IMPROVING OUR UNDERSTANDING OF MEDICINE 2.0 COMMUNITIES: COMBINING CONTENT AND SOCIAL NETWORK ANALYSIS

Samuel Alan Stewart[1], Syed Sibte Raza Abidi[1]

[1]NICHE Research Group, Dalhousie University, Halifax, Canada

Abstract

This paper presents an analytic framework of Social Network Analysis and Content Analysis for better understanding the Knowledge Translation activities within an online community of surgeons. Through core-periphery and agglomerative clustering methods we have identified a core group of power users that control the flow of conversation within the community. Semantic mapping to the MeSH lexicon has resulted in Knowledge Maps that provide insight into the content of the messages shared within the community, and these semantic mappings also provide a detailed look at the content of the conversation within the user clusters. The result is greater insight into how the community functions and who runs it, which will be a valuable resource moving forward for the list administrators.

Keywords

Social Network Analysis; Content Analysis; Semantic Mapping; Knowledge Translation; Online Communication

Introduction

The synthesis and dissemination of new knowledge to the existing medical community is vital to providing more effective health services and to strengthen the health care system [1]. The need for Knowledge Translation (KT) in healthcare is well established [2], but KT is often hampered by physical and temporal barriers between users, and these barriers can be overcome using online communication technologies. Online 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 knowledge translation. Web-based knowledge translation 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.

The Leveraging Internet for Knowledge Sharing (LINKS) model [3] provides a formal framework for developing knowledge-based online communities, but there is little known about how these communities function. The objective of this project is to develop analytic methods for better understanding the knowledge sharing dynamics of LINKS-guided online communities. The community will be analyzed in two dimensions. First, the communication patterns will be analyzed using Social Network Analysis (SNA) to determine who the leaders are within the community and where subgroups may exist using core-periphery and agglomerative clustering analysis. Second, the content of the communications themselves will be mapped to the Medical Subject Headings (MeSH) lexicon using a semantic mapping program called Metamap [4]. These mappings will be used to develop *Knowledge Maps*, or high-level summaries of the content being discussed within the community. The mappings will then be used to provide insight into the discussions within the subgroups identified through the SNA.

The methods in this paper will be tested on SURGINET, a medical mailing list comprised of general surgeons from around the world, and insights from the community will be presented.

Methods

The methods for this paper are broken into two distinct sections. First, SNA will provide insight into the leaders of the community, and second content analysis will provide insight into the content of the messages within the community.

Social Network Analysis (SNA)

SNA utilizes the principles of graph theory to represent communication networks in terms of actors (nodes) and ties between actors (edges) [5].

The structure of the network is a key component of the analytic process, and there is no accepted standard for how to design a network from a mailing list. This project will study the network as a 2-mode structure, in which there are two classes of nodes, users and threads, and a tie between a user and a thread indicates that a specific user has contributed a message to a thread. From this 2-mode network a 1-mode network of users can be created, in which a tie between two users has a number that indicates how many threads those two users have both communicated on.

Core-periphery analysis, or coreness, is a measure that tries to partition the community into two groups: a "core" group that performs the majority of the communication, and a "periphery" group that largely listens without contributing. Users are part of the core if they are well connected to other core members. With appropriate normalization this recursive definition can be solved using eigenvector decomposition which produces a "coreness" number on a [0,1] scale that measure how vital a specific user is to the core of the community [5].

Agglomerative clustering will also be applied to the 1-mode user network to attempt to identify potential subgroups of users based on their shared communications. The clustering will be done using the AGNES package in R using Ward's distance to produce disparate user clusters.

Content Analysis

The objective of the content analysis is to move beyond the communication patterns of the users to study what is being said. We will use semantic mappings of the content to a formal medical lexicon (MeSH) in order to investigate what is being said within the community. Metamap [4], developed by the National Library of Medicine, is a program that maps unstructured text to formal medical terms from the Universal Medical Language System (UMLS), or one of its component lexicons, such as MeSH. Each mapped term is assigned a score that is a measure of how strongly the term represents the source text on a [0,1000] scale [4].

Using the mapped terms we can investigate the knowledge base of the community. An online community is usually centred around a medical topic, but within those fields, 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 will be done using Knowledge Maps.

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. Combined 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.

Finally, the mappings can be used to provide insight into the content of the clusters detected in the SNA section, by investigating the most common terms within each cluster.

Results

SURGINET is a community of 865 clinicians from around the world that use the forum to discuss general surgical issues. The archives of the community from 2012-01-01 to 2013-04-05 were extracted, a sample that comprised 17,000 messages in 2,111 threads by 231 users. Once the threads were cleaned and non-medically relevant threads were removed the sample contained 13,404 messages on 948 threads with 50,597 total semantic mappings returned by Metamap.

SNA: Identifying Community Leaders

The core-periphery analysis results in a loose partitioning of the network into a core group of very active users and a periphery group of inactive users. Figure 1 presents the 1-mode network, re-organized such that the most active users (based on their coreness value) are at the centre of the network.

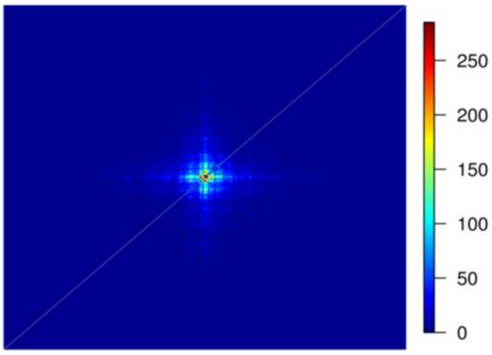

Fig. 1: The 1-mode network, sorted by coreness

As the figure demonstrates there is little traffic outside of the core group of active users. It is also worth noting that this figure only includes those 231 users that contributed at least one message, so there are an additional 634 users not included in this figure that have no ties to any other users. The actual definition of the core is variable, and table 1 presents the contributions of three different definitions of core.

Table 1: 3 Potential definitions of "core" and their contribution to the community

Cutoff	**Users**	**Messages**	**%**	**Threads**	**%**
0.84	26	9870	73.6	939	99.1
0.60	59	12143	90.6	947	99.9
0.52	77	12634	94.3	947	99.9

At its most strict definition the core of the community is comprised of 26 users that account for 71% of all the messages on the community and communicate on 99% of the threads (all by 9). As the definition of core is loosened more users are added that represent a larger proportion of all the messages. Regardless of what definition is used, the core is a relatively small group of users (3-9% of users overall) that comprise the bulk of the conversation within the community.

The agglomerative clustering was performed on the 1-mode network. The clustering resulted in 4 clusters, three small clusters and 1 large. The densities of the clusters are presented in table 2.

Tab. 2: Average number of shared threads between clusters

	A (18)	**B (21)**	**C (130)**	**D (26)**
A (18)	0.97	2.08	0.2	7.26
B (21)	2.08	3.53	0.39	13.6
C (130)	0.2	0.39	0.04	1.36
D (26)	7.26	13.6	1.36	49.52

The cluster analysis revealed four potential clusters, and confirmed the core-periphery structure from earlier. Cluster *D* has 26 users in it, the same 26 users that were at the core of the community in the previous analysis. These users have strong connections both to one another and to the other two small, somewhat active clusters, *A* and *B*. Cluster *C* represents the periphery, with 130 users that have very few connections to the rest of the community. Cluster *A* is also somewhat interesting, as it has a low inter-cluster density, but a somewhat strong connection to the core cluster *D*. Figure 2 is a re-arrangement of the adjacency matrix from figure 1 according to the clustering results.

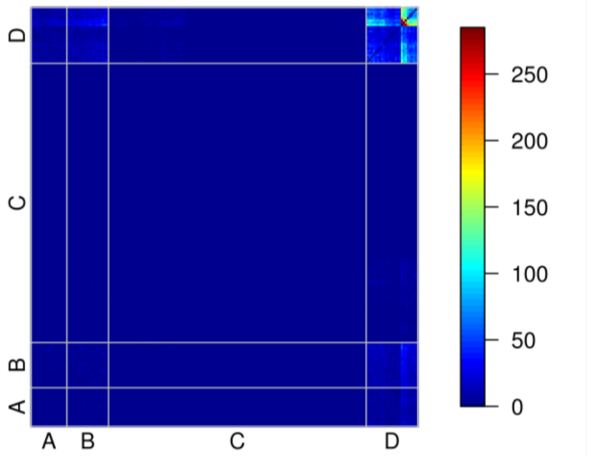

Fig. 2: The 1-mode network, ordered according to the clustering results

Content Analysis: Exploring Community Knowledge

Figure 3 presents the knowledge map for the SURGINET community, summarizing the overall mapping scores for the roots of the MeSH tree.

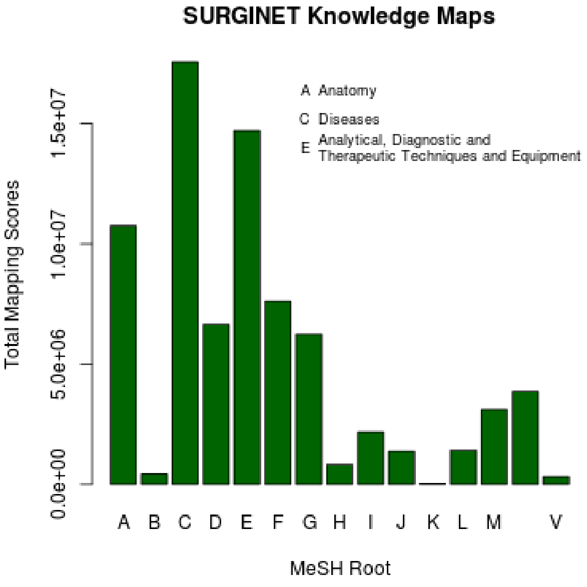

Fig. 3: Knowledge Maps for SURGINET data

The SURGINET users seem to spend the majority of their time discussing issues related to anatomy, specific diseases and techniques and equipment for analysis and diagnosis. Further investigation reveals that the anatomy issues (root *A*) focus on specific MeSH terms: *Breast*, *Back* and *Abdomen*, along with a series of terms related to the *Digestive System*. The disease mappings (*C*) are to a specific subgroup of the disease tree: *C23: Signs and Symptoms*, which suggests that a lot of effort is directed at the indications of disease within the list. Finally, the mappings to root *E* are related to specific surgical procedures, as would be expected on a surgical mailing list.

We can take the semantic mappings of the messages and use them to gain further insight into the structure of the core and periphery from the social network analysis. Table 3 presents the 10 most popular mappings for each of the four clusters.

Tab. 3: Most popular MeSH terms in each of the four main clusters

A	**B**	**D**
Fasting	Appendix	Duodenum
Hospitals	Laparotomy	Hernia, Inguinal
Comprehension	Reading	Fistula
Eating	Homosexuality	Carbidopa
Work	Appendicitis	Herniorrhaphy
Paper	Unemployment	Ileus
Drainage	Poverty	Needles
Wound Healing	Neoplasm Met.	Duodenum

Cluster *C* was omitted from the summary. There are very few messages in the cluster, so studying the content of those messages does not provide meaningful results. The contents of the three clusters here presents some significant separation between the cluster contents. Cluster *A* seems focused around hospital work, along with eating and wound management. Cluster *B* seems interested in the appendix and its surgical management, along with some non-medical content. Cluster *D* is the densest cluster and focuses largely around specific surgical discussions. This suggests that the most active subjects are those that are around very specific surgical content.

Conclusion

The SNA identified a core group of users that control the bulk of the communication traffic within the community. This is in line with previous research [6] that have investigated the "Pareto Principle" within online communities, or the hypothesis that 20% of the users do 80% of the work. It is not necessarily a bad thing that there is a core group of power users guiding the community, but knowing who these users are is important to controlling the future of the community.

The clustering was not overly successful. It correctly identified the core, but beyond that it found little clustering, with some loose clustering within the periphery. There may be potential for more sophisticated clustering methods to detect some patterns, but the problem may be that there are no inherent clusters within the community, which results in no clustering.

The knowledge maps were very successful in identifying what subject areas and what terms in particular were popular within the community. Knowing what is and is not being talked about is essential to directing future discussions within the community, and if there is material that is not being covered that the community administrators are interested in, they can now identify it and incorporate it into future conversations.

The synthesis of the SNA and content analysis methods in table 3 presents the full power of the framework. Using the content analysis we can now not only identify who the most active users are, but we can investigate what they are discussing, thereby determining what the community leaders are most interested in. What we found is a community that gathers around specific medical subjects and areas. This suggests that, while many subject areas are popular, the ones around specific procedures and processes are the most active.

Overall the analytic tools presented here provide a powerful addition to the LINKS model, and can provide real, tangible insight into the community as a whole.

Acknowledgement

The authors would like to thank Dr. Tom Gilas and the rest of the SURGINET community for their data. This project is partially funded by the Teasdale-Corti program.

References

[1] Sharon E Straus, Jacqueline Tetroe, and Ian Graham. Defining knowledge translation. CMAJ, 181:165–168, 2009

[2] Richard Grol. Successes and failures in the implementation of evidence-based guidelines for clinical practice. Med Care, 39:46–54, 2003.

[3] AbidiLINKS

[4] Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. Proceedings of the AMIA Symposium, pages 17–21, 2001.

[5] Stanley Wasserman and Katherine Faust. Social Network Analysis: Methods and Applications. Cambridge University Press, 1994.

[6] 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