the traffic clusters because they have few messages may be tightly clustered in the content clusters because the content of their messages is homogeneous.

Tables 4.24 and 4.25 present summaries of the most prevalent terms in each of the content clusters, based on the highest average TF-IDF scaled score for each user in the cluster.

The clusters present some interesting findings. From DS_1 cluster D is the strongest, and seems to be grouped around Opiate pain medications, with several terms related to strong pain management medications, measurments and side-effects. Cluster C is similar, with a similar but different set of pain medications. Cluster B is the other relatively strong cluster presence, but seems more interest in pain management, with terms around chronic pain an psychology, pain clinics, and questions about research and behaviour.

When we move onto DS_2 cluster D is split into subclusters d, e, f which all seem similar though not identical. The divide of this single cluster into three clusters may not be appropriate, as the sub-clusters seem to be grouped around the same concepts.

(a)

A	B	C	D	E	F
1.61	2.74	3.17	6.17	0.92	1.36
1.62	2.76	3.28	6.26	0.93	1.38
1.65	2.78	3.29	6.27	1.01	1.38
1.72	2.80	3.66	6.30	1.02	1.43
1.72	2.88	3.79	6.39	1.06	1.48
1.74	3.11	3.83	6.47	1.11	1.51
1.75	3.35	3.90	6.50	1.13	1.60
1.79	3.52	3.94	6.50	1.17	1.60
1.98	3.66	4.25	6.63	1.27	1.61
2.17	3.79	4.82	7.75	1.32	1.97

(b)

A	B	C
Pediatrics	Pain Clinics	Pharmaceutical Preparations
Pharmaceutical Preparations	Research	Emotions
Methadone	Behavior	Fentanyl
Clinical Trials as Topic	Work	Lidocaine
Methods	Psychology	Analgesics, Opioid
Pain Management	Pediatrics	Patients
Thinking	Child	Transdermal Patch
Patients	Education	Thinking
Helping Behavior	Ambulatory Care Facilities	Morphine
Child	Chronic Pain	Acetaminophen

D	E	F
Morphine	Wounds and Injuries	Vascular System Injuries
Weights and Measures	Infection	Myelin-Associated Glycoprotein
Sleep	Methods	Silicon Dioxide
Work	Videotape Recording	Weaning
Ketamine	Computers	Hospitals, Group Practice
Organization and Administration	Pain	Hospital Planning
Methods	Helping Behavior	Interleukin-11
Pharmaceutical Preparations	CRPS	Thoracic Injuries
Aptitude	Internet	Rupture, Spontaneous
Analgesics, Opioid	Pediatrics	Cells

Table 4.24: The highest average TF-IDF score within each cluster (a) and the highest scoring terms (b) for the DS_1 user clusters on the PPML

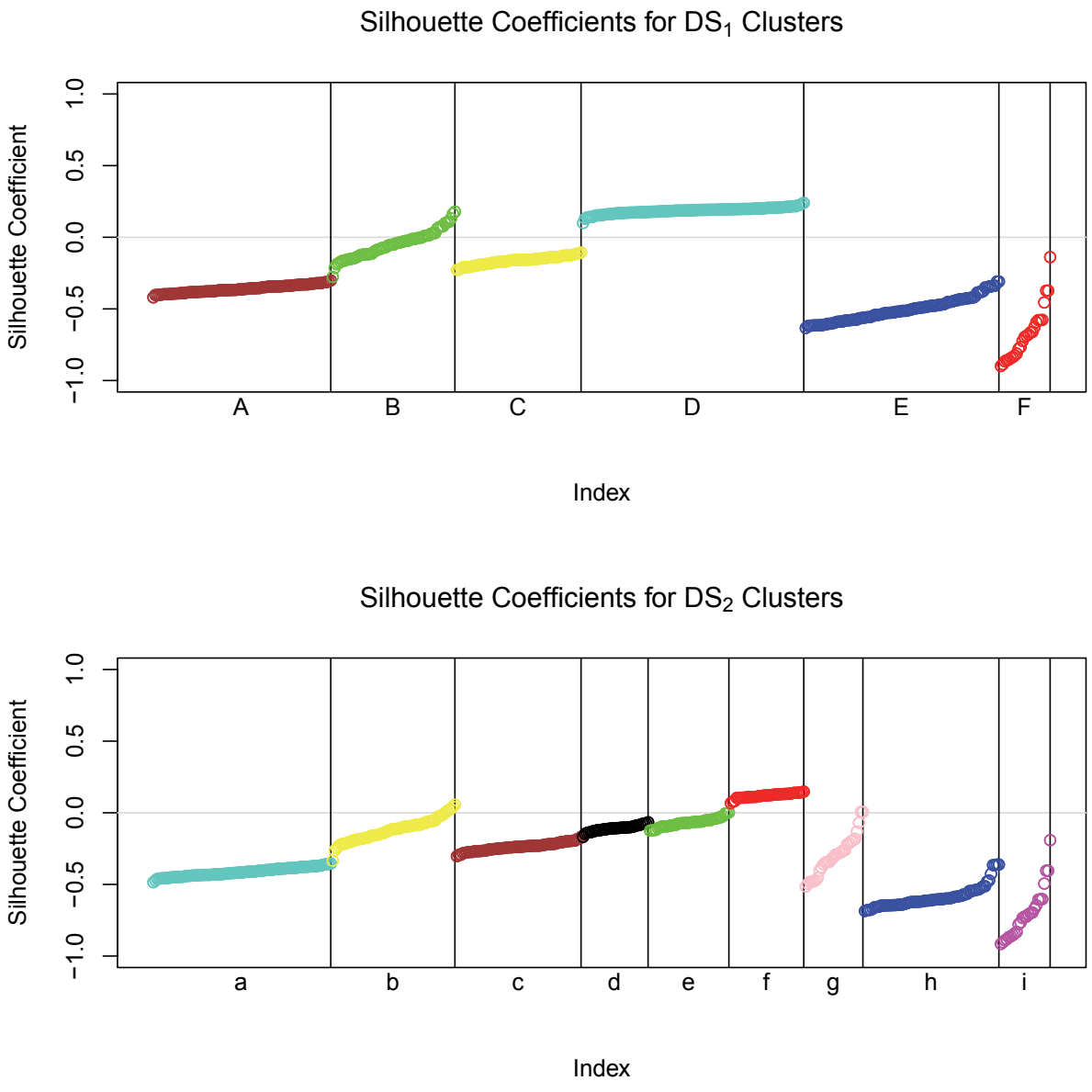


Figure 4.40: The silhouette coefficients for the PPML user clusters

(a)

d	e	f	g	h	i
6.18	4.94	9.04	1.55	0.97	1.36
6.59	4.95	9.06	1.75	0.98	1.38
6.62	4.96	9.10	1.80	1.05	1.38
6.67	5.12	9.32	1.84	1.06	1.43
6.71	5.60	9.38	1.86	1.10	1.48
7.12	5.62	9.52	2.11	1.11	1.51
7.26	6.14	9.76	2.29	1.14	1.60
7.32	6.18	9.77	2.31	1.33	1.60
8.51	6.81	9.96	2.38	1.34	1.61
9.13	7.34	10.07	3.03	1.47	1.97

(b)

d	e	f
Work	Pain Clinics	Ambulatory Care Facilities
Emotions	Adult	Clinical Trials as Topic
Thinking	Analgesics, Opioid	Lead
Midazolam	Organization and Administration	Weights and Measures
Nausea	Pain Management	Hydromorphone
Pharmaceutical Preparations	Aptitude	Publishing
Methadone	Hearing	Fentanyl
Morphine	Sleep	NSAIDs
Analgesics, Opioid	Weights and Measures	Codeine
Ketamine	Pain Measurement	Analgesics, Opioid

g	h	i
Restaurants	NLM	Vascular System Injuries
Computers	Pain	Myelin-Associated Glycoprotein
Rehabilitation	Thinking	Silicon Dioxide
Volition	Helping Behavior	Weaning
Epidermolysis Bullosa	Needles	Hospitals, Group Practice
Fruit	Wounds and Injuries	Hospital Planning
Pediatrics	Catheters	Interleukin-11
Internet	Infection	Thoracic Injuries
CRPS	Methods	Rupture, Spontaneous
Postal Service	Videotape Recording	Cells

Table 4.25: The TF-IDF scores for the highest scoring terms in each cluster of the PPML data (a) and the terms themselves (b) for the DS_2 user cut. Clusters a, b and c were omitted, as they are the same as A, B and C respectively from the DS_1 cut, table 4.24

4.5.2.2 Content Clustering on the SURGINET Users

The user clustering is presented in figure 4.41, and presents a network that seems to be partitioned into 4 or 5 clusters (DS_1 or DS_2).

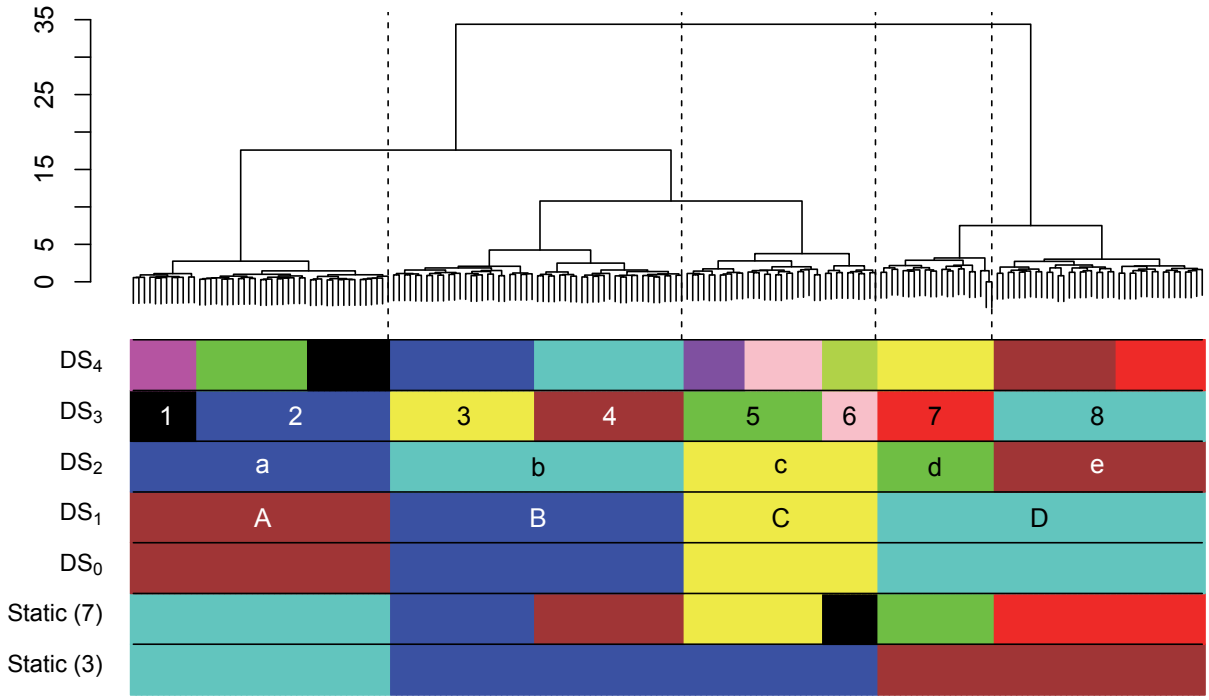


Figure 4.41: The SURGINET user clustering along with several potential cuts in the dendrogram

The image matrices for the two potential cuts are presented in figure 4.42, and the present the same large-core structure that was discovered in the SNA, which was to be expected. The silhouette coefficients (figure 4.43) demonstrate this same structure, with cluster A/a representing the core of the community.

As with the PPML data, comparing the content based clusters with the connection clusters from section 4.4.3 reveals no significant overlap between the two methods. Table 4.26 presents the overlap. We can see that the core group from the connection clusters (D) is all part of a single content cluster, along with 21 other users from the other two small clusters. The large, sparse connection cluster (C) is spread relatively evenly across the other clusters.

The content of the SURGINET content clusters DS_2 are presented in table 4.27. The

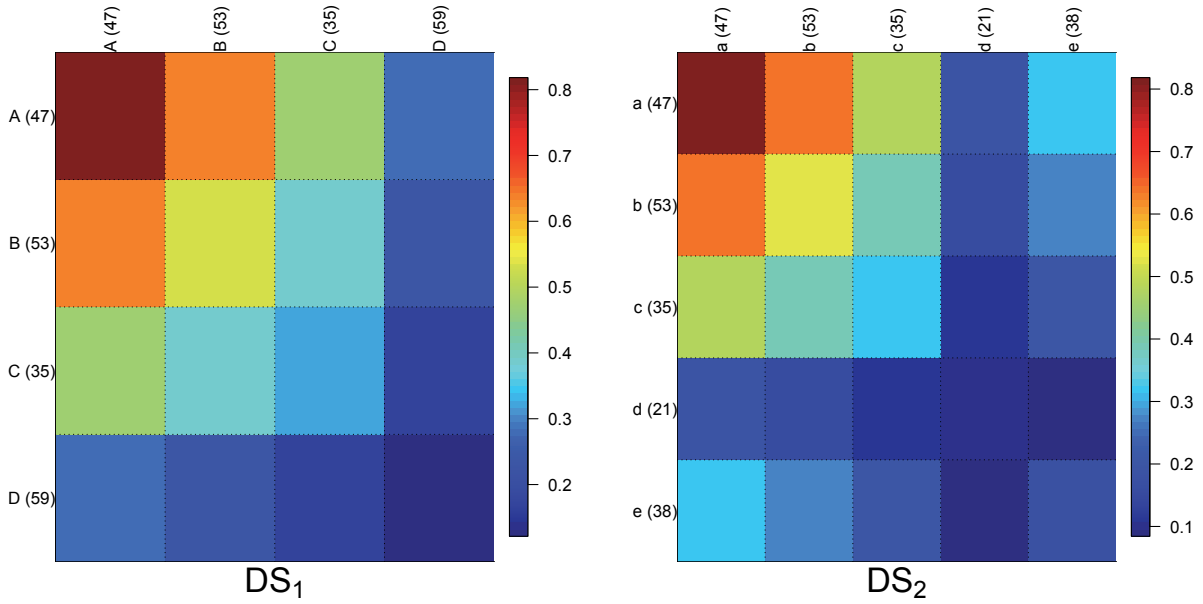


Figure 4.42: The densities of the user-clusters in SURGINET

	DS_1 *Semantic Cluster*			
	A	B	C	D
A	6	10	2	0
B	15	5	1	0
C	0	38	32	59
D	26	0	0	0

	DS_2 *Semantic Cluster*				
	a	b	c	d	e
A	6	10	2	0	0
B	15	5	1	0	0
C	0	38	32	21	38
D	26	0	0	0	0

Table 4.26: Comparing the connection clusters (rows) to the content clusters (columns) for the SURGINET users

strong core-periphery structure of these clusters precludes strong conclusions about the content, but there are still some interesting findings. Clusters *a* and *b* are the densest (which mirrors the results from figure 4.42), and both seem to be centred around general surgery contents. Cluster *c* seems to focus on surgery as well, though more around side effects (Hemorrhage, fractures, necrosis) and *d* and *e* seem to have no noticeable pattern, though that is largely due to the lack of density in those clusters, i.e., there are not many meaningful threads in the smaller clusters.

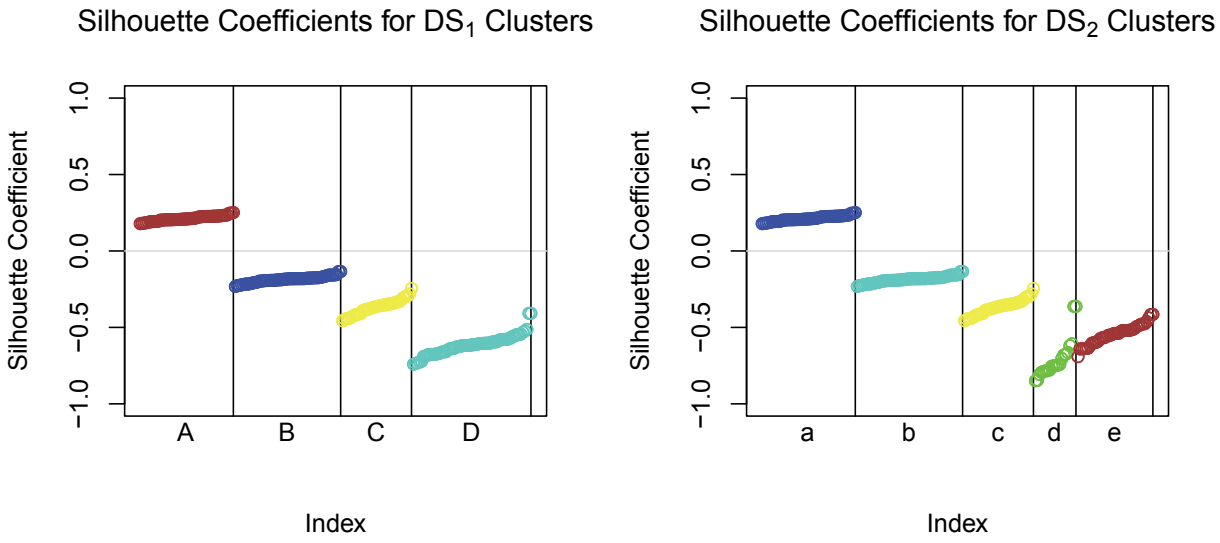


Figure 4.43: The silhouette coefficients for the user-clusters in SURGINET

(a)

a	b	c	d	e
8.35	3.04	1.51	1.45	0.88
8.38	3.07	1.52	1.45	0.88
8.38	3.08	1.55	1.46	0.89
8.45	3.14	1.62	1.48	0.89
8.48	3.16	1.66	1.51	0.91
8.57	3.23	1.72	1.56	0.93
8.57	3.29	1.84	1.60	0.95
8.62	3.30	1.91	1.61	0.96
8.70	3.80	1.93	1.64	0.96
8.72	3.99	1.98	1.70	1.32

(b)

a	b	c	d	e
Sound	CT as Topic	Hemorrhage	Angioedema	Schizophrenia
Appendix	Reading	Lifting	Satellite Viruses	Erythromycin
Fistula	Laparoscopy	Uridine	Ambulances	Claudins
Dissection	Happiness	Fractures, Bone	Traumatology	Epithelial Cells
Drainage	Cholecystectomy	Patients	Bronchial Spasm	Carotid Artery Thrombosis
Abscess	Back	Colectomy	Pancreatic Elastase	Angiogenesis Inhibitors
Running	Work	Necrosis	Balloon Occlusion	Single-Payer System
Appendicitis	Pain	Rectum	Spinal Fractures	Ribosomes
Ileus	Abdomen	Cholecystectomy	Hemodynamics	Risperidone
Liver	Thinking	Intestines	Quinine	Helping Behavior

Table 4.27: The highest average TF-IDF score within each cluster (a) and the terms (b) for the content-based SURGINET users clusters

4.5.3 Summary

The knowledge-based clustering has produced some mixed results. The communication clustering suggested that the users in the community formed more of a core-periphery structure than a clustering structure, and this was confirmed with the content analysis. One of the problems with trying to cluster users based on their content is that the most active users are heterogeneous in their knowledge, so trying to fit them into only one cluster is a difficult task, and may result in tightly clustering the homogeneous users, which are the users that are less active within the community.

The thread clusterings again revealed a core-periphery structure, with tight clustering around a small group of threads. Within the PPML data there was some evidence, even among the periphery, of contextual knowledge clusters, which is a promising finding. The SURGINET data revealed a larger and partitioned core, with some threads around specific surgical ideas and other being more general, which does summarize the community well. Beyond the core, however, there is little evidence of strong activity.

4.6 Detecting Content Expertise: A Network Analysis of the BICGM

One of the objectives of the research was to identify content experts. The communication pattern analysis did not reveal any satisfactory definitions of content expertise, but the BICGM network, in which a tie between user A and B measures how interested A is in B's content, may provide the means for identifying expert users within the community. The BICGM network was used to construct a directed network by creating a directed tie between all users with a BICGM value > 0.5, the *expertise* network described in section 3.5.4.

The directed centrality metrics for the PPML are presented in figure 4.44 and the highest scoring users are presented in table 4.28. The table and figures present a network with, again, a small number of elite users that make up a central core. The users with the highest authority measures are similar to the users with the highest coreness measures, but there are some variations. Figure 4.45 compares the coreness and authority measures, colour-coded by the difference between the two scores. With a correlation value of 0.86 the coreness and authority are highly related to one another, but looking at the figure there are a couple of notable differences. At the high end of both centrality measures are the same users, those power users that are at the centre of both the communication network (coreness) and the

expertise network (authority). Moving beyond the highest level there are some users that have a moderately high value in one metric and a low value in the other. The red dots represent users that have high authority but low coreness, which suggests that they are experts in a certain field, but their overall contribution rates are low, and thus they have low coreness centrality. Conversely, the blue dots represent users that have low authority, so they have not demonstrated expertise in any particular field, but their contributions are higher so they have higher coreness measures.

	Out Degree	In Degree	Proximity	Authority
S2509	8	485	0.498	1.000
S2122	17	484	0.498	0.999
S2100	19	476	0.494	0.990
S2523	13	475	0.493	0.987
S2105	18	473	0.492	0.985
S2176	23	470	0.491	0.982
S2340	21	471	0.491	0.980
S2078	18	467	0.489	0.978
S2225	20	461	0.487	0.974
S2111	24	462	0.487	0.970
S2132	24	462	0.487	0.969
S2101	23	452	0.482	0.959
S2271	27	452	0.482	0.949
S2155	30	442	0.478	0.944
S2130	24	441	0.477	0.943
S2119	40	440	0.477	0.941
S2236	30	438	0.476	0.940
S2085	38	427	0.471	0.928
S2153	29	421	0.468	0.910
S2410	31	399	0.458	0.879
S2578	22	404	0.460	0.877
S2095	46	386	0.452	0.862
S2161	21	388	0.454	0.844
S2091	33	370	0.446	0.830
S2121	29	365	0.444	0.820

Table 4.28: The most central PPML users in the directed network, sorted by authority centrality

User S2578 is the darkest red point in the figure, with a coreness of 0.241 and an authority of 0.877. He has shared 21 messages on the mailing list on 15 threads, but the threads he has communicated on have not been overly popular, so he is only connected to 41 other users in

the shared threads network (the 1-mode network), resulting in the low coreness measure. His messages themselves, however, have been knowledge rich, with an average of 13.7 mapped terms per message (greater than 81% of the other users), so his knowledge contribution has been much higher than his network connectivity. From a knowledge perspective he is much more important to the community than his collaboration patterns would suggest.

At the opposite end of the spectrum is the darkest blue point in the figure, user S2237 with a coreness of 0.366 and an authority of 0.053. She has shared only 8 messages (on 8 threads), but they have been much more active threads, so her shared threads network has 55 total connections. Her messages contributed to the community have been short (only 4.2 mappings per message), and looking through the content of the messages she seems to have largely responded to knowledge seeking activity with short answers that directly answer a question, or facilitation-type responses that direct the knowledge seeker to the correct knowledge sources.

Both users contribute to the community. Users with high authority are contributing more knowledge to the community, but those users like S2237 that are participating in the community without contributing as much knowledge are still important. She is facilitating and participating in active conversations with important but short contributions.

For the SURGINET data the distribution of metrics are presented in figure 4.46 and the highest scoring users are presented in table 4.29, and present a similar pattern to the PPML data. The correlation between authority and coreness is 0.82, so similar. Due to the distributions of the two centrality measures the raw comparison does not make sense (left side of figure 4.47), but the right side comparing the relative rankings of the measures presents the same pattern, with a users in the middle level of the rankings on with either high coreness and low authority or vice versa.

	outDeg	inDeg	proximity	auth
S0951	23	193	1.000	1.000
S1020	33	192	0.995	0.998
S0952	32	192	0.995	0.997
S0968	32	191	0.990	0.996
S0959	26	190	0.985	0.995
S0998	33	192	0.995	0.995
S0956	33	190	0.985	0.993
S1009	32	189	0.980	0.990
S0992	31	189	0.980	0.990
S0989	33	189	0.980	0.989
S0970	30	191	0.990	0.989
S0966	34	188	0.975	0.986
S0972	32	188	0.975	0.986
S0993	33	187	0.970	0.980
S0980	33	187	0.970	0.980
S0957	33	185	0.960	0.978
S1008	33	185	0.960	0.977
S0953	33	185	0.960	0.974
S1037	36	184	0.955	0.972
S0969	36	184	0.955	0.970
S1022	34	184	0.955	0.970
S0971	35	183	0.951	0.965
S0977	33	183	0.951	0.964
S0975	32	184	0.955	0.962
S0950	35	181	0.941	0.959

Table 4.29: The most central users in the SURGINET BICGM network, sorted by authority centrality

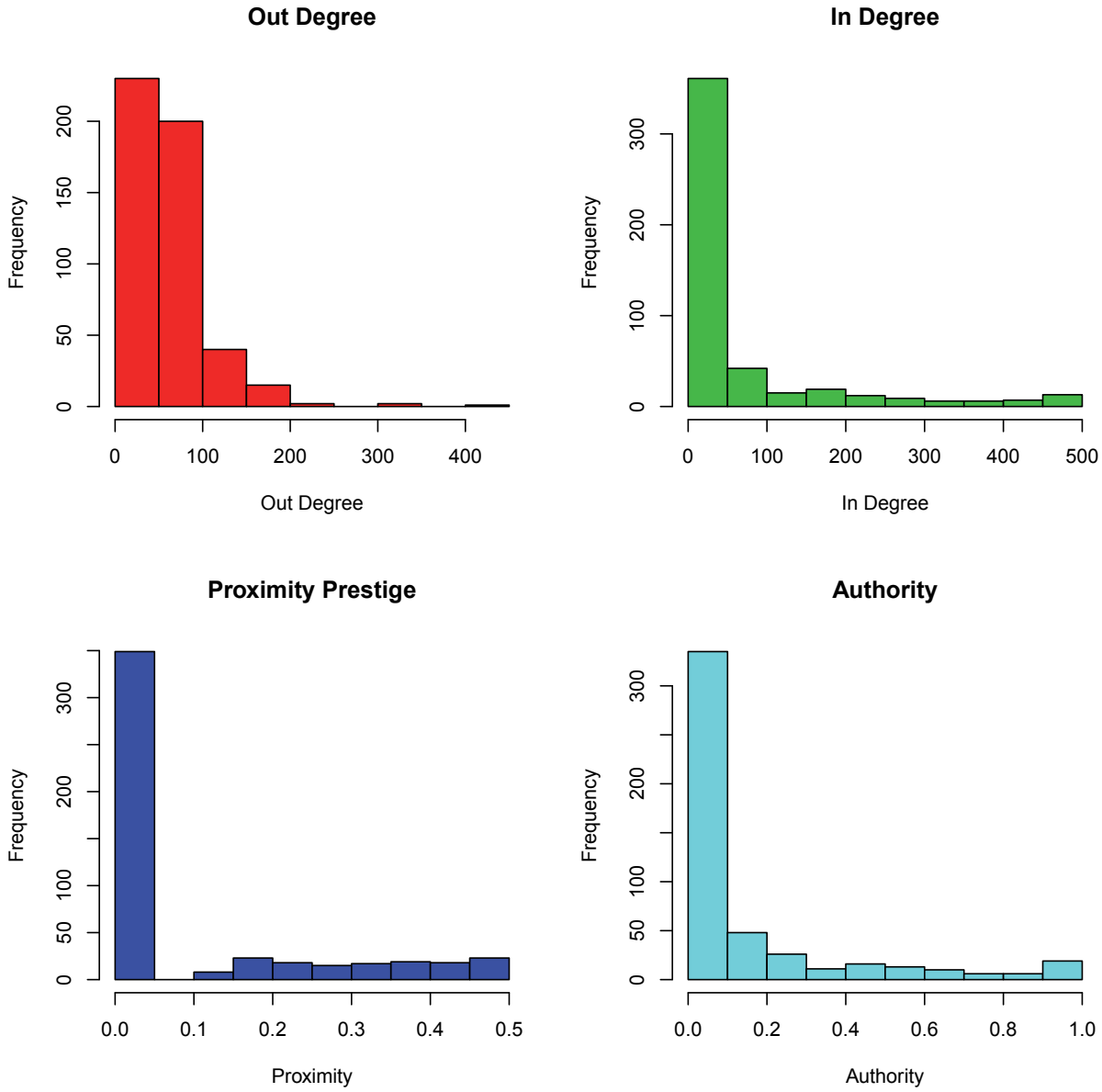


Figure 4.44: The distribution of the four directed centrality measures for the PPML BICGM network

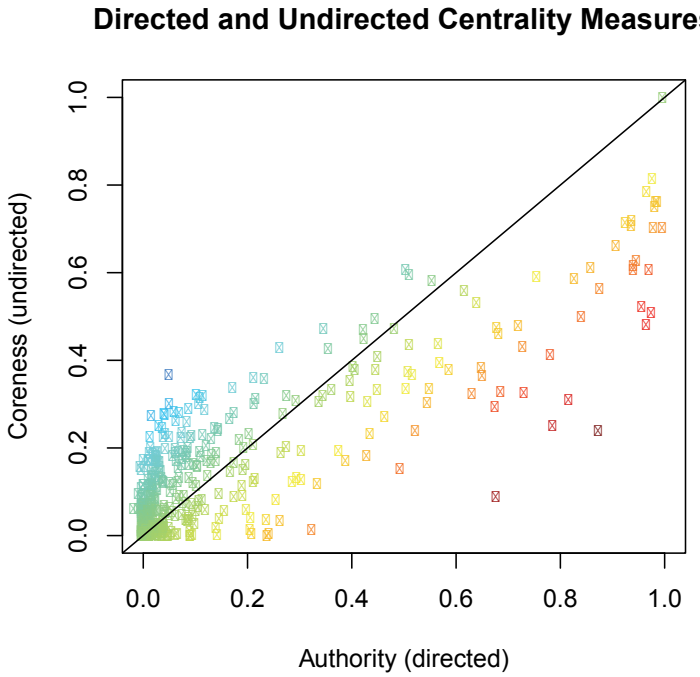


Figure 4.45: Comparing the PPML collaboration centrality (coreness) to the knowledge-based centrality from the BICGM network (authority). The colour of the points denotes the difference between the two values, with blue indicating higher coreness to red indicating higher authority

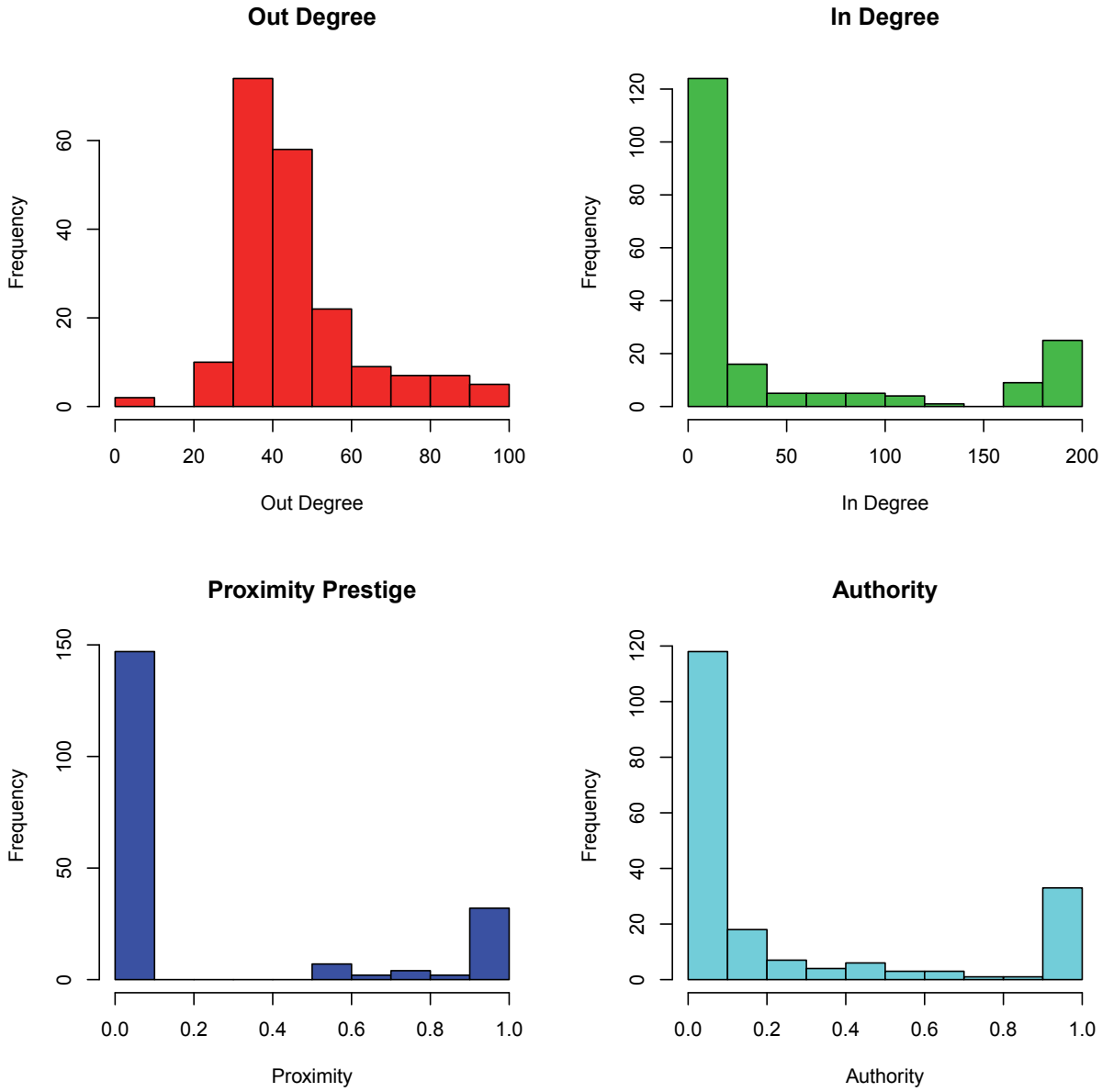


Figure 4.46: The distribution of the four directed centrality measures for the BICGM network for SURGINET

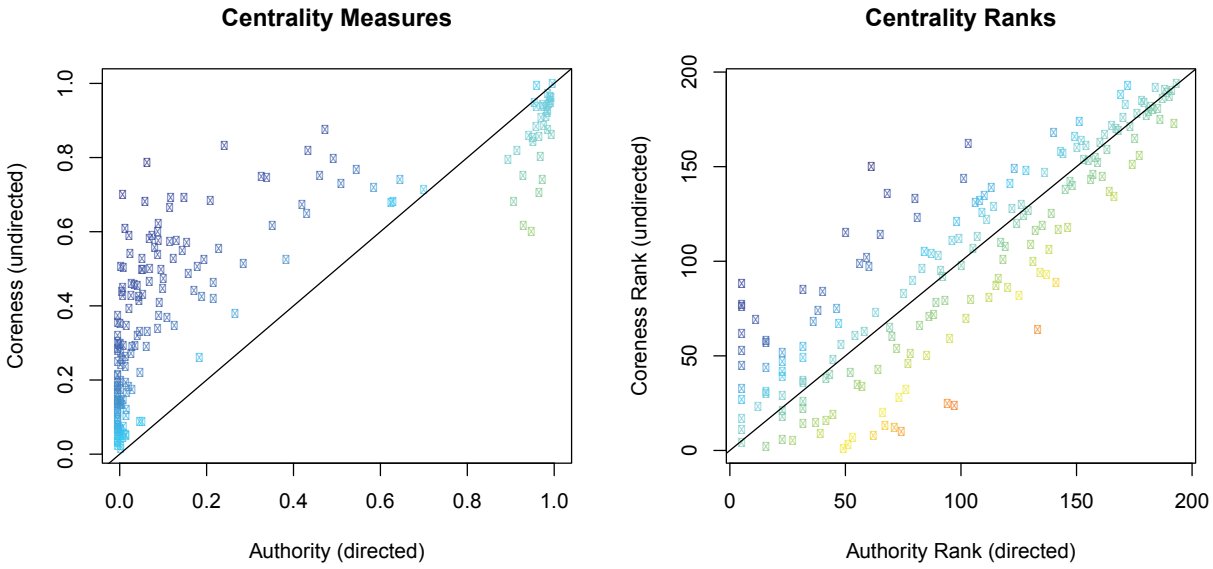


Figure 4.47: Comparing the coreness and authority measures on SURGINET. The left figure presents the raw values, but due to the distribution of the two metrics the raw differences are not as informative as the right figure, which compares the ranks of the two metrics.

4.7 Conclusion

This chapter has evaluated the methods presented in the previous chapter on two test datasets. For the PPML the isolate analysis has detected pendants and determined an algorithm for preventing future pendants, but an investigation into the cause of pendant threads found no evidence of systemic bias based on either user activity levels or pendant content. The response analysis determined that the communities have similar attention spans, which is an interesting finding for two communities with vastly different activity levels.

The clustering of the knowledge-based user similarities was not overly successful, but the clustering of the threads identified some potential clusters within both communities. The core-periphery structure identified amongst the users was also present amongst the threads, but for the PPML the non-core threads demonstrated some clustering, and the core of the SURGINET data could be relatively well partitioned into two disparate groups.

The BICGM network provided the means to build a knowledge-based expertise network, a directed network in which an authority measure could be calculated to study which users are considered the experts within the community. Comparing this measure to the coreness measure revealed the similarities and differences between leadership detected based on communication patterns and communication content.

Chapter 5

Discussion and Conclusion

KT is a vital component of the modern healthcare community, and online communication tools can provide a valuable addition to the KT process by providing a venue for community members to meet and discuss clinically relevant issues in a way that bypasses the traditional obstructions to face-to-face communication. In order for these communities to function, however, they must be guided by formal implementation guidelines, and the LINKS model [1] provides such a framework. The objective of this thesis was to develop methods for understanding and improving the culture of collaboration and the knowledge context of a LINKS-guided online community of practitioners. Through a combination of communication and content analysis we have developed methods that address these two issues, providing a more detailed understanding of the community as a whole and developing the means for guiding the KT process.

This chapter will summarize the thesis in two sections. The first will discuss the methods developed in this thesis, explore how they addressed the original objectives of the research and how the combination of SNA and content analysis can provide unique insights into KT processes. The second section will investigate the applications of this research, how the methods developed here fit into KT frameworks, where future research may go, and the limitations of the methods developed here and how to address them.

5.1 Summary of Methods

The objectives for this thesis were given in table 1.1, and the analytic approach for answering them was given in figure 1.1. This figure demonstrates how communication and content analysis can be used to answer questions about the community, and also how they can combine to provide additional insights that SNA and content analysis cannot provide on their own. Table 5.1 presents the objectives from table 1.1 along with how they were answered in this thesis, and what the overall contribution of these methods is to the KT process. The following sections will investigate these four objectives in a detailed manner, investigating

the results from the application of the methods to the two test datasets (the PPML and SURGINET).

Objective	**Method(s)**	**Contribution**	**Impact on KT**
Identifying Community Leaders and Content Experts	Coreness and authority	Methods for identifying community leaders based on communication patterns or communication content; identification of content experts and facilitators through authority analysis	Identifying opinion leaders, change agents, content experts and facilitators
Calculating User/Thread Similarity	GVSM and BICGM correlation	Two novel methods for calculating content-based similarities between users	Providing advanced archive and community navigation tools for improved connectivity
Detecting Clusters	SNA clustering; GVSM clustering	Content investigations of SNA clusters and contrasts of SNA vs GVSM clusters provides insight into potential structure of subgroups	Preventing segmentation and monitoring/managing potential sub-groups
Content summaries	Knowledge Maps	Providing high level insight into the content being discussed within the community	Ensuring that the knowledge context of the project is being met
Isolate Detection	Detection algorithm; content analysis	Algorithm of preventing pendants; methods for studying causes of pendants	Increasing connectivity and ensuring a culture of collaboration

Table 5.1: Contributions of the methods developed in this thesis

Many of the methods in this thesis are dependent on the mapping of the unstructured text in the messages to MeSH terms using Metamap to provide formal representation of the knowledge contained within the threads. These mappings are the basis for the GVSM and BICGM correlations and they provide the detailed insight into the content of the pendants and the SNA clusters along with the knowledge maps.

5.1.1 Identifying Community Leaders

Community leaders are the guiding forces of the knowledge sharing practices within the community, so identifying and engaging them is essential to furthering the KT process. Coreness (section 3.2.3.2) is a measure of how central a user is to the community based on the communication patterns, while authority (section 3.5.4) is a measure of how central a user is to the community based on the content of their messages. When compared to one another the metrics provide the means to separate facilitators, those with high coreness and low authority, from inactive content experts, low coreness and high authority, with core members having high values of both metrics.

When applied to the PPML and SURGINET the coreness and authority measures revealed a core group of users that dominate the bulk of the communication within both communities. Both networks demonstrated a core-periphery structure, though the PPML's structure was tighter, with fewer core members and a slightly more active periphery. Figures 4.45 and 4.47 compare the coreness and authority metrics and the three groups of users named (facilitators, inactive experts and core members) can be seen in both figures. This type of insight provides added information not available in previous studies [16, 21, 26] that looked only at the centrality measures as indicators of leadership within the community.

5.1.2 Calculating Content-based Similarities

The content mappings provide an improved method for calculating user and thread similarities than network based methods because they are based on what the users have said and not only where they said it. The BICGM and GVSM are two different approaches to calculating user or thread similarity that make full use of the taxonomic structure of the MeSH lexicon. The BICGM took the BGM [29], improved upon the existing algorithm, and adapted it to incorporate context-specific relationships using information content. The GVSM similarity is dependent on a term correlation matrix, and this thesis developed a term correlation calculation that combines the semantic and context-specific relationships between MeSH terms to fully capture the relationships between them. When plugged into the GVSM calculation the result is a set of user similarities that fully reflect the inherent relationships between the representative MeSH terms. The effect of the BICGM adaptation was evident in the specific examples provided, but at the macro level it did not differentiate from the adapted BGM in the way it ranked users. The adaptation to the term correlation

was found to have a significant effect at both the individual and the macro level, suggesting that it is having a positive effect on the calculation of user and thread similarity.

Section 3.4 compared the new methods to more traditional approaches. In both the GVSM and BICGM examples specific cases were found where the new metrics provided an improved measure from the traditional values. For the BICGM the overall effect seemed to be minor compared to the BGM, but for the GVSM the addition of contextual similarity was found to significantly influence the list of similar users. Future work should pursue these differences via survey-based evaluation of the community (see section 5.2.2 for a more detailed exploration of these tests).

5.1.3 Finding User and Thread Clusters

The clustering of the users within the PPML and SURGINET was disappointing, as no real clusters were detected. More advanced clustering algorithms may provide a result (though several methods were investigated in this project beyond what was presented), but ultimately clustering methods only work when there are true clusters within the data, and the nature of mailing lists may preclude user clustering. Since every user receives every message, there may be a tendency for the leaders in the community to dominate the conversations, creating a homogeneity of knowledge. This is not necessarily a problem, as the objectives of establishing a community of practice are to bring interested users together around a common subject, so in that scenario the existence of clusters would represent the amalgamation of multiple communities that may be better off split into their own KT groups. Perhaps larger or more diverse communities would provide a better test bed for clustering methods.

At the thread level both networks resulted in a surprisingly strong core-periphery structure, but within those clusters there were some valuable findings. The PPML network had a core centred around generic pediatric pain management content, but in the periphery there were loose clusters around specific concepts, suggesting the potential for some slight subgroups within the community. For SURGINET the core-periphery structure was much stronger with no real pattern in the periphery, but the core was divided into two clusters, one focused on procedure-specific content and the other on more general information that was less surgery-specific. This is representative of the SURGINET community as a whole, which is less context-focused than the PPML, and the clustering reflected that finding.

5.1.4 Presenting Content Summaries

The knowledge maps identified what content each community spent their time discussing. SURGINET conversations centred more around anatomy and surgical procedure issues, while the prevlance of pain and medications was popular within the PPML. Both mailing lists had a significant presence of terms related to the diagnosis of diseases, suggesting that there might be some overlap between the two communities. Summarizing these mappings using knowledge maps provides a broad overview of the content of the conversations within the community, and can be used in future research to identify desired knowledge areas in which the community seems to be lacking.

5.1.5 Detecting and Analyzing Pendants

Combining pendant and response analysis resulted in a pendant detection algorithm that could be used to prevent future pendants from occurring. The pendant detection algorithms were applied to the PPML (no pendants were present in SURGINET), and though they identified both the pendants and a way to prevent future pendants, no systemic bias was detected in the community based on either the pendant author's activity history or the content of the pendants themselves. The two worries with pendants are that they are not being responded to because the sender is not a respected member of the community, or because the content of the messages are not in concordance with the knowledge base of the community, but neither of those problems arose on the PPML. The content analysis was used to supplement to pendant analysis in this section to go beyond looking at what messages were pendants to try and determine why they were pendants. The lack of findings in this section suggests that there are no systemic causes of pendants within the community, and the methods developed here could be applied to other medical communities.

5.2 Applications to KT Programs

The use of KT frameworks is becoming increasingly common within the medical community, and supplementing these frameworks with online tools is an important next step toward ensuring a healthy and active KT process, but without methods to understand the online KT process we cannot fully leverage it or incorporate it into the KT process as a whole. Answering the objectives stated in table 1.1 are important to improving our understanding

of KT overall, and this thesis has outlined specific analytic methods for addressing those objectives in table 5.1 that have not previously been explored within the literature.

From a methodological perspective the BICGM and the term correlation methods developed in this thesis are an important contribution to the literature both within and beyond the KT community. The BGM [29] was a useful theoretical model, but the methods in the original specification needed clarification in certain spaces and improvement in others. This thesis has adapted the BGM to non-leaf mappings and issues of homonymity, and clarified specific issues in the original specification that were not clear. Most importantly, it has improved the model by moving from edge-based to information-content based measures of term similarity. The analysis of the adaptation suggest that, at a macro level, the information-content adaptations may not have been as significant an outcome as was initially hoped, but even using edge-based methods the BGM needed improvement that this thesis has provided. Even without empirical evidence for an effect the movement from edge-based to information-content-based methods is a theoretical improvement.

Little work has been done on using the semantic mappings from programs like Metamap [5] to calculate similarities, but those that have usually ignored the semantic relationships between MeSH terms. The GVSM is designed to leverage those similarities, but needs a term correlation matrix to function. The calculation of the term correlation matrix in this thesis, using a combination of semantic similarity and contextual similarity, leverages the inherent relationships between terms while incorporating the context-specific relationships between terms. With the methods combined the GVSM applied to the user and thread representations makes full use of all the knowledge about the users/threads. Both the GVSM and BICGM correlations can be applied to any project that uses semantically similar terms to represent objects and needs to calculate similarities between them.

5.2.1 Augmenting the KT Process

The LINKS model [1] provides a framework for developing online KT environments. In order for the LINKS model to be successful it needs analytic tools that can provide feedback to the community members, ensuring that their KT needs are being met. This thesis has developed methods to monitor two components of the LINKS model, the knowledge sharing context and the culture of collaboration. Table 5.2 presents the components of the LINKS model addressed by this thesis. These same components can be directly tied to parts of the PARIHS

Level	Element	Contribution
Conceptual Level	Knowledge Modality	
	Knowledge Sharing Context	Knowledge Maps; Content Experts; User/Thread Correlations
	Knowledge Sharing Medium	
Operational Level	Technical Infrastructure	
	Culture of Collaboration	Community Leaders; Pendants; Subgroups; User/Thread Clusters
Compliance Level	Trust	

Table 5.2: Summary of the contributions to the LINKS model [1]

framework as well. The PARIHS framework considers KT as a 3-part process (see figure 5.1) involving evidence, context and facilitation [72]. This thesis has developed methods for directly addressing the facilitation and context components, providing feedback on how the community functions and what may need to be done in order to improve knowledge sharing.

Figure 5.1: The PARIHS Model [72], supplemented with the analytic additions from this thesis

The PARIHS framework depends on leaders and facilitators for transmitting knowledge through the community [72], so detecting leaders is paramount to facilitating KT. For experiential knowledge in an online community these leadership roles are filled organically rather than explicitly, so identifying who fills these roles is important to monitoring the community. The coreness and authority metrics are two measures that can provide insight into who forms the core of the community, who the content experts are and who the facilitators

are. An analytic framework could be developed to apply to online communities or mailing lists to identify these users so that the community can be better understood and future knowledge-based interventions can be applied in an optimal manner.

Detecting user and thread similarities provides important additional connections within the community. Identifying similar users can facilitate the interaction component, and identifying similar threads can help identify what specific subjects are being discussed, helping to understand what drives the community. In order to fully leverage the user and thread similarities there needs to be a viable way to navigate the archives of the community. Within the archives similar users could be presented to community members as "potential collaborators", and similar threads could be presented to supplement the knowledge in the current thread.

Clusters are a potential problem within a KT-based community, as research has suggested that the segmentation of communities leads to the "death" of those communities [39], and identifying and separating clusters can ensure that the "common subject" component of the Communities of Practice is maintained [94]. Clusters can be identified using the methods presented in this thesis, and if the content area is found to be sufficiently different then the cluster can be split off into its own group. Within an online community this could be a matter of creating a new area within the community, while segmentation in a mailing list may require the spawning of a new mailing list around a more specific subject.

The content summaries provide important insight into the knowledge context of the community. The knowledge maps provide insight into the content of the community, and linking these summaries to the archives of the community could provide a novel archive navigation tool that could increase the connectivity of the community as a whole. Finally, detecting and preventing pendants is essential to establishing a culture of collaboration within the LINKS model [1]. The algorithm presented in this thesis can provide the means for preventing messages from going unanswered, and monitoring of the content of the potential pendants using the methods presented here can provide insight into potential explanations as to why these messages may not have been resolved.

Overall online communication media can provide an alternative avenue to face-to-face communication for facilitating KT, bypassing communication barriers, connecting disparate groups and providing archives of KT activities, but there are challenges in moving to a new communication venue. Without face-to-face communication establishing trust-based

connections is a challenge, and the presence of a larger community, though a potential benefit overall, can be overwhelming and a challenge to navigate. The methods presented in this thesis provide the means for improving the online KT process and addressing the issues that may arise in implementing and using such a system. Combining the analytic methods here with the LINKS model in implementing online KT using the PARIHS framework can ensure optimal connectivity between community members and an improved sharing environment overall.

5.2.2 Future Work

The creation of usable, navigable archives is paramount to making full use of the knowledge shared within the communities, and several of the approaches for leveraging the methods in this thesis require the use of navigable archives. It is not feasible to consume all conversations within an active community, particularly as the majority of conversations last less than three days, which results in the same questions arising multiple times. For online communities these archives are normally present, and the methods presented above can improve the navigation of those communities, but for mailing lists there is a need for an online archive that can be linked to the conversations to provide a usable archive of previous threads.

The application of all methods presented in this thesis could be applied to online discussion forums, and forums have the additional details of who consumed knowledge (i.e. who read specific threads), providing another level of both network and semantic information. It is difficult to incorporate users without them contributing to a conversation, but "Lurkers", those that read the contents of online communities without contributing, are a large part of the community, and incorporating them into the analysis could provide additional insight. Our previous research [79] investigated the roles of lurkers in a foreign language community, and combining those methods with the content-based methods developed here would be a valuable extension of this research.

As with all algorithmic research there is always the potential for improving the details of the algorithms. There are coefficients and components of the term correlation and BICGM algorithms that could be tweaked and studied. Other clustering algorithms could be investigated for detecting content-based clusters of users or threads, and more advanced SNA methods may provide additional insight into the community.

The methods presented in this thesis look at the communities in a cross-sectional manner

(with some temporal investigations of individual threads), but there is a larger temporal structure within the communities that may be worth investigating. Studying the evolution of a community, both at the user level and at the community level as a whole, may provide methods for measuring the growth of knowledge within the community, for observing users transforming from knowledge seekers to content experts, or for detecting the development of subgroups.

This thesis addressed two major components of the LINKS model, the culture of collaboration and the knowledge context, but other dimensions warrant further investigation. Looking at figure 2.1 and table 2.1 the Knowledge Sharing Medium and Technical Infrastructure are not issues that need to be studied, as they are specified and implemented at the beginning of a project, but evaluating the Knowledge Modality and the Trust in the system are components that may warrant further investigation in future studies. Studying the knowledge modality involves parsing the content of the messages and determining what kind of knowledge is being shared, while evaluating trust requires contacting community members to determine if their practice is changed based on the knowledge they received from the mailing lists. Both projects require a different analytic approach than the methods in this thesis and are therefore beyond the scope of this work, but future investigation into those components could be combined with this research to provide a complete analytic framework for the LINKS model.

There are several results in this thesis that could be validated through engagement with the communities themselves. A survey to evaluate the utility of the GVSM and BICGM similarities could be useful. There are many approaches to take for this kind of project, but the objective is to determine if the BICGM rankings are better than the BGM rankings, and if the contextual supplement to the term correlation improves the similarities. The community members would be presented a list of users and asked "how similar are these users to you?" along with a second question "who are your peers in the community?", and these results would be compared to the rankings. The issue of community leadership would be investigated by a similar survey, in which the users are asked "who are the community leaders?" and "what is a leader?". The second question is particularly interesting, to see if users lean toward a coreness or authority-oriented definition. Engaging the community could also provide feedback on the best way to provide a usable archive of old conversations, and how best to present additional information (similar threads, for example) in the most

effective format possible.

5.2.3 Limitations

There are limitations to this research that should be addressed. The methods developed within this thesis have not been incorporated into a KT framework at the process level. This chapter has outlined how the methods can be used to supplement KT, but future work should investigate the process of incorporating them into a KT framework and deploying them in a real-world environment.

Most of the methods described here are dependent on semantic mapping programs like Metamap [6], and therefore the majority of the methods here cannot be extended beyond the healthcare community. Some of the communication analysis would be applicable to non-medical communities, and extensions to non-medically oriented lexicons like Wordnet [61] may be possible, but the methods here are designed largely for medical communities. Specifically, the methods here are most useful for closed, professional communities. Non-professional communities, such as physician-patient or patient-patient communities, are very common, but the language used within these communities is different so semantic mapping approaches are more difficult, and the objectives of these communities are often very different from the KT communities that are the focus of this thesis. The objective of this thesis is to improve KT in online communities, so if KT is not the objective then these methods are not optimal. The response analysis and coreness metrics would still work, and for certain communities the content mapping may be successful, but the implications of a non-secure community, or a community in which the trust dimension of the LINKS model is not present, are wide ranging and warrant their own research pursuits.

MeSH was used as the target semantic language for this project over potentially more rich medical lexicons like SNOMED or UMLS. The decision was made to use MeSH because it is the same lexicon that is used to index PubMed, which was used in a previous project [78]. The theoretical implications are minimal: Other lexicons still have a hierarchical (or at least somewhat hierarchical) structure of relationships between terms, and most of them can be mapped to using Metamap or other semantic mapping programs. The major effects of using a different and potentially more complex lexicon are the calculations of the semantic correlation from equation 3.23, the information content for the BICGM in equation 3.27, and the aggregation of terms in the knowledge maps. These issues could all be addressed

without altering the theoretical structure of the thesis, but comparisons of the effect of using MeSH, SNOMED or other lexicons on the outcomes may be pursued in future work.

Bibliography

[1] Syed Sibte Raza Abidi. *Healthcare Knowledge Sharing: Purpose, Practices, and Prospects*, chapter 6, pages 65–86. 2006. ix, xii, xiii, 2, 3, 8, 14, 15, 16, 99, 181, 186, 187, 188

[2] Syed Sibte Raza Abidi. *Healthcare Knowledge Management: The Art of the Possible*, pages 1–21. Springer, 2008. 1

[3] L.A. Adamic, J. Zhang, E. Bakshy, and M.S. Ackerman. Knowledge sharing and yahoo answers: Everyone knows something. pages 665–674, 2008. 17

[4] Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, pages 211–230, 2003. 55

[5] Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *Proceedings of the AMIA Symposium*, pages 17–21, 2001. 8, 19, 99, 186

[6] Alan R Aronson and Francois-Michel Lang. An overview of metamap: historical perspective and recent advances. *JAMIA*, 17:229–236, 2010. 37, 191

[7] R Aviv, Z Erlich, G. Ravid, and A Geva. Network analysis of knowledge construction in asynchronous learning networks. *Journal of Asynchronous Learning Network*, 7(3), 2003. 17

[8] R Beck, W Fitzgerald, and B Pauksztat. Individual behaviors and social structure in the development of communication networks of self-organizing online discussion groups. In B Wasson and U. Hoppe, editors, *Designing for change in networked learning environments.*, pages 313–322, 2003. 35

[9] Stephen P. Borgatti and Martin G. Everett. Network analysis of 2-mode data. *Social Networks*, 19:243–269, 1997. 32, 35

[10] Stephen P Borgatti and Martin G Everett. Models of core/periphery structures. *Social Networks*, 21(4):375–395, 1999. 35

[11] Wendy W Chapman, Marcelo Fiszman, John N Dowling, Brian E Chapman, and Thomas C Rindflesch. Identifying respiratory findings in emergency department reports for biosurveillance using metamap. *MEDINFO*, 2004. 20

[12] Herbert S Chase, David R Kaufman, Stephen B Johnson, and Eneida A Meddonca. Voice capture of medical residents' clinical information needs during an inpatient rotation. *Journal of the American Medical Informatics Association*, 16:387–394, 2009. 20

[13] Hugh Chipman and Robert Tibshirani. Hybrid hierarchical clustering with applications to microarray data. *Biostatistics*, 7:302–317, 2006. 66

[14] Bipasha Choudhury and Ingrid Gouldsborough. The use of electronic media to develop transferable skills in science students studying anatomy. *Anat Sci Educ*, 2012. 10

[15] James J Cimino, Anthony Aguirre, Stephen B Johnson, and Ping Peng. Generic queries for meeting clinical information needs. *Bulletin of the Medical Library Association*, 81(2):195–206, 1993. 18

[16] Nathan K Cobb, Amanda L Graham, and David B Abrams. Social network structure of a large online community for smoking cessation. *American Journal of Public Health*, 100(7):1282–1289, 2010. 17, 183

[17] David Constant, Lee Sproull, and Sara Kiesler. The kindness of strangers: The usefulness of electronic weak ties for technical advice. *Organization Science*, 7:119–135, 1996. 17

[18] Francisco M. Couto, Mario J. Silva, and Pedro M. Coutinho. Measuring semantic similarity between gene ontology terms. *Data & Knowledge Engineering*, 61:137–152, 2007. 52

[19] David G Covell, Gwen C Uman, and Phil R Manning. Information needs in the office practice: are they being met? *Annals of Internal Medicine*, 103(4):596–599, 1985. 1, 18

[20] M Dai, N.H. Shah, W. Xuan, M.A. Musen, S.J. Watson, B Athey, and F. Meng. An efficient solution for mapping free text to ontology terms. *AMIA Summit on Translational Bioinformatics,*, 2008. 21

[21] James A Danowski. Identifying networks of semantically-similar individuals from public discussion forums. pages 144–151, 2010. 17, 183

[22] Elnaz Davoodi, Keivan Kianmehr, and Mohsen Afsharchi. A semantic social network-based expert recommender system. *Applied Intelligence*, pages 1–13, 2012. 17

[23] Joshua C Denny, Jeffrey D Smithers, Randolph A Miller, and Anderson Spickard. understanding medical school curriculum content using knowledgemap. *JAMIA*, 10:351–362, 2003. 19

[24] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 55

[25] Patrick Doreian, Vladimir Batagelj, and Anuska Ferligoj. Generalized blockmodeling of two-mode network data. *Social Networks*, 26:29–53, 2004. 35

[26] Erlin, Norazah Yusof, and Azizah Abdul Rahman. Students' interactions in online asynchronous discussion forum: A social network analysis. pages 25–29, 2009. 17, 183

[27] Carole A. Estabrooks, David S. Thompson, J. Jacque E. Lovely, and Anne Hofmeyer. A guide to knowledge translation theory. *Journal of Continuing Education in the Health Professions*, 26:25–36, 2006. 11

[28] Gunther Eysenbach, John Powell, Marina Englesakis, Carlos Rizo, and Anita Stern. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *British Medical Journal*, 328, 2004. 9, 10

[29] Prasanna Ganesan, Hector Garcia-Molina, and Jennifer Widom. Exploiting hierarchical domain structure to compute similarity. *ACM Transactions on Information Systems*, 21(1):64–93, January 2003. 38, 46, 47, 60, 70, 183, 186

[30] Jeremy Grimshaw, Martin Eccles, and Jacqueline Tetroe. Implementing clinical guidelines. *Journal of Continuing Education in the Health Professions*, 24:S31–S37, 2004. 1

[31] Richard Grol. Successes and failures in the implementation of evidence-based guidelines for clinical practice. *Med Care*, 39:46–54, 2003. 1

[32] Robert A. Hanneman and Mark Riddle. *Introduction to social network methods.* University of California, Riverside, Riverside, CA, 2005. 26, 32

[33] Margaret M Hansen. Versatile, immersive, creative and dynamic virtual 3-d healthcare learning environments: A review of the literature. *Journal of Medical Internet Research*, 10(3), 2008. 8

[34] Ronald G Havelock. *Planning for innovation through dissemination and utilization of knowledge*. Center for Research on Utilization of Scientific Knowledge, Institute for Social Research, University of Michigan, 1976. 12

[35] Caroline Haythornthwaite. Social network analysis: An approach and technique for the study of information exchange. *Library & Information Science Research*, 18(4):323 – 342, 1996. 17

[36] Caroline Haythornthwaite, Barry Wellman, and Marilyn Mantei. Work relationships and media use: A social network analysis. *Group Decision and Negotiation*, 4:193–211, 1995. 17

[37] William R Hersh and Robert A Greenes. Saphire - an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. *Comput Biomed Res*, 23:410–425, 1990. 18

[38] Xiaodi Huang and Wei Lai. Clustering graphs for visualization via node similarities. *Journal of Visual Languages and Computing*, 17:225–253, 2006. 55

[39] Alicia Iriberri and Gondy Leroy. A life-cycle perspective on online community success. *ACM Comput. Surv.*, 41(2):1–29, 2009. 34, 188

[40] Jay J Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics*, 1997. 51

[41] Clement Jonquet, Nigam H. Shah, and Mark A. Musen. The open biomedical annotator. *Summit of Translational Bioinformatics*, pages 56–60, 2009. 19, 20, 21, 23

[42] Charles E Kahn and Daniel Rubin. Automated semantic indexing of figure captions to improve radiology image retrieval. *Journal of the American Medical Informatics Association*, 16:280–286, 2009. 20

[43] Leonard Kaufman and Peter J Rousseeuw. *Finding Groups in Data*. Wiley Interscience, 1990. 66

[44] Sung-jin Kim, Jong-yi Hong, and Eui-ho Suh. A diagnosis framework for identifying the current knowledge sharing activity status in a community of practice. *Expert Systems with Applications*, 39(18):13093–13107, 2012. 17

[45] Alison Kitson, Gill Harvey, and Brendan McCormack. Enabling the implementation of evidence based practice: a conceptual framework. *Quality in Health Care*, 7:149–158, 1998. 12, 13

[46] Jon M Kleinberg. Hubs, authorities, and communities. *ACM Computing Surveys*, 31, 1999. 32

[47] Johnathan Kuhn, Barbara Hasbargen, and Halina Miziniak. Pretest online discussion groups to augment teaching and learning. *Comput Inform Nurs*, 28:297–304, 2010. 10

[48] Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics*, 24(5):719–720, 2008. 65, 67

[49] Gondy Leroy and Hsinchun Chen. Meeting medical terminology needs–the ontology-enhanced medical concept mapper. *IEEE Transactions on Information Technology in Biomedicine*, 5(4):261–270, 2001. 18

[50] Yanyan Li and Ronghuai Huang. Analyzing peer interactions in computer-supported collaborative learning: Model, method and tool. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5169 LNCS:125–136, 2008. 17

[51] Yuhua Li, Zuhair A Bandar, and David McLean. An approach for measuring semantic similarity between words using multiple information sources. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4):871 – 882, july-aug. 2003. 49, 70

[52] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, 1998. 52

[53] Yongjing Lin, Wenyuan Li, Keke Chen, and Ying Liu. A document clustering and ranking system for exploring medline citations. *JAMIA*, 14(5):651–661, 2007. 66

[54] Jo Logan and Ian D Graham. Toward a comprehensive interdisciplinary model of health care research use. *Sci Commun*, 20:227–246, 1998. 12

[55] Henry J. Lowe and G. Octo Bamett. Micromesh: a microcomputer system for searching and exploring the national library medicines medical subject headings (mesh) vocabulary. *Proc Annu Symp Comput Appl Med Care*, pages 717–20, 1987. 18

[56] Susan Lowes, Peiyi Lin, and Yan Wang. Studying the effectiveness of the discussion forum in online professional development courses. *Journal of Interactive Online Learning*, 6(3):181–210, 2007. 17

[57] Leah P Macfadyen and Shane Dawson. Mining lms data to develop an "early warning system" for educators: A proof of concept. *Computers and Education*, 54(2):588–599, 2010. 17

[58] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2013. R package version 1.14.4 — For new features, see the 'Changelog' file (in the package source). 67

[59] Brendan McCormack, Allison Kitson, Gill Harvey, Jo Rycroft-Malone, Angie Titchen, and Kate Seers. Getting evidence into practice: the meaning of 'context'. *Journal of Advanced Nursing*, 38:94–104, 2002. 8, 12, 13, 30, 33

[60] Robin S McLeod, Helen M MacRae, Margaret E McKenzie, J Charles Victor, and Karen J Brasel. Evidence based reviews in surgery steering committee. a moderated journal club is more effective than an internet journal club in teaching critical appraisal skills: results of a multicenter randomized controlled trial. *Journal of the American College of Surgeons*, 211:769–776, 2010. 10

[61] George A Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41, 1995. 17, 191

[62] Randolph A. Miller, Filip M. Gieszczykiewicz, John K. Vries, and Gregory F. Cooper. Chartline: Providing bibliographic references relevant to patient charts using the umls metathesaurus knowledge sources. *Proc Annual Symposium of Comput Appl Med Care*, pages 86–90, 1992. 18

[63] Prakash Nadkarni, Roland Chen, and Cynthia Brandt. Umls concept indexing for production databases: a feasibility study. *JAMIA*, 8:80–91, 2001. 19

[64] Blair Nonnecke, Dorene Andrews, and Jenny Preece. Non-public and public online community participation: Needs, attitudes and behavior. *Electronic Commerce Research*, 6:7–20, 2006. 35

[65] Else Nygren. Simulation of user participation and interaction in online discussion groups. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7472 LNAI:138–157, 2012. 17

[66] Jerome A. Osheroff, Diana E. Forsythe, Bruce G. Buchanan, Richard A. Bankowitz, Barry H. Blumenfeld, and Randolph A. Miller. Physicians' information needs: analysis of questions posed during clinical teaching. *Annals of Internal Medicine*, 114(7):576–581, 1991. 1, 18

[67] Ulrike Pfeil, Knut Svangstu, Chee Siang Ang, and Panayiotis Zaphiris. Social roles in an online support community for older people. *International Journal of Human-Computer Interaction*, 27(4):323–347, 2011. 17

[68] Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on*, 19:17 –30, 1989. 49

[69] Traian Rebedea, Mihai Dascalu, Stefan Trausan-Matu, Dan Banica, Alexandru Gartner, Costin Chiru, and Dan Mihaila. Overview and preliminary results of using polycafe for collaboration analysis and feedback generation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6383 LNCS:420–425, 2010. 17

[70] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, 1995. 49, 57, 70

[71] Everett M Rogers. *Diffusion of Innovations*. Free Press, New York, 5th ed. edition, 1962. 8, 11, 12

[72] Jo Rycroft-Malone, Gill Harvey, Kate Seers, Allison Kitson, Brendan McCormack, and Angie Titchen. An exploration of the factors that influence the implementation of evidence into practice. *Journal of Clinical Nursing*, 13:913–924, 2004. xviii, 8, 12, 13, 30, 33, 187

[73] Mark A Schuster, Elizabeth A McGlynn, and Robert H Brook. How good is the quality of health care in the united states? *Milbank Q*, 76:517–563, 2003. 1

[74] Nigam H Shah, Nipun Bhatia, Clement Jonquet, Daniel Rubin, Annie P Chiang, and Mark A Musen. Comparison of concept recognizers for building the open biomedical annotator. *BMC Bioinformatics*, 10 (suppl 9):S14, 2009. 8, 19, 21

[75] Cuihua Shen and Peter Monge. Who connects with whom? a social network analysis of an online open source software community. *First Monday*, 16(6), 2011. 17

[76] Suresh Srinivasan, Thomas C. Rindflesch, William T. Hole, Alan R. Aronson, and James G. Mork. Finding umls metathesaurus concepts in medline. *Proc AMIA Symp*, pages 727–731, 2002. 19

[77] Michael A Stefanone and Geri Gay. Structural reproduction of social networks in computer-mediated communication forums. *Behaviour and Information Technology*, 27(2):97–106, 2008. 17

[78] Samuel Alan Stewart and Syed Sibte Raza Abidi. An infobutton for web 2.0 clinical discussions: The knowledge linkage framework. *IEEE Transactions on Information Technology in Biomedicine*, 16(1):129–135, 2012. 20, 33, 191

[79] Samuel Alan Stewart, Maia Elizabeth von Maltzahn, and Syed Sibte Raza Abidi. Comparing metamap to mgrep as a tool for mapping free text to formal medical lexions. In *Proceedings of the 1st International Workshop on Knowledge Extraction & Consolidation from Social Media*, 2012. 189

[80] Sharon E Straus, Jacqueline Tetroe, and Ian Graham. Defining knowledge translation. *CMAJ*, 181:165–168, 2009. 1

[81] Craig A. Struble and Chitti Dharmanolla. Clustering mesh representations of biomedical literature. In *HLT-NAACL 2004 Workshop: Biolink 2004, Linking Biological Literature, Ontologies and Databases*, pages 41–48, 2004. 67

[82] Toomas Timpka, Marie Ekstrom, and Per Bjurulf. Information needs and information seeking behavior in primary health care. *Scandanavian Journal of Primary Health Care*, 7(2):105–109, 1989. 1, 18

[83] Sergio L Toral, M Rocio Martinez-Torres, Federico Barrero, and Francisco Cortes. An empirical study of the driving forces behind online communities. *Internet Research*, 19(4):378–392, 2009. 17

[84] Wouter S Tuli. *IVF and Internet: Evaluation of an Interactive Personal Health Record for IVF Patients*. PhD thesis, Nijmegen, The Netherlands: Radboud University, 2008. 9

[85] Ruta K Valaitis, Noori Akhtar-Danesh, Fiona Brooks, Sally Binks, and Dyanne Semogas. Online communities of practice as a communication resource for community health nurses working with homeless persons. *Journal of Advanced Nursing*, 67:1273–1284, 2011. 10

[86] Tom H Van De Belt, Lucien JLPG Engelen, Sivera AA Berben, and Lisette Schoonhoven. Definition of health 2.0 and medicine 2.0: A systematic review. *Journal of Medical Internet Research*, 12(2), 2010. 8, 9

[87] Pamela Vercellone-Smith, Kathryn Jablokow, and Curtis Friedel. Characterizing communication networks in a web-based classroom: Cognitive styles and linguistic behavior of self-organizing groups in online discussions. *Computers and Education*, 59:222–235, 2012. 35

[88] Shenghui Wang and Paul Groth. Measuring the dynamic bi-directional influence between content and social networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6496 LNCS(PART 1):814–829, 2010. 17, 35

[89] Joe H Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963. 66

[90] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994. 26, 32

[91] Barry Wellman. *An electronic group is virtually a social network.* Lawrence Erlbaum, 1997. 17

[92] Barry Wellman, Janet Salaff, Dimitrina Dimitrova, Laura Garton, Milena Gulia, and Caroline Haythornthwaite. Computer networks as social networks: Collaborative work, telework, and virtual community. *Annual Review of Sociology*, 22:213–238, 1996. 9, 10

[93] Etienne Wenger. Knowledge management as a doughnut: Shaping your knowledge strategy through communities of practice. *Ivey Business Journal*, pages 1–8, 2004. 2, 13

[94] Etienne C. Wenger and William M. Snyder. Communities of practice: The organizational frontier. *Harvard Business Review*, pages 139–145, 2000. 8, 13, 188

[95] S. K. M. Wong, Wojciech Ziarko, and Patrick C. N. Wong. Generalized vector spaces model in information retrieval. In *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '85, pages 18–25, New York, NY, USA, 1985. ACM. 37, 48

[96] Illhoi Yoo and Xiaohua Hu. Biomedical ontology mesh improves document clustering qualify on medline articles: A comparison study. *19th IEEE International Symposium on Computer-Based Medical Systems*, pages 577–582, 2006. 66

[97] Meixun Zheng and Hiller Spires. Teachers' interactions in an online graduate course on moodle: A social network analysis perspective. *Meridian*, 13(2), 2011. 17

[98] Erping Zhu. Interaction and cognitive engagement: An analysis of four asynchronous online discussions. *Instructional Science*, 34(6):451–480, 2006. 17

[99] Ales Ziberna. Generalized blockmodeling of valued networks. *Social Networks*, 29(1):105–126, 2007. 35

[100] Qinghua Zou, Wesley W. Chu, Craig Morioka, Gregory H. Leazer, and Hooshang Kangarloo. Indexfinder: A method of extracting key concepts from clinical texts for indexing. *AMIA Annu Symp Proc*, pages 763–767, 2003. 19