JOURNAL OF COMPUTER-MEDIATED COMMUNICATION

Automated Delineation of Subgroups in Web Video: A Medical Activism Case Study

Alvin Chin*

Department of Computer Science, University of Toronto, Toronto, Canada (alvin.chin@utoronto.ca)

Jennifer Keelan
George Tomlinson

Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
(jenn.keelan@utoronto.ca, george.tomlinson@utoronto.ca)

Vera Pavri-Garcia

Division of Natural Sciences, York University, Toronto, Canada (pavri@yorku.ca)

Kumanan Wilson

Canada Research Chair in Public Health Policy, University of Ottawa, Ottawa, Canada
(kwilson@ohri.ca)

Mark Chignell

Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Canada
(chignell@mie.utoronto.ca)

Web 2.0 tools in general, and Web video in particular, provide new ways for activists to express their viewpoints to a broad audience. In this paper we deployed tools that have been used to find subgroups automatically in social networks and applied them to the problem of distinguishing between two sides of a controversial issue based on patterns of online interaction. We explored the problem of distinguishing between anti- and pro-vaccination activists based on a social network of videos and associated comments posted on YouTube. Videos for the analysis were selected by submitting the term "vaccination" to a search on YouTube. A content analysis of the selected videos was then performed (Keelan et al, 2007) to classify videos as pro- or anti-vaccination. Then, a modified version of the SCAN method (Chin and Chignell, 2008) for identifying cohesive subgroups in social networks was applied to the social network inferred from the discussions about the videos. Results showed that a cohesive subgroup of anti-vaccination people existed in discussions around anti-vaccination videos, whereas discussions around pro-vaccination videos included both anti-vaccination and pro-vaccination people. Implications of the method and results for

*Alvin Chin's current position is at the Nokia Research Center, Beijing, China. He may be contacted at alvin.chin@nokia.com. The work reported in this paper was done while he was a PhD student at the University of Toronto.

more general delineation of types of medical activism and the opposing camps within those camps are discussed.

Key words: medical activism; social network analysis, cluster analysis, Web Video, hierarchical clustering, subgroups.

doi:10.1111/j.1083-6101.2010.01507.x

Introduction

Web video has become a powerful tool for communicating ideas and expressing opinions. Activists are using Web videos, blogs, podcasts and a number of other tools to disseminate their ideas and to win over converts to any number of causes. In the case of healthcare, activism may sometimes be damaging, if high quality clinical evidence and policies set to benefit society as a whole are over-ruled based on emotional appeals and over-reliance on isolated cases. However, it should also be recognized that activism may sometimes have a salutary effect on medical practice.

Recently a number of healthcare issues have become a matter of debate between activists who question the current medical approach, and healthcare professionals who seek to maintain current medical practices. The research reported in this paper studied online medical activism using a case study of YouTube videos that can be divided into anti-vaccination and pro-vaccination videos. The research question asked was: can automated social network analytic methods distinguish between the pro and con sides of a medical issue based on social networks constructed from online interactions?

The particular medical issue examined in this case study was vaccination (immunization). While childhood immunization programmes are among the most successful public health interventions, and are based on compelling scientific evidence, public support for universal immunization continues to be undermined by controversy over vaccine safety and efficacy. Champions of a now discredited link between routine immunizations and autism have been successful in garnering public support to change public policy and stimulate public discourse about the safety of routine childhood immunization (Taylor et al, 1999; Hvid et al, 2003). Celebrity-advocates such as Jenny McCarthy have successfully mobilised public campaigns arguing for a moratorium on standard immunization practices. Organized resistance to routine childhood immunization and the defence of immunization by prominent scientists and public health organizations has led to two polarized camps, one comprised of staunch vaccine critics on the one hand and the other consisting of defenders of immunizations on the other. In order to answer the research question, a content analysis was used to distinguish the pro and anti stances of commentators on the vaccination videos, and then a modified version of the SCAN method (Chin and Chignell, 2008) was used in order to delineate subgroups of pro- and anti-vaccination people from online discussions concerning the videos. In addition to providing new

results concerning the applicability of the SCAN method to analysis of activist debates, the results of the case study that we carried out also provide insights into how video sharing is being used by activists in medical debates and in the debate surrounding the issue of vaccination in particular.

The remainder of this paper is structured as follows. After reviewing the Background research literature on vaccination, Web Video, and automated subgroup identification, the following section then briefly describes methods, of which the SCAN method was developed and used in the case study. The YouTube vaccination video case study is then presented, along with the results obtained by applying the SCAN method to the social network derived from online conversation concerning the vaccination-related videos. The implications of the findings are then discussed, followed by a short concluding section.

Background

The emergence of Web 2.0 (O'Reilly, 2005) has allowed for the easier formation and tracking of groups of people with common interests. The research reported in this paper is concerned with the use of Web video as a means to delineate subgroups of activists with respect to vaccination.

Vaccination on the Internet

The Internet is a frequently used source for health information by the public (Fox, 2006; Fox and Lee, 2000; Harris Interactive, 2002; Liszka et al., 2006). Polls from 2006 show that 80% of Internet users in the United States (Fox, 2006) and 27% of users in Great Britain have sought health-related information online (Pollard, 2006). Health care professionals have expressed concerns about the quality and veracity of information that individuals receive from Internet-based sources (Hardey, 1999; Kunst et al., 2002; Silberg et al., 1997; Walji et al., 2004). One particular area of concern is the use of Internet sites to communicate information to discourage the uptake of routine immunization. Four recent studies examining the quality of information about immunization on the World Wide Web found that search engines returned a high percentage of websites that promoted viewpoints that contradicted conventional medical opinion and advice (Davies et al., 2002; Nasir, 2000; Wolfe et al., 2002; Zimmerman et al., 2005). Davies et al. (2002) found that 43% of websites contained explicit anti-vaccination content and in another study, sites appearing in the top ten results for Internet searches frequently disseminated viewpoints critical of immunization (Nasir, 2000). All four studies found that the Internet is being used by vaccine critics to share health experiences, provide a forum for people who share similar medical opinions, and to disseminate alternative medical viewpoints. For example, Zimmerman et al. (2005) and Wolfe et al. (2002) found common design attributes on vaccine critic's websites indicating social networking, along with marketing features such as links to other anti-vaccination websites. Other features of anti-vaccination Websites included provisions for legal advice for avoiding

required immunizations, mechanisms to allow users to donate money to support the community, solicitation of personal stories, active listservs or chatrooms, and opportunities to purchase materials such as books and tapes.

Using YouTube Web Video For Sharing and Communication

Websites offering video streaming to users, such as YouTube (http://www.youtube.com), are becoming increasingly popular and those that allow user tagging, viewer rating, commenting and ranking provide a novel platform for sharing health information and enabling the formation of groups around certain health viewpoints (Cheng et al., 2007; Keelan et al, 2007; Lange, 2007). Sharing health information on YouTube is also becoming increasingly popular (Keelan et al, 2007; Khamsi, 2007). YouTube and other video services therefore have the potential to communicate health information to a large segment of the population. YouTube provides users with a new Internet-based tool to convey powerful images of both the risks and benefits of immunization (Keelan et al, 2007).

Electronic media such as Internet Relay Chat, web-based bulletin boards, and e-mail can be used to connect people together into a social network (Haythornthwaite, 2005). Previous research has examined whether social networks among YouTube users form when users interact with videos that are uploaded to YouTube (Geisler and Burns, 2007; Halvey and Keane, 2007; Lange, 2007). Other researchers have studied the communication patterns within video conversations (Molyneaux et al, 2008a; Molyneaux et al, 2008b). Halvey and Keane (2007) discovered through an analysis of the frequency distributions of views and number of uploaded videos on YouTube, that while many YouTube users did not participate in social networks, they did form social groups. In order to facilitate understanding of the discussion that follows, the terms social network and social group will now be defined. A social network is a graph where the edges between the nodes (typically people) reflect relationships between those nodes. More recently the term "social network" has been co-opted by social networking sites where the social network is a graph constructed by linking all the contact lists of site members together. For instance, if Person B was within the contact list of someone who was in turn in the contact list of Person A, Then Person B would be added to Person A's social network as a "friend of a friend". Note that in this type of social network, many of the members connected through friends of friends may not know each other, just as many guests at a wedding may not know each other even if they all know the bride or groom. A social group on the other hand, contains people who are linked by a shared purpose and who will generally know each other. Often there may be a group member list and their may be explicit criteria as to who is inside, or outside, the group. While the distinction between communities and groups is somewhat blurred, communities tend to be larger, and they may contain different social groups nested within them. Groups that coalesce around video sharing represent a particular kind of technology enabled group (Lange, 2007).

Social Network Analysis and Cohesive Subgroups

A social network is a graph where the edges between the nodes (typically people) reflect relationships between those nodes. More recently the term ''social network'' has been co-opted by social networking sites where the social network is a graph constructed by linking all the contact lists of site members together. For instance, if Person B was within the contact list of someone who was in turn in the contact list of Person A, Then Person B would be added to Person A's social network as a ''friend of a friend''. Note that in this type of social network, many of the members connected through friends of friends may not know each other, just as many guests at a wedding may not know each other even if they all know the bride or groom. A social group on the other hand, contains people who are linked by a shared purpose and who will generally know each other. Often there may be a group member list and there may be explicit criteria as to who is inside, or outside, the group. Groups that coalesce around video sharing represent a particular kind of technology enabled group (Lange, 2007). Wasserman and Faust (1994) defined a cohesive subgroup as a set of actors (nodes) that are relatively dense and directly connected through reciprocated (bi-directional) relationships (links). Previous research has shown that cohesive subgroups form communities of interest (Dixon, 1981), have weak ties (Garton, Haythornthwaite, and Wellman, 1997), and have cohesive bonds that bring people together (Piper, Marrache, Lacroix, Richardsen, and Jones, 1983).

A variety of methods have been used to identify subgroups in social networks, including in-degree screening for potential subgroups (Kumar, Raghavan, Rajagopalan, and Tomkins, 1999), content analysis of text and tags associated with Web pages (Flake, Lawrence, Giles, and Coetzee, 2002) and threaded conversations (Gruzd and Haythornthwaite, 2008), link analysis (Chau, Shiu, Chan and Chen, 2005), graph theory for finding densely-connected subgraphs within larger graphs (Gibson, Kumar, and Tomkins, 2005), and optimization (Tantipathananandh, Berger-Wolf and Kempe, 2007). However, previous research has tended to involve static networks, web pages as seeds, or links between text content, rather than links between people as reflected in their online interactions.

Social network analysis is a useful method for finding groups in online social networks (Garton, Haythornthwaite, and Wellman, 1997). Clique analysis and related methods look directly at the links that occur in a network and identify specific patterns of connectivity (e.g., subgroups where everyone in the subgroup has a direct connection to everyone else). Cliques and k-plexes have been used to characterize groupings in social networks (Alba, 2003; Balasundaram, Butenko, Hicks, and Sachdeva, 2007; Chin and Chignell, 2007, Du et al, 2007; Reffay and Chanier, 2003; Sterling, 2004; Wasserman and Faust, 1994), where in a clique (Wasserman and Faust, 1994) each member has a direct connection to every other member in the subgroup but in a k-plex, each member in the subgroup has direct ties to at least n-k other members where n is the number of members in the subgroup and k is a parameter. However, finding cliques and k-plexes in large networks is a computationally expensive and exhaustive process that is NP-complete

(Balasundaram, Butenko, Hicks, and Sachdeva, 2007), i.e., the computational effort scales exponentially with the number of nodes in the network.

Network centrality (or centrality) (Freeman, 1978) measures how important or central an individual node is to a network. People who are actively involved in one or more subgroups will generally score higher with respect to centrality scores for the corresponding network. Centrality has been used for identifying possible members of cohesive subgroups in networks derived from online interactions (Chin and Chignell, 2007b; Tyler, Wilkinson, and Huberman, 2005; Welser, Gleave, Fisher, and Smith, 2007). Clustering algorithms using centrality measures such as betweenness centrality (Freeman, 1978) have been used for repeatedly dividing the network into clusters and thus automatically identifying cohesive subgroups (Girvan and Newman, 2002; Gloor, Laubacher, Dynes, and Zhao, 2003; Marlow, 2004; Tremayne, Zheng, Lee, and Jeong, 2006; Tyler, Wilkinson, and Huberman, 2005). Betweenness centrality measures the extent to which a node can act as an intermediary or broker to other nodes (Freeman, 1978). In addition, closeness centrality (Crucitti, Latora, and Porta, 2006; Kurdia, Daescu, Ammann, Kakhniashvili, and Goodman, 2007; Ma and Zeng, 2003), and degree centrality (Fisher, 2005; Welser, Gleave, Fisher, and Smith, 2007; Chin and Chignell, 2007b), have also been used to infer community structure and find members of cohesive subgroups. Closeness centrality measures how many steps on average it takes for an individual node to reach every other node in the network, whereas degree centrality measures the number of direct connections that an individual node has to other nodes within a network (Freeman, 1978). Although network centrality measures are easy to calculate using social network analytic programs, there has been no consensus among researchers as to the most meaningful centrality measure to use for finding subgroup members (Costenbader and Valente, 2003).

Finding Cohesive Subgroups

While video content analysis by experts is the best way to assess the attitudes and goals of the people who post videos, video content analysis can be difficult, and time consuming due to its visual and temporal nature (Geisler and Burns, 2007). Videos often have to be watched all the way through before they can be classified, since some videos hide their real message until just before the end of the video. Thus an alternative to content analysis involves the use of data about patterns of tagging and video use (and associated conversations and discussions about vidos) to infer points of view and subgroup membership.

Automated methods for finding cohesive subgroups (subgroups of tightly connected people that interact with each other within the subgroup than with others outside of the subgroup) in online discussions around Web video should be useful for addressing needs such as the following:

- New users need a way to quickly find out what the main arguments are on each side of the controversy so that they can make up their minds about what to believe or do,

- Policy experts need to understand the structure of online debate so that they can better promote or advocate their viewpoint, and
- Researchers need a way to track changing patterns in how people view topics, and how they promote their ideas with respect to those topics.

The Social Cohesion Analysis of Networks method or SCAN (Chin and Chignell, 2008) was developed to identify cohesive subgroups within large social networks. This method involves selecting possible members of subgroups and filtering out the other members from the network (Select), collecting the possible members to form subgroups (Collect), and determining which of the subgroups are the most cohesive over time (Choose). A brief overview of the SCAN method will now be provided. A more detailed account of the method may be found in Chignell and Chin (2008).

Select

In the first step, the possible members of cohesive subgroups are identified. A cutoff value is selected on a measure that is assumed to be correlated with the likelihood of being a subgroup member, and those people who fail to reach the cutoff value on that that measure are removed from further consideration. Betweenness centrality has been used as the cutoff measure due to its comparatively good performance in previous research (Chin and Chignell, 2006; Girvan and Newman, 2002; Tyler, Wilkinson, and Huberman, 2005). Other centrality measures such as degree and closeness centrality may also be used. After filtering out those people that fail to reach the cutoff level of centrality in the social network, a list of potential active members of subgroups is obtained.

Collect

In the second step of the SCAN method, subgroups are formed using cluster analysis, specifically weighted average hierarchical clustering because of its computational efficiency (Chin and Chignell, 2008). This results in a set of nested, non-overlapping clusters in a tree-type format, called a dendrogram. The extraction of the hierarchy shows potential cohesive subgroups, but it does not actually partition the people into a particular set of non-nested subgroups.

Choose

For the Choose step, a similarity score is created that compares the cohesiveness of subgroups over time and determines the most appropriate cohesive subgroups using the highest similarity. The SCAN method implements the three steps (Select, Collect, Choose) Since the research in the present paper examines static networks over a single period of time, the Choose step is not applied.

Finding Cohesive Subgroups in Vaccination Videos on YouTube

In this section, we describe our method for finding cohesive subgroups in a set of vaccination videos on YouTube. First, the vaccination videos are classified and

characterized using content analysis (Keelan et al, 2007), and then a modified version of the SCAN method (Chin and Chignell, 2008) that omittedthe Choose step, was applied for automatically identifying cohesive subgroups from the conversations centred around the vaccination videos.

Using Content Analysis to Classify and Characterize Vaccination Videos on YouTube

The analysis of videos relating to vaccination on YouTube as described in Keelan et al(2007) shows that Web video is being used by activists to challenge the practice of routine immunization. While the arguments contained in these videos are similar to those found in other Internet sources, the video medium offers significant new rhetorical, emotive and artistic tools for persuasion. It allows users to share their own personal experiences with immunization and to disseminate their beliefs about immunization in a more intimate and interactive fashion. Face-to-face interviews with those purportedly injured by vaccination are posted with extensive background text detailing the risks of vaccination and alleged collusion between pharmaceutical companies and medical experts to cover up side-effects from vaccination. Using the features in YouTube, viewers can rate the videos, add supportive commentary, or challenge the information or health experience presented. Users can also share mainstream media clips or annotate them to make particular points.

Anti-vaccination groups are active in speaking out against the use of vaccines. In addition to their online activity, they sometimes hold meetings in different cities. Opposed to the anti-vaccination group is a pro-vaccination group that primarily contains medical doctors, public health experts, and drug companies that argue in favour of vaccination. The groups use whatever electronic means that they can to convey their message, with the anti-vaccination group being particularly aggressive in trying to get their message across.

In spite of the allianceof healthcare professionals and drug companies arrayed against them, the anti-vaccination group seemed to achieve a significant presence on YouTube during the period in which this research was conducted, with searches on YouTube video for the term ''vaccination'' typically returning almost as many videos relating to anti-vaccination than those that related to pro-vaccination (as reported by Keelan et al, 2007).

Content Analysis for Classifying Videos

A content analysis was performed on 339 YouTube videos retrieved in early 2007 using the keywords ''vaccination'' and ''immunization'' in the detailed description of the videos. Detailed analysis was carried out on 174 of those videos. The videos were classified based on methods used in earlier studies examining information about immunization on the World Wide Web and the vaccine criticism movement (Wolfe and Sharp, 2005; Wolfe et al., 2002; Zimmerman et al., 2005).

Videos were categorized as ''negative'' (anti-vaccination) if the main message of the video portrayed immunization negatively and ''positive'' (pro-vaccination) if the central message supported immunization, portraying it positively (refer to

Table 1 Criteria for Classifying Videos

Video type	Characteristics of video	Examples of video type
Negative (anti-vaccination)	• Emphasizes the risk of immunization • Advocates against immunizing • Argues immunization science cannot be trusted • Alleges collusion between supporters of vaccination and manufacturers • Alleges conspiracy or cover up of serious adverse events	Personal video diaries, commentaries, multi-media clips of news segments and documentaries
Positive (pro-vaccination)	• Describes the benefits and safety of immunizing (balance of risks) • Describes immunization as a social good • Encourages people to receive immunizations	Personal video diaries, personal commentaries, multi-media clips of news segments, public service announcements, tutorials and manufacturer-sponsored commercials
Ambivalent	• Indicates that immunization is a routine practice/social norm • Focuses on the pain of immunizing a child • Reveals parental anxieties over immunization of children	Personal video diaries
Debate	• Central message of the video was a debate over immunization • An effort was made to present different viewpoints or "both sides" of a controversy	Primarily news segments or small-scale media productions (e.g., a university debate club)

Table 1). Based on the content analysis, two additional types of video were identified. "Ambiguous" videos had conflicting information (i.e., a beneficial/social good was countered by negative experiences such as anxious parents and crying infants) while debates attempted to portray both sides of the argument. The videos were reviewed and classified independently by two experts (JK & VP). The weighted kappa for agreement on classification of videos was 0.93 showing a high level of reliability for the classification scheme.

After categorizing the videos, user interactions with these videos were compared using view-counts and viewer reviews indicated by the star-rating system from 1 to 5 stars (1 = "Poor" to 5 = "Awesome"). 73 (48%) of the videos were classified as positive (i.e., pro-vaccination), 49 (32%) were negative (i.e., anti-vaccination), and

31 (20%) were ambiguous. YouTube also allows for users to rate the video on a 5-star scale. We discovered that negative (i.e., anti-vaccination) videos were more likely to receive a rating, and had mean higher star ratings and more views, than positive videos.

Classification of Vaccination Videos Using Tag Analysis

In order to characterize and classify what the anti- and pro-vaccination videos are generally about, we performed a tag analysis by recording all the tags of each video and then doing a frequency distribution of tags for the videos. Tags in YouTube are words that members use for characterizing and describing the video, for example if a video is about a child getting a polio vaccine, possible tags could be "vaccine", "vaccination", "polio". Adding tags to video is designed to make the search process easier and quicker to find that type of video because it adds semantics to the video. Tags are a form of informal index intended to make video easier to find. From counting the frequency of each of the tags from the anti-vaccination videos (190 tags in total), and pro-vaccination videos (294 tags in total), we discovered that some of the most popular tags for anti-vaccination videos deal with disputed harmful effects or agents of harm related to vaccination (as expected) with the top tags being "autism", "mercury", "thimerosal" and "HPV". Autism is a harmful effect (Wilson et al, 2003), mercury and thimerosal are disputed agents supposedly harmful agents (according to anti-vaccination activists)., and HPV is the human papilloma virus where.a new vaccine against HPV was released in this time period. For pro-vaccination videos, most of the popular tags deal with vaccines and the illnesses that vaccines address such as "HPV" (papilloma virus), "cancer", "influenza", and "flu". There were general tags that were shared with the pro-vaccination and anti-vaccination videos, these included "vaccine", "vaccination" and "health".

Thus content analysis can be used to identify the vaccination video as anti-vaccination or pro-vaccination, and how tag analysis can be used to create meta-data that characterize and describe the content and semantics of the anti-vaccination and pro-vaccination videos respectively. The next part will take this classification of videos and look into identifying cohesive subgroups by examining the comments to the videos.

Identifying Subgroups Automatically in Vaccination Video Conversations

Once the vaccination videos were classified, the next step is to analyze the social interactions occurring via conversations centered around the videos. Only the pro- and anti-vaccination videos from the original sample were chosen for the data analysis since our purpose was to delineate subgroups of pro- and anti-vaccination people behind the video conversations. A crawler (i.e., a programme used to harvest information from online sites) was written that automatically traversed each of the videos from the list of videos identified by Keelan et al(2007. The crawler then generated the social network graph for each of the pro- and anti-vaccination videos using the following process. For each video posted by user u, a link from v to u

Table 2 Characteristics of each of the generated video conversation social networks

Network	# of members	# of videos	# of comments	# of tags
Anti-vaccination	177	34	217	190
Pro-vaccination	222	66	246	294

is inferred for each user v that made a comment to the video posted by user u. Additional links are inferred between the person that made a comment (v) and the person who made the preceding comment (x) in reverse time chronological order, with respect to a particular video. The links inferred in the above manner results in a directed graph (social network) where each edge in the graph is a directed connection from v to u or from v to x and w is the number of replies between v and u or between v and x. The names in the study have been anonymized in accordance with the ethics protocol filed with the University of Toronto Research Ethics Board.

The dataset comprised video spanning a time period of 30 months starting from the earliest time when the video was first posted on YouTube (April 2006) until the time of the crawl (September 2008) table 2 summarizes social networks generated from running the crawler on the anti- and pro-vaccination videos and their comments respectively, are shown in Table 2 along with the number of tags for all the videos.

The number of videos for anti-vaccination and pro-vaccination video conversations is less than the number of videos analyzed in Section 4.1.1 by Keelan et al(2007) because the video conversations were analyzed at a later time period than the videos themselves, with some of the videos having been removed from YouTube in the interim. The number of comments indicates the total number of comments for all videos during that time period, and the number of tags is the total number of tags that was posted for all videos.

Applying the SCAN Method: Select

Cutoff betweenness centrality can be used to select the active YouTube members, following the Select step of the SCAN method (Chin and Chignell, 2008). Since we wanted to delineate subgroups of anti-vaccination and pro-vaccination members, the commenters had to be labeled as having an anti-vaccination or pro-vaccination stance. This labeling of commenters was performed by having two independent coders read the comments and classify them as either anti-vaccination or pro-vaccination. A level of agreement for each comment was then calculated. If both coders classify the comment as anti-vaccination or pro-vaccination, then the level of agreement was ''Agree''. If both coders classified the comment differently (one considers as anti-vaccination and the other as pro-vaccination), then the level of agreement was ''Disagree''. Other comments, where at least one of the coders was unsure if the comment is anti-vaccination or pro-vaccination, were discarded from the sample. A commenter was then considered as an anti-vaccination member or pro-vaccination member if both coders agreed that the comment was positioned

against, or in favour of, vaccination, respectively. Anti-vaccination comments were more frequent, for both anti-vaccination and pro-vaccination videos, demonstrating that the anti-vaccination people were more active and vocal than the pro-vaccination people, regardless of the type of video.

Applying the SCAN Method: Collect

Figures 1a and 1b illustrate the social networks generated from comments on YouTube anti-vaccination and pro-vaccination videos after being processed usingthe Select step and then anonymized,. These figures show visual evidence of subgroup formation based on the conversational interactions around the videos, where the visualization is performed using the spring-embedded force algorithm in the NetDraw social network analytic software (Borgatti, 2002). From the Select step, we use circles and the prefix 'Anti' to indicate anti-vaccination members and squares and the prefix 'Pro' to indicate pro-vaccination members.

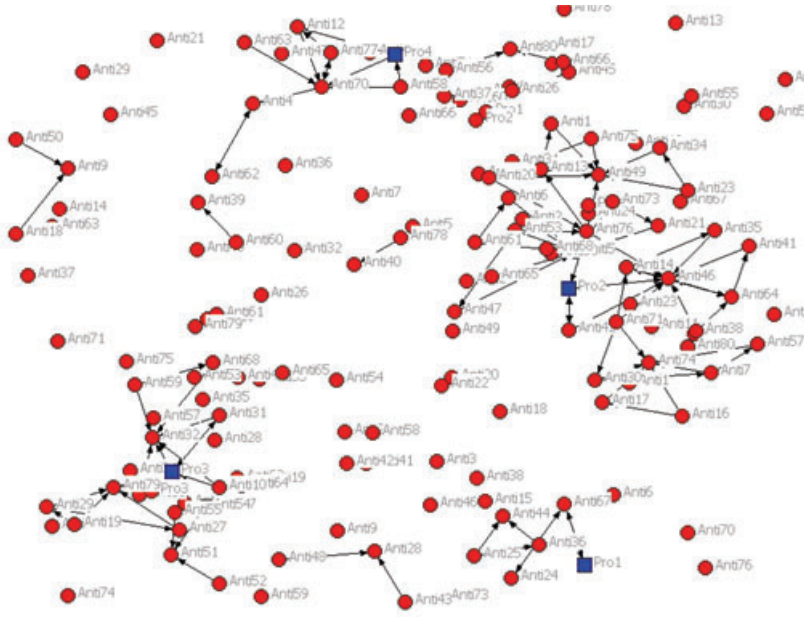
It can be seen from Figure 1a that anti-vaccination members are present in the comment discussions around anti-vaccination videos and there are very few pro-vaccination members (as expected). However, from Figure 1b, almost equal numbers of anti-vaccination members and pro-vaccination members are present within the comment discussions around pro-vaccination videos.

Next, weighted average hierarchical clustering was performed on the networks to find cohesive subgroups to delineate anti-vaccination members and pro-vaccination members. A dendrogram was created for both anti-vaccination and pro-vaccination video conversation networks, where the members were anonymized and identified as either anti-vaccination or pro-vaccination in order to easily determine the order of clustering for anti- and pro- members. From the dendrograms (omitted here), we discover that the anti-vaccination members in the anti-vaccination video conversations form subgroups that generally contain only anti-vaccination members, whereas for pro-vaccination videos, pro-vaccination and anti-vaccination members were combined together based on the conversations that occur around the videos, as shown visually on the graph in Figure 1b.

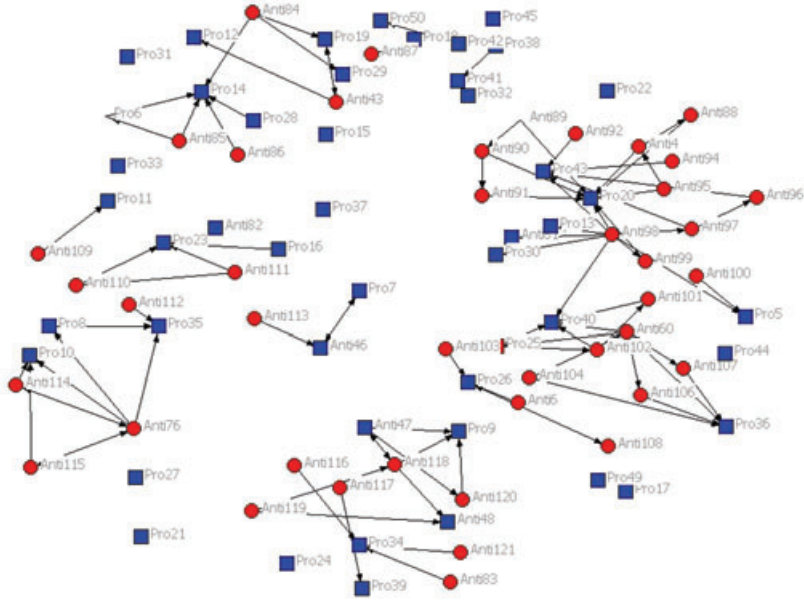
Anti-vaccination members were the most important or influential members in the anti-vaccination video comments network and they were also very active in the pro-vaccination video comments network.

Discussion and Implications

In the case study that was carried out it was impossible to identify cohesive subgroups of anti-vaccination and pro-vaccination people respectively using only the modified version of the SCAN method in an unsupervised learning fashion. Additional content analysis was required to disentangle the subgroups. Since activist debate typically involves discussions amongst heterogenous people, methods that consider only graph structure perform poorly in looking for subgroups in such debates, even though those methods have been found to perform quite well with topic-oriented and



(a) Comment discussions network from YouTube anti-vaccination videos



(b) Comment discussions network from YouTube pro-vaccination videos

Figure 1 Social networks generated from comment discussions to (a) YouTube antivaccination videos and (b) pro-vaccination videos using content analysis to classify comments

interest-driven (Chin and Chignell, 2008) interaction. In general, pro-vaccination videos have more comments and tags than anti-vaccination videos and have more members in their networks, due to the larger number of pro-vaccination videos that were crawled than anti-vaccination videos and also because of the intense debates that pro-vaccination videos caused, compared to anti-vaccination videos. Subgroups of anti-vaccination people exclusively arise from conversations around anti-vaccination videos with no subgroups of pro-vaccination people, whereas pro-vaccination video conversations generate subgroups that contain a mixture of both anti-vaccination and pro-vaccination members. In addition, the most influential and important members can be identified from the subgroups using betweenness centrality. Almost all of the important members in the anti-vaccination video conversations have an anti-vaccination stance, in contrast to pro-vaccination videos where there is considerable activity by both anti- and pro-vaccination people.

In this case study it was found that pro-vaccination members were infiltrated (and perhaps drowned out to some extent) by the anti-vaccination members who were more vocal and active. While it was difficult without analyzing the content of the discussion to automatically differentiate between pro and anti- people in pro-vaccination video conversations , when someone commented on an anti-vaccination video, it was a strong indicator that the person who made the comment had an anti-vaccination attitude.

Conclusions and Future Work

In the case study reported here, a method based only on social network analysis failed to clearly delineate opposing sides of a medical debate. However, labeling groups with the tags that people use provided an indication of which topics and issues divide the opposing camps.

The results in this paper demonstrate the value and pitfalls of studying and identifying cohesive subgroups in online health communities. One area of research which we did not explore is the tracking of changing patterns in how people view topics and the evolution of subgroup members using similarity measures as explained in the Choose step of the SCAN method. Since the SCAN method was only tested on one example of health-related online community (the YouTube vaccination videos), it is not possible to generalize the SCAN method to all health-related online communities. Thus, a second area for future work is to test the method with other health-related online communities and media (besides video). Another area for future research is the development of methods that combine content analysis and social network analysis in delineating subgroups and tracking the evolution and behaviour of online communities.

Acknowledgements

We would like to acknowledge the assistance of Bell University Laboratories in funding our research on Web video and social network analysis. We would also like

to thank the coders, Andrew Chignell and Sanny Chen, for classifying the vaccination videos and associated comments into anti-vaccination and pro-vaccination.

References

Alba, R. D. (2003). A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology*, *3*, 113–126.

Balasundaram, B., Butenko, S., Hicks, I., and Sachdeva, S. (2007). *Clique relaxations in social network analysis: The maximum k-plex problem. Technical report,* Texas A and M Engineering.

Bird, C. (2006). Community structure in oss projects [online]. *Technical report*, University of California Davis.

Borgatti, S. et al. (2002). *Ucinet for Windows: Software for Social Network Analysis*. Analytic Technologies.

Borgatti, S. (2002). *Netdraw. Retrieved* from http://www.analytictech.com/downloadnd.htm [accessed 30 September 2008].

Chau, M., Shiu, B., Chan, I., and Chen, H. (2005). Automated xidentification of web communities for business intelligence analysis. In *Proceedings of the Fourth Workshop on E-Business (WEB)* (New York, NY, USA, 2005).

Cheng, X., Dale, C. and Liu, J. (2007). "Understanding the Characteristics of Internet Short Video Sharing: YouTube as a Case Study." *CoRR* abs/0707.3670.

Chin, A., and Chignell, M. (2006). A social hypertext model for finding community in blogs. In *Proceedings of the 17th International ACM Conference on Hypertext and Hypermedia: Tools for Supporting Social Structures* (Odense, Denmark), ACM, 11–22.

Chin, A., and Chignell, M. (2007a). Identifying subcommunities using cohesive subgroups in social hypertext. In *HT '07: Proceedings of the eighteenth conference on Hypertext and hypermedia* (New York, NY, USA), ACM, 175–178.

Chin, A., and Chignell, M. (2007b). Identifying active subgroups within online communities. In *Proceedings of the 17th IBM Centre for Advanced Studies Annual International Conference on Computer Science and Software Engineering (CASCON 2007)*, 280-283.

Chin, A., and Chignell, M. (2008). Automatic detection of cohesive subgroups within social hypertext: A heuristic approach. *New Review of Hypermedia and Multimedia*, *14* (1), 121–143.

Costenbader, E., and Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social Networks*, *25*, 283–307.

Crucitti, P., Latora, V., and Porta, S. (2006). Centrality measures in spatial networks of urban streets. Physical Review E *73*, 036125.

Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W. (1992). Scatter/gather: a cluster-based approach to browsing large document collections. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval* (Copenhagen, Denmark), ACM, 318–329.

Davies, P., Chapman, S., and Leask, J. (2002). "Antivaccination activists on the world wide web." *Archives of Disease in Childhood*, *87*(1), 22–25.

Dixon, J. (1981). Towards an understanding of the implications of boundary changes - with emphasis on community of interest, draft report to the rural adjustment unit. *Technical report*, University of New England.

Du, N., Wu, B., Pei, X., Wang, B., and Xu, L. (2007). Community detection in large-scale social networks. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (New York, NY, USA, 2007), ACM, 16–25.

Fisher, D. (2005). Using egocentric networks to understand communication. *IEEE Internet Computing*, *9*:5, 20–28.

Flake, G. W., Lawrence, S., Giles, C. L., and Coetzee, F. M. (2002). Self organization and identification of web communities. *IEEE Computer 35*:3, 66–71.

Fox, Suzanna. (2006). "Most internet users start at a search engine when looking for health information online. Very few check the source and date of the information they find.", *Pew Internet & American Life Project*.

Fox, Suzannah, and Lee, Rainie. (2000). "The Online Health Care Revolution: How the Web helps Americans take better care of themselves." *Pew Internet & American Life Project*.

Freeman, L. C. (1978). Centrality in social networks: Conceptual clarification. *Social Networks*, *1*, 215–239.

Garton, L., Haythornthwaite, C., and Wellman, B. (1997). Studying online social networks. *Journal of Computer-Mediated Communication*, *3*(1), 1–30.

Geisler, G., and Burns, S. (2007). Tagging video: conventions and strategies of the youtube community. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (New York, NY, USA, 2007), ACM, 480–480.

Gibson, D., Kumar, R., and Tomkins, A. (2005). Discovering large dense subgraphs in massive graphs. In *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, VLDB Endowment, 721–732.

Girvan, M., and Newman, M. E. (2002). Community structure in social and biological networks. *PROC.NATL.ACAD.SCI.USA*, *99*, 7821.

Gloor, P. A., Laubacher, R., Dynes, S. B. C., and Zhao, Y. (2003). Visualization of communication patterns in collaborative innovation networks - analysis of some w3c working groups. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management (New York, NY, USA, 2003)*, ACM Press, 56–60.

Gorski, D. (2008). Jenny McCarthy, Jim Carrey, and "Green Our Vaccines": Anti-vaccine, not "pro-safe vaccine". *Science-Based Medicine*, http://www.sciencebasedmedicine.org/?p=139.

Gruzd, A., and Haythornthwaite, C. (2008). Automated discovery and analysis of social networks from threaded discussions. Paper presented at the *International Network of Social Network Analysts*.

Halvey, M. J., and Keane, M. T. (2007). Exploring social dynamics in online media sharing. In *WWW '07: Proceedings of the 16th international conference on World Wide Web* (New York, NY, USA, 2007), ACM, 1273–1274.

Hanneman, R. A., and Riddle, M. (2005). Introduction to social network methods (online textbook). University of California, Riverside, CA.

Hardey, Michael. (1999). "Doctor in the house: the Internet as a source of lay health knowledge and the challenge to expertise." *Sociology of Health & Illness*, *21*(6), 820–835.

Harris Interactive. (2002). "Cyberchondriacs Update:110 million people sometimes look for health information online, up from 97 million a year ago. On average, they do so three times a month Most use a portal or search engine." *The Harris Poll 21*.

Haythornthwaite, C. (2005). Social networks and internet connectivity effects. *Information, Communication and Society*, *8*:2, 125–147.

Hvid et al. (2003). Association between thimerosal-containing vaccines and autism, *JAMA*, *290*:1763–66.
Keelan, J., Pavri-Garcia, V., Tomlinson, G., and Wilson, K. (2007). Youtube as a source of information on immunization: A content analysis. *JAMA: The Journal of the American Medical Association* 298 (21), 2482–2484.
Khamsi, Roxanne. (2007). "Is YouTube just what the doctor ordered?" *New Scientist*, 1 April 2007.
Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). Trawling the web for emerging cyber-communities. *Computer Networks*.
Kunst, Heinke, Diederik Groot, Pallavi M. Latthe, Manish Latthe and Khalid S. Khan. (2002). "Accuracy of information on apparently credible websites: survey of five common health topics." *BMJ*, *324*(7337), 581–582.
Kurdia, A., Daescu, O., Ammann, L., Kakhniashvili, D., and Goodman, S. (2007). Centrality measures for the human red blood cell interactome. *Engineering in Medicine and Biology Workshop*, IEEE Dallas (Nov. 2007), 98–101.
Lange, P. (2007). "Publicly Private and Privately Public: Social Networking on YouTube." *Journal of Computer Mediated Communication*, 13(1), article 18.
Liszka, Heather A., Terrence E. Steyer and William J. Hueston. (2006). "Virtual medical care: how are our patients using online health information?" *Journal of Community Health*, *31*(5), 368–378.
Ma, H.-W., and Zeng, A.-P. (2003). The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, *19*:*11*, 1423–1430.
Manning, C. D., Raghavan, P., and Schutze, H. (2007). Introduction to Information Retrieval. Cambridge University Press, New York, NY.
Marlow, C. (2004). Audience, structure and authority in the weblog community. In *International Communication Association Conference* (New Orleans, LA).
Molyneaux, H., Gibson, K., ODonnell, S., and Singer, J. (2008). New visual media and gender: A content, visual and audience analysis of youtube vlogs. In *Proceedings of the International Communication Association Annual Conference (ICA 2008)*.
Molyneaux, H., ODonnell, S., Gibson, K., and Singer, J. (2008). Exploring the gender divide on youtube: An analysis of the creation and reception of vlogs. *American Communication Journal*, 10.
Nasir, L. (2000). "Reconnoitering the antivaccination web sites: news from the front." *Journal of Family Practice*, *49*(8), 731–733.
Newman, M. E. (2006). Modularity and community structure in networks. *In Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582.
Nielsen/NetRatings. (2006). "The hottest online brands in 2006: User generated content dominates the fasting growing online brands in the UK." *Nielsen/NetRatings*.
O'Reilly, T. What is web 2.0? Retrieved from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html [accessed 30 September 2008], 2005.
Piper, W. E., Marrache, M., Lacroix, R., Richardsen, A. M., and Jones, B. D. (1983). Cohesion as a Basic Bond in Groups. *Human Relations*, *36*:2, 93–108.
Pollard, M. (2006). "Internet Access: Households and Individuals." London: Office for National Statistics.
Reffay, C., and Chanier, T. (2003). How social network analysis can help to measure cohesion in collaborative distance learning. In *Proceedings of Computer Supported Collaborative Learning 2003 (Kluwer Academic Publishers, 2003)*, ACM Press, 343–352.

Silberg, W.M., Lundberg, G.D. and Musacchio, R.A. (1997). "Assessing, Controlling, and Assuring the Quality of Medical Information on the Internet." *JAMA*, *277*(15), S1244–1245.

Sterling, S. (2004). Aggregation techniques to characterize social networks. Master's thesis, Air Force Institute of Technology.

Tantipathananandh, C., Berger-Wolf, T. Y., and Kempe, D. (2007). A framework for community identification in dynamic social networks. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 2007), ACM, 717–726.

Taylor et al.(1999), Autism and measles, mumps, and rubella vaccine: no epidemiological evidence for a causal association, *Lancet* 1999 *353*:2026–29.

Tremayne, M., Zheng, N., Lee, J. K., and Jeong, J. (2006). Issue publics on the web: Applying network theory to the war blogosphere. *Journal of Computer-Mediated Communication*, *12*:1.

Tyler, J. R., Wilkinson, D. M., and Huberman, B. A. (2005). E-mail as spectroscopy: Automated discovery of community structure within organizations. *The Information Society*, *21*:2, 143–153.

Walji, M., Sagaram, S., Sagaram, D., Funda, M.B., Johnson, C., Nadeem Mirza, Q. and Elmer Bernstam, V. (2004). "Efficacy of Quality Criteria to Identify Potentially Harmful Information: A Cross-sectional Survey of Complementary and Alternative Medicine Web Sites." *J Med Internet Res 6*(2), e21.

Wasserman, S., and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Welser, H. T., Gleave, E., Fisher, D., and Smith, M. (2007). Visualizing the signatures of social roles in online discussion groups. *Journal of Social Structure*, 8.

Wikipedia. (2009). Definition of NP-complete. http://en.wikipedia.org/wiki/NP-complete.

Wilson, K et al. (2003). Association of autistic spectrum and the measles, mumps, and rubella vaccine: a systematic review of current epidemiological evidence. *Archives of Pediatric Adolescent Medicine* 2003;*157*(7): 628–34.

Wolfe, R. M. and Sharp, L.K. (2005). "Vaccination or immunization? The impact of search terms on the internet." *Journal of Health Communication 10*(6), 537–551.

Wolfe, R. M., Sharp, L. K. and Lipsky, M. S. (2002). "Content and design attributes of antivaccination web sites." *JAMA*, *287*(24), 3245–3248.

Zimmerman, R. K., Wolfe, R. M., Dwight, E. F., Fox, R., Nowalk, M. P. and Troy, J. (2005). "Vaccine Criticism on the World Wide Web." *Journal of Medical Internet Research*, *7*(2), e17.