Understanding Cancer-based Networks in Twitter using Social Network Analysis

Dhiraj Murthy Alexander Gross
Sociology Department, Social Innovation Laboratory
Bowdoin College
Brunswick, ME
{dmurthy, agross}

Daniela Oliveira
Computer Science Department
Bowdoin College
Brunswick, ME
doliveir@bowdoin.edu

Abstract—**Web-based social media networks have an increasing frequency of health-related information, resources, and networks (both support and professional). Although we are aware of the presence of these health networks, we do not yet know their ability to (1) influence the flow of health-related behaviors, attitudes, and information and (2) what resources have the most influence in shaping particular health outcomes. Lastly, the health research community lacks easy-to-use data gathering tools to conduct applied research using data from social media websites. In this position paper we discuss and sketch our current work on addressing fundamental questions about information flow in cancer-related social media networks by visualizing and understanding authority, trust, and cohesion. We discuss the development of methods to visualize these networks and information flow on them using real-time data from the social media website Twitter and how these networks influence health outcomes by examining responses to specific health messages.**

Keywords: social network analysis, virtual social networks, cancer-based Twitter networks, data visualization, trust inference, information-flow.

I. Introduction

Traditionally patients have received health-related information by meeting personally with their doctors or medical staff and could only share details of their condition to their families or to someone close to them. As the popularity of online social network services has increased, individual patients, their families, and their caregivers have bypassed the traditional controls of the healthcare and life science industries by volunteering private information about themselves on publicly accessible Internet sites. Additionally, they have become more open to considering health messages on the sites. According to data from the 2007 Health Information National Trend Survey (HINTS), 23% of respondents reported using a social networking site [1]. As Orsini [2] observes, people are able to use new media to create support communities such as those found at websites such as Patients-LikeMe. Chou et al. [1] found that cancer-related 'secondary audiences', family members of individuals who have/had cancer, have a high prevalence of social media use. This is unsurprising given that 61% of adult Americans look online for health information [3]. Of these 'e-patients', 41% 'have read someone else's commentary or experience about health or medical issues on an online news group, website, or blog' [3]. Additionally, 15% of e-patients 'have posted comments, queries, or information about health or medical matters' [3].

In particular, Twitter, a popular social media site, has had a recent impact on the ways in which health information and resources are shared. It is a microblogging (i.e. short message) service that enables its users to send and read short tweets. Rather than employing the taxonomy of 'friends', Twitter has 'followers' and 'followees'. A follower is someone who considers the followee interesting for any reason. The relationships are often asymmetric, consisting of unidirectional (arcs) as 'follower' users choose to 'follow' other users ('followed') but the followed user does not have to follow the follower. A key difference of social media and the main reason for its popularity is that responses are often almost synchronous and can occur regularly throughout the day as individuals check their social media feeds at work, home, and on their smart phones. These type of social media foster telepresence, the perception of mediated communication as face-to-face communication [7]. As McNab [8] notes, Twitter provides a unique historical opportunity for more accurate health information to be disseminated 'to many more people than ever before', adding that 'one fact sheet or an emergency message about an outbreak can be spread through Twitter faster than any influenza virus' [8]. Lastly, Twitter changes the relationship between health institutions (including individual doctors) and the public in that previously monologic health dictums and warnings can now be interrogated, individually situated, or affirmed through an interaction with the institution or person tweeting that information. In this way, Twitter can foster better health outcomes like, for instance, someone deciding to schedule colonoscopy or mammogram after receiving tweets discussing the successful cases of patients who beat cancer that was discovered at an early stage. Twitter and similar social media also present new opportunities for patient support networks.

In Twitter, the illnesses which tend to have the most active networks are either chronic or life-changing. Twitter networks surrounding cancer are highly active and some users insert the phrase 'cancer survivor' into their user biographies. Survivors of cancer are shaped by their illness experience and, for some, this becomes a part of their Twitter persona. The case of cancer networks on Twitter presents a glimpse not only of how doctors and health institutions are dialogically interacting with individuals, but also how these networks have an international reach and, most of the time, involve strangers, rather than strengthening existing off-line relationships. Though existing patients do follow their doctor's Twitter timeline, most often doctors and health institutions are interacting with 'far-flung' colleagues or members of the public [11]. In the case of cancer, Butcher argues that Twitter is 'transforming the cancer care community' by engaging individuals in one-to-one conversations, connecting with oncology professionals, as well as assisting oncology researchers in finding clinical trial participants [12]. Butcher gives the example of the Vanderbilt-Ingram Cancer Center and how they are planning to use Twitter to recruit participants for an upcoming lung cancer clinical trial [12] by locating clusters of people who are interested in lung cancer as well as lung cancer survivors and using these networks to inform these targeted individuals about the clinical trials they will be running.

Cancer networks on Twitter have a far reach. Indeed, some individual oncologists have large followings as well. For example, Butcher [13] gives the example of Prof. Naoto T. Ueno (@TeamOncology), a doctor at the M.D. Anderson Cancer Center and a cancer survivor, who tweets in English and Japanese and has over 4100 followers. Prof. Ueno's use of Twitter is interesting for several reasons. First, he tweets in the evening in Japanese and during the day in English so that Twitter users in Japan are online when he is tweeting in Japanese. In other words, his Twitter timeline straddles two distinct sociolinguistic spaces within Twitter itself. Second, he makes a point of tweeting about other aspects of his daily life. He believes that by tweeting about non-cancer topics, he draws an audience 'that has nothing to do with cancer', but when he tweets about cancer (which he puts it about 40% of the time), these followers still pay attention to his cancer-related tweets [13]. Ueno also uses Twitter to correct misinformation regarding cancer and, in fact, one of his tweets which criticized a breast cancer screening program in Japan led to the program undergoing a 'rethink' [13].

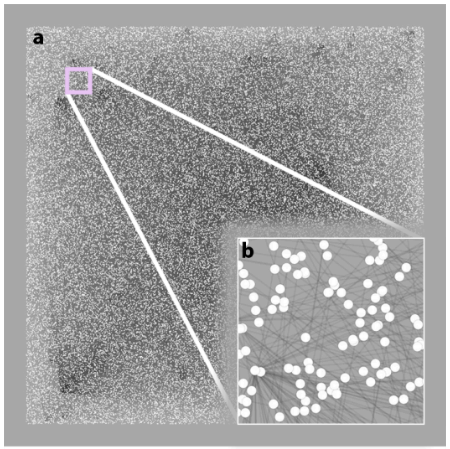

Figure 1. Visualizing Large Networks. (a) This network graph contains more than 70,000 users and 90,000 connections, only 0.16% of the size of the complete distance-2 network around Dr. Anas Younes. (b) Up-close, node distinction improves, but it remains difficult to distinguish which nodes are connected by which edges.

Individuals diagnosed with cancer, carers, and family members have been using Twitter to gather information on particular cancer treatment options and clinical trials, but also ask questions about their specific cases to leading oncologists in the field. This level of agency, from the perception of patients, carers, and family, is a critically important utility of Twitter. Importantly, social media has also facilitated an increase of patient-generated content, which is 'seen as more democratic and patient controlled, enabling users to exchange health-related information that they need and therefore making information more patient/consumer-centered' [1]. However, despite a potential democratizing turn [19], a key issue in these networks is trust. Talking about personal or family diseases to people online and seeking advice necessitates elements of trust in the network. In an open environment such as Twitter, the issue of how much to trust a source of information (a person someone follows in Twitter) or a health-related statement made by such a person is critical. For instance, some recommendations on treatments for statements about symptoms and recovery may be contradictory or made by users without the relevant professional training or experience. Because of the uncensored and collaborative nature of such online networks, requiring any previous validation of the statements and content is unfeasible.

With these significant issues in mind, we describe our ongoing work on studying the dynamics of health-related networks on social media and addressing fundamental, 'building block' questions about these networks including: (i) how these networks influence the flow of health-related information and resources, (ii) how these networks influence health-related attitudes and outcomes, (iii) how we can model and infer trust in users and belief of health-related statements in these networks, and (iv) how trust and beliefs influence information propagation and changes in health-related habits and attitudes.

II. PRELIMINARY STUDIES ON VISUALIZING HEALTH-RELATED TWITTER NETWORKS

In the absence of a large body of accepted practices and methods for studying health networks in social media, we felt that a preliminary study was critical to (i) understand the nature of health networks in social media and what types of information they contain, (ii) develop methods for capturing data, (iii) evaluate whether this data revealed potentially positive health outcomes, and (iv) anticipate potential research challenges. Twitter was chosen for our preliminary study, as it is a not only prominent example of emergent social media [20], but has been meaningful to the development of health-related communities. These investigations so far have been two-fold. The first component has consisted of investigations into the nature of directional communication in Twitter as related to particular topical contexts by the keywords 'chemo', 'mammogram', 'melanoma', and 'lymphoma'. The second area of study has focused on the size, connectivity, and structure of specific social clusters in the network focused around a particular topic or person (e.g. cancer-related communities). We conducted a six-month preliminary investigation in our Social Network Innovation Lab (SNIL). From this, we have preliminary data which includes a filtered dataset of 195,915 tweets containing 'chemo' (88,293 tweets), 'mammogram' (18,443), 'lymphoma' (39,215), and 'melanoma' (49,961). The dataset also includes over 30 million user nodes over which these messages have traversed. This project also took as its goal to capture and visualize the structure of social networks focused around one individual "seed" user, within Twitter. The "seed" user chosen, Dr. Anas Younes, is an oncologist and cancer researcher at the MD Anderson Cancer Research Center as well as a respected member of a variety of cancer-focused networks within a number of different social media platforms including Twitter (See Figure 1).

We also wanted to understand structure and information flow in health networks on Twitter and test the requirements for data storage, but also to develop methods for future research interactions with such networks. To do so, we created a basket of tools to capture data about the network of friends and followers around the seed to a distance of two degrees of separation. It was found that network clusters identified using this method can be quite substantial. The network at a distance of 2 from our chosen seed consisted of approximately 30 million users and over 72 million unique connections between these users. Current estimates put the total number of Twitter users at about 175-200 million [21] so this seed network represents roughly one-sixth of the entire Twitter network. Networks of this size resist visualization both because of the processor intensive problem of laying out over 100 million visual objects; but also because once rendered, the information visualized would be near impossible to understand in a meaningful way without restricting one's field of view within the network. Early manipulations of this large dataset presented poor performance which was later improved through database optimization and formatting. Further, the need for visualization also posed issues due to their high processing power/memory demands: most currently accepted tools for network visualization were not developed with such large networks as a requirement. Initially we used the SNA software package Pajek, which through experiments proved to be largely insufficient to the needs of the project. We then turned to the Java programming language-based Cytoscape which achieved better performance for larger datasets.

In order to analyze structural significance of the cancer-related network surrounding Dr. Younes, we developed a methodology to categorize connections between nodes into one-way "links" and reciprocal "peer" connections. We applied this methodology to the existing network to investigate and visualize more focused and specific subnetworks within the total network and useful visualization was created with degree 1 network of "peers" around the "seed", where we could also visualize the "peer" connections between any two nodes of that subset. This network consists of 176 nodes connected on 2200 arcs and its visualization is illustrated in Figure 2. The results of this pilot revealed not only the value of visualization for understanding social medial network but some issues that need to be addressed. First, we will need to careful design our data structures with performance as a requirement and also optimize operations on this data (e.g, data base indexing). Our preliminary results also showed the important of choosing and optimizing visualization tools as they play a critical role on meaningful visual analysis. Finally it become clear the need for categorizing various classes of Twitter users as well as classes for categorizing connections between any two network nodes.

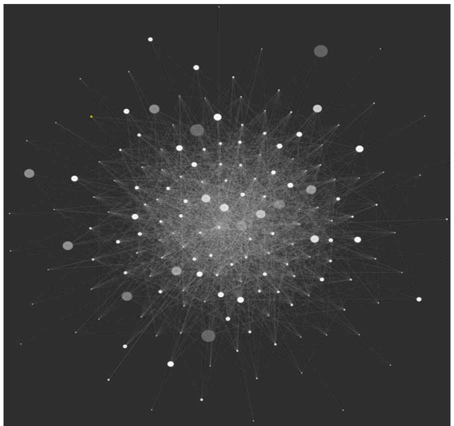

Figure 2. Distance-1 network of "peers" around @DrAnasYounes. Nodes near the center are highly connected with others nodes in this network. Nodes at the edges share only a few peer connections with Dr. Anas Younes.

The development of such taxonomies would allow for visualization of more focused networks around a given user or set of users, as well as providing additional details about the networks themselves and allowing new hypotheses to be drawn in future social network analysis research.

III. INNOVATION VS STATE OF THE ART

Our approach seeks to innovates Social Network Analysis (SNA) to understanding complex health networks in social media which are fluid, resist traditional notions of trust, and often lack explicit bidirectional relationships. SNA is a quantitative social scientific method for measuring social relations through an emphasis of structural relations [22], which posits that the structure of social networks affects 'perceptions, beliefs, and actions through a variety of structural mechanisms that are socially constructed by relations among entities' [22]. The literature on SNA is well established and so are the metrics and modes of visualization [22-40]. However, SNA is still evolving as a method to understand social media networks such as the health networks on Twitter. A recent development to update SNA to understand online networks has been VSNA, the analysis of virtual social networks using methods derived from SNA [23]. VSNA, like SNA, is built upon mapping and observing relationships rather than merely aggregating data on members of their attributes [41].

We extend VSNA from its nascency to a new stage which would help future researchers understand and visualize health-related online social networks. We explore fundamental questions about social networks formed in social media and provides innovative methods for future researchers to conduct applied research in the health sciences using SNA. Our approach is divided into three key areas of innovation: (i) to address fundamental questions about social networks in emergent social media regarding information flow, authority, and cohesion, (ii) to analyze the dynamics of social trust and its influence in social media, and (iii) to provide methods for future researchers to gather data from social media networks for analysis using SNA.

Specifically, our goal is to develop new methods fusing SNA, Natural Language Processing (NLP), and machine learning [48, 49] to analyze confirmatory and negatory mentions in social media. For example, if a 'stop smoking today' tweet circulates across Twitter, how is it being responded to in Twitter as a whole as well as in specific subnetworks (e.g. ones identified as formed around cancer or health-related issues). Additionally, we plan to examine how this tweet is being responded to, what percentage of users respond, what percentage of users respond positively, and what percentage of users respond negatively. Though work has been done on gauging mood from Twitter using sentiment analysis [50-52], no work has been done at such a focused level. This is a gap in the literature and is a critical step to giving future researchers a better understanding of micro-/macro-network flows. Another innovation is using SNA to map out a large Twitter network using influential cancer experts as seeds and then to trace particular tweets and mentions. Further, we will track nodes who have received the message in their Twitter feed, but who have abstained from responding (i.e. tweeting a mention) and analyze if these abstainers tweet messages using keywords/phrases which indicate a confirmatory behavior (e.g. 'finally decided to schedule a mammogram'). This method enables us to not only see how information flows from cancer 'authorities' in Twitter, but also how this potentially shapes behaviors. Ultimately, this method will provide a platform for future researchers to test information flow, influence, and reach. Additionally, because SNA is used to visualize these networks, the authority of individual Twitter users can be analyzed using SNA software and cohesion can be visualized and analyzed by looking not only across meshing between nodes, but also can be discerned by examining whether a specific tweet traverses the network along these tight nodal interconnections.

Our final key area of innovation is to analyze social trust and its influence on health-related social media. Trust is an important issue in these networks given their collaborative and uncensored nature. Volunteering personal medical information, seeking advice, or supporting strangers requires elements of trust which is itself a subjective concept [53-55]. Computing with social trust is a relatively new research area, but has developed initial models and systems capable of defining, modeling and employing social trust in applications. Golbeck [56] studied the problem of utilizing the structure of an online social network (OSN) and the trust relationships within it to infer how much two people that are not directly connected trust one another and to integrate this data in applications. She considers only networks such as LinkedIn where individuals that are directly connected explicitly assign trust to one another in a binary scale and infers trust relationships from individuals that are not direct connected. Her model does not consider social networks such as Twitter because her definition of social network requires that there exists a relationship between two connected users and that they explicitly state their relationship with users they are connected in the network. Richardson et al. [57], introduce a mathematical and a probabilistic model to infer trust and belief in statements made by users in the Web. Their model is general in the sense that they do not specifically consider a social network, but a system where users explicitly assign trust values to other users and statements these users make in this system. They run their experiments in Epinions, a user-oriented product review website where users specifically specify which users and statements they trust and use this model to order the product reviews seen by each person. Guha et al. [58] also used Epinions to study whether distrust can be propagated and inferred like trust by converting ratings to binary values representing trust and distrust. Our approach, although based on Richardson et al. model [57], does not require that users explicitly rate or assign trust to followers or their tweets. This is an important development.

In terms of work done on Twitter, Ye and Wu [59] present a measurement study of 58 million messages collected from 700,000 users on Twitter to analyze propagation patterns of general messages and show how breaking news (for instance Michael Jackson's death) spread through Twitter. However, a major gap in this analysis is that it did not consider social trust. Golder and Yardi [60] conducted a web-based experiment on Twitter in which users were asked to rate their interest in forming ties to other users without any previous information about existing connections between them. Their study showed that transitivity and mutuality are significant predictors of the desire to form new ties. Phelan et al. [61] developed a system that uses real-time Twitter data for ranking and recommending news articles from a collection of RSS feeds. The proposed project sees these emerging trends in social media research as significant, but lacking more developed algorithms for social trust. Additionally, there is a need to not only understand the trust structure of these networks and how information, ideas, and behaviors flow across it, but to also develop tools for meaningful applied research to be conducted from social media-derived health/health outcome data.

IV. VISUALIZING HEALTH NETWORKS IN SOCIAL MEDIA

In this section we describe how we seek to map large and small-scale health-related networks on Twitter. A purpose of this aim is to not only view the internal structure of these networks, but also their interconnections with other cancer-related and health-related networks on Twitter. We also consider here the crawling from nodes in these networks to three degrees so that interconnections between these networks to a degree of three can be observed.

Our preliminary data indicates the presence of discernible and meaningful cancer-related networks on Twitter. Additionally, the structure of Twitter enables the construction of topical groups ('lists') which consist of other Twitter users who have an interest in a specific topic. For example, there are numerous cancer survivor lists which are extremely active and consist of a critical mass of followers (e.g. ones working with the cancer organization Livestrong). These lists would be visualized using the visualization methods outlined in Section III. Selected Twitter users who are oncologists and other cancer-related medical professionals will be selected as seeds and this is feasible given their biographical information (e.g. oncologist at X hospital).

We will generate Cytoscape visualization files of selected 'list' networks identified by keyword, number of followers, and offline institutional affiliations (e.g. the case of the Livestrong cancer organization mentioned above). Specific lists to be visualized will be those relevant to (i) cancer survival networks, (ii) cancer support groups (including breast cancer groups), and (iii) lists based on treatment advice/options. These lists will be visualized as complete networks and arcs/edges will be visualized based on singular or reciprocal relationships. Network diagrams will also be generated from 25 selected seeds and will be visualized using Cytoscape to illustrate first degree connections only. This will provide visualizations of networks surrounding seed users who have been identified as having authority in 'traditional' medical institutions (i.e. doctors and cancer-related medical professionals). Among these seeds, we will select ten nodes that are cancer survivors and visualizations of these networks are additional expected outcomes. Lastly, a large-scale network will also be constructed by crawling three degrees from these 25 seeds. This network is anticipated to contain 100-125 million nodes with an unknown number of links between nodes.

We anticipate some research challenges such as the fact that networks on social media are highly transitory and subject to regular change. Using a seed on one-day versus a week later could yield quite different network visualizations if the seed gains 100 followers during the week. Not only is the size of the network affected, but also its internal structure is subject to change. A solution to this problem is to have regularly scheduled services on our server which will incrementally update the data files for these networks so that, though not real-time, a reasonably accurate visualization can be analyzed.

A. *Authority and Cohesion*

We also seek to understand how resources and ideas flow through health networks in social media. By looking at particular health outcomes and behaviors of interest (e.g. getting a mammogram or undergoing chemotherapy), we hope to identify what ideas and resources are considered important. Specifically, we will investigate how the authority and cohesion of these networks influences how importance is formulated in the first place.

Sub-networks in social media have both a cohesive network structure with identifiable hubs as well as having identifiable nodes of authority. We will track particular keywords like mammogram real-time by leveraging the Twitter API. Our goal is to generate Cytoscape datafiles which map health networks and identify authorities and hubs are a key expected outcome of this Aim. These files will correlate authorities and hubs with the propagation of tweets by particular health-related keywords and sets of files by dates and time over a longitudinal period are an expected outcome. Further, we will identify what health-related keywords are considered to be important in Twitter and to reveal what specific resources are considered to be essential to propagating these ideas. Specifically, it is expected that status as a hub or authority will affect what ideas are considered important to both the larger network of Twitter as well as health-related sub-networks. Additionally, it is also expected that other resources such as mentions and retweets will be critical to identifying particular health information as important.

One issue is that nodes identified as authorities and hubs will be potentially changing on a frequent basis depending on what keywords are traversing these networks as well as how these networks are structured. A second anticipated problem is that malicious users could potentially be identified as hubs and authorities due to extremely high levels of mentions, retweets, and links to other nodes. A solution to this is to implement Twitter-based spam detection methods [62-64] which have a proven track record in the literature of eliminating nodes identified as spammers.

V. DYNAMICS OF HEALTH-RELATED BEHAVIORS

We will explore whether and how these networks have the ability to influence health-related behaviors, attitudes, and information. Specifically, how are healthy behaviors (e.g. regular Pap smears and mammograms for women) facilitated and unhealthy ones (encouraging smoking) impeded. Keywords in specific tweets will be used to trace the life of particular tweets and will be used to see if they result in a confirmation of these behaviors. For example, does a tweet of 'Don't forget to schedule a mammogram' lead to responses back such as 'just scheduled my mammogram'

Our strategy is to employ sentiment analysis to identify the mood of tweets and extending it to identifying sentiment in response to tweets (e.g. 'just scheduled my mammogram').

Further, we will incorporate mechanisms to track specific tweets and to look for responses back. Additionally, by using machine learning to evolve our understandings of what terms in tweets and in what contexts correlate with positive, negative, neutral responses, the scope of this aim, though challenging, is feasible.

One issue which is anticipated is that related phrases/keywords suggested by the crawler could be either malicious or simply irrelevant. The solution which we would implement in this likely scenario is to combine random checking by humans using student research assistants and human checkers with the Amazon Mechanical Turk (AMT), a service which has people around the world ('turkers') who perform simple tasks through the AMT API. Wee would send keywords and phrases to our student research assistants and turkers for human error checking. A second issue is that the accuracy of sentiment analysis-based algorithms may be low. As with the first issue, we would incorporate the same solution. The idea is that this human input would help the software learn how to better code confirmatory, negatory, neutral, and abstaining responses.

VI. MODELING AND INFERRING TRUST

Our goal here is to understand the dynamics of social trust in a health-related Twitter network. We will be modeling, inferring and storing computed trust values for users and beliefs, and also analyzing how social trust influence the information flow of messages. Our first challenge is how to estimate initial trust values t_{ij} for all users *i* following user *j*. Given a Twitter network we must first estimate initial trust values t_{ij} for all users *i* in the system following user *j*, where t_{ij} means how much user *i* trusts user *j* based on the number of tweets from user *j* positively replied or re-tweeted by user *i*. We will develop a machine learning-based classifier [65, 66] to estimate (classify) a trust value in the range [0,1] for all users *i* and *j* where *i* follows *j*. We will first select (semi-automatically) a set of features in the tweets that indicate trust and train our machine learning classifier with this dataset. Machine learning techniques have been successfully used, for example, in network security to classify a stream of network packets as malicious or benign [67].

A second challenge is how to infer how much a user trusts another user in the network even if she is not directly connected to them (follower relationship). We adapt the model proposed by [57] as follows. If user *i* follows user *j*'s tweets, this means that *i* has some trust t_{ij} in what *j* has to say. Also, if user *j* follows user *k's* tweets, then *j* has some trust t_{jk} in user *k* and, then user *i* should have some trust t_{ik} in user *k*, which is a function of t_{ij} and $t_{jk.}$. We assume a network of *N* users. The result of our machine learning classification will be a NxN matrix T, called the personal trust values matrix, where t_{ij} contains the trust user *i* has on user *j* he/she follows. In this matrix t_{ij} is not necessarily equal to t_{ji}, and t_i represents the row vector of user *i* estimated trust in other users. Thus, in this matrix, t_{ik} represents how much user *i* trusts user *k* and t_{kj} represents how much user *k* trusts user *j* and ($t_{ik} . t_{kj}$) represents the amount user *i* trusts user *j* via *k*. The amount that user *i* trusts user *j* via any single other node is thus given by $\sum_k$ ($t_{ik} . t_{kj}$). Using the idea of web of trust from Richardson et al. [57] we can compute for any user *i* her trust on any user *j* in the network. The trust between any two users is given by a trust matrix T (merged trusts matrix) to compute the merged trusts on the same Twitter graph where there is a path between user *i* and *j* if *i* follows *j*. We infer trust values between any user *i* and *j*, independently if *i* follows *j* using an aggregation function which concatenates trusts along every path between them by applying the following algorithm [57]:

```
$T^{(0)} = \mathrm{T}$
Repeat {
    $T^{(n)} = \mathrm{T} \,.\, T^{(n-1)}$
    }
Until ($T^{(n)} = T^{(n-1)}$)
```

Here, $T^{(i)}$ is the value of *T* in iteration *i*. Also, we borrow the matrix multiplication definition from [57]: C = AB is such that $C_{ij} = \sum_k (A_{ik} . B_{kj})$ The result is a data structure with per-user inferred trust values dataset for analysis.

Our second research challenge is how to estimate initial personal beliefs for each tweet based on replies and retweets that user *i* reads from user *j* that she follows. We also make use of machine learning for this initial estimation and adapt a model for beliefs in general Web documents using a path algebra interpretation as proposed by Richardson et al [57]. Given a network with *N* users and *M* health-related tweets, we must first estimate personal beliefs values in the range [0,1], for each user and each statement this user has access to. We will first select (semi-automatically) a set of features in the tweets that indicate belief and train our machine learning system with this dataset to estimate a per-user belief value on a tweet. Let b_i represent user *i*'s personal belief on a particular tweet. If user *i* has no belief on the tweet (or no access to it), b_i is set to 0. The collection of personal beliefs in a particular statement is the column vector **b**. Then we infer how much a user believes in any tweet in the network. The trust values computed above will allow us to compute for any user *i*, her belief in any tweet using a structure called merged beliefs ($\boldsymbol{\beta}$) The merged beliefs structure can be calculated as follows:

```
$\beta^{(0)} = \mathrm{b}$
Repeat {
    $\beta^{(n)} = \mathrm{T} \,.\, \beta^{(n-1)}$ or,  $\beta_i^{(n)} = \sum_k t_{ik} \,.\, \beta_k^{(n-1)}$
}
Until ($\beta^{(n)} = \beta^{(n-1)}$)
```

Here, $\beta^{(i)}$ represents the value of β in iteration *i.* The research challenge here is that these online social network structures are fluid and, as a result, their structure may change as new users become part of the network and others leave. Moreover we may observe changes in social influence over time. We use a snapshot of the network to infer initial trust values for our inferred per user trust structure and these values might become outdated with time. A solution to this problem is recalculate the initial trust values and re-infer the web of trusts periodically. To accomplish this we need to conduct experiments to estimate an optimal frequency for the network checkpoints. Further, our analysis methods should be as automated as possible as we may need to re-analyze the new data after each checkpoint. Malicious users might also introduce errors to our data and we plan to counter this using anti-spam approaches [62-64].

We also plan to track tweet information flow using Ye and Wu's [59] algorithm and analyze how propagation relates to the belief in a tweet or trust in a user that replied or retweeted it. In Twitter a reply message has the fields "in_reply_to_status_id" and "in_reply_to_user_id", which allows us to track a reply with the message being replied. First all replies are sorted according to their timestamps with the earliest message as the first one. Then, we walk through the sorted message list in a top-down fashion. Supposing the original message is *j*; for each message *i* replying *j,* if there is a tree data structure which has *j* as node, create a node *i,* with *j* as parent node. Otherwise, create a new tree data structure with node *j* as root and then create a new node *i*, with node *j* as parent.

VII. Conclusions

This position paper describes how, for the first time, SNA can be combined with natural language processing and machine-learning methods to be able to determine the flow of health information, trust, resources, and ideas on social media and their specific impact on health outcomes. This allows us to better understand why users trust particular health messages (for example based on a keyword or originating user) and what impact these trust relationships have on bettering health outcomes. We believe our approach will greatly improve our understanding of health networks in social media. Until work is done to not only understand the structure and ways in which health information flows on these emergent networks, but also provide easy-to-use basic research tools, it will be impossible to gauge what impact these networks are having on health outcomes.

References

[1] Chou, W.-y.S., et al., *Social media use in the United States: implications for health communication.* Journal of medical Internet research, 2009. **11**(4).

[2] Orsini, M., *Social Media: How Home Health Care Agencies Can Join the Chorus of Empowered Voices.* Home Health Care Management & Practice, 2010. **22**(3): p. 213-217.

[3] Fox, S. and S. Jones, *The social life of health information*, in *Pew Internet & American Life Project*. 2009, Pew Research Center: Washington DC.

[4] Madden, M., *Older Adults and Social Media*, in *Pew Internet & American Life Project*. 2010, Pew Research Center: Washington DC.

[5] Krowchuk, H.V., *Should Social Media be Used to Communicate With Patients?* MCN The American Journal of Maternal Child Nursing, 2010.

[6] Crumb, M.J., *Twitter Opens a Door to Iowa Operating Room*, in *The Associated Press*. 2009.

[7] Licoppe, C., *'Connected' presence: the emergence of a new repertoire for managing social relationships in a changing communication technoscape.* Environment and Planning D: Society and Space, 2004. **22**(1): p. 135-156.

[8] McNab, C., *What social media offers to health professionals and citizens.* Bulletin of the World Health Organization, 2009. **87**: p. 566-566.

[9] Hawn, C., *Take Two Aspirin And Tweet Me In The Morning: How Twitter, Facebook, And Other Social Media Are Reshaping Health Care.* Health Affairs, 2009. **28**(2): p. 361-368.

[10] Vance, K., W. Howe, and R.P. Dellavalle, *Social Internet Sites as a Source of Public Health Information.* Dermatologic Clinics, 2009. **27**(2): p. 133-136.

[11] Victorian, B., *Nephrologists Using Social Media Connect with Far-Flung Colleagues, Health Care Consumers.* Nephrology Times, 2010.

[12] Butcher, L., *How Twitter Is Transforming the Cancer Care Community.* Oncology Times, 2009. **31**(21): p. 36-39.

[13] Butcher, L., *PROFILES IN ONCOLOGY SOCIAL MEDIA: Naoto T. Ueno, MD, PhD - @teamoncology.* Oncology Times, 2010. **32**(13): p. 38-39.

[14] Butcher, L., *Oncologists Using Twitter to Advance Cancer Knowledge.* Oncology Times:, 2010. **32**(1): p. 8-10.

[15] Manfredi, C., et al., *Are Racial Differences in Patient-Physician Cancer Communication and Information Explained by Background, Predisposing, and Enabling Factors?* Journal of Health Communication, 2010. **15**(3): p. 272-292.

[16] Celik, H., et al., *Maintaining gender sensitivity in the family practice: facilitators and barriers.* Journal of Evaluation in Clinical Practice, 2009. **15**(6): p. 1220-1225.

[17] Singh, H., et al., *Older Patients' Enthusiasm to Use Electronic Mail to Communicate With Their Physicians: Cross-Sectional Survey.* Journal of Medical Internet Research, 2009. **11**(2): p. 13.

[18] Lenhart, A., et al., *Social media & mobile internet use among teens an young adults*, in *Pew Internet & American Life Project*. 2010, Pew Research Center: Washington DC.

[19] Turner, G., *Ordinary people and the media : the demotic turn*. 2010, London: SAGE. vi, 189 p.

[20] Naaman, M., J. Boase, and C.-H. Lai, *Is it really about me?: message content in social awareness streams*, in *ACM conference on Computer supported cooperative work*. 2010, ACM: Savannah, Georgia, USA.

[21] Raby, M. (2010) *Twitter on pace to reach...200 million users by 2011*. TG Daily.

[22] Knoke, D., S. Yang, and D.N.a. Knoke, *Social network analysis*. 2nd ed. ed. 2008, Los Angeles ; London: Sage. viii ,132 p.

[23] Abraham, A., A.E. Hassanien, and V. Snásel, eds. *Computational social network analysis : trends, tools and research advances*. 2010, Springer: New York ; London. 1 v.

[24] Carrington, P.J., J. Scott, and S. Wasserman, *Models and methods in social network analysis*. 2005, Cambridge: Cambridge University Press.

[25] Dekker, A.H., *C4ISR architectures, social network analysis and the FINC methodology : an experiment in military organisational structure*. 2002, Edinburgh, S. Aust.: DSTO Electronics and Surveillance Research Laboratory. 29 p.

[26] Freeman, L.C., *The development of social network analysis : a study in the sociology of science*. 2004, Vancouver, BC: Empirical Press. xii, 205 p.

[27] Freeman, L.C., *Social network analysis*. 2008, Los Angeles ; London: SAGE. 4 v.

[28] Freeman, L.C., A.K. Romney, and D.R. White, *Research methods in social network analysis : Conference on methods of research in social networks : Papers*. 1992, Fairfax, Va.: George Mason University Press ; London : Eurospan, 1989.

[29] Kendrick, A., *Applied social network analysis*. 1992, Scotland: Social Services Research Group, Scottish Branch. ii,31p.

[30] Lörincz, A., G.N. Gilbert, and R. Goolsby, *Social network analysis : measuring tools, structures and dynamics*. 2006: Elsevier. 1 v.

[31] Nooy, W.d., A. Mrvar, and V. Batagelj, *Exploratory social network analysis with Pajek*. 2005, Cambridge: Cambridge University Press.

[32] Papaioannou, T. *Using social network analysis to examine organizational use of electronic mail*. [1 v. ; 31 cm.] 2004.

[33] Pastor, J.-C., J.R. Meindl, and M. Mayo, *A social network analysis of attributions of charisma*. 1996, Buffalo: School of Management, State University of New York at Buffalo. 48, [9] leaves.

[34] Rodkin, P.C. and L.D. Hanish, *Social network analysis and children's peer relationships*. 2007, San Francisco: Jossey Bass. 112 p.

[35] Scott, J., *Social network analysis : a handbook*. 1991, London: Sage.

[36] Wasserman, S. and K. Faust, *Social network analysis : methods and applications*. 1994, Cambridge: Cambridge University Press. xxxi,825p.

[37] Huisman, M. and T.A.B. Snijders, *Statistical analysis of longitudinal network data with changing composition.*Sociological Methods and Research, 2003. **32**: p. 253-287.

[38] Wellman, B., *For a social network analysis of computer networks: a sociological perspective on collaborative work and virtual community*, in *ACM SIGCPR/SIGMIS conference on Computer personnel research*. 1996,

[39] Wellman, B., et al., *Computer Networks as Social Networks: Collaborative Work, Telework, and Virtual Community.* Annual Review of Sociology, 1996. **22**(1): p. 213-238.

[40] van Duijn, M.A.J. and J.K. Vermunt, *What Is Special about Social Network Analysis?* Methodology, 2006. **2**(1): p. 2-6.

[41] D'Andrea, A., F. Ferri, and P. Grifoni, *An Overview of Methods for Virtual Social Networks Analysis*, in *Computational social network analysis : trends, tools and research advances*, A. Abraham, A.E. Hassanien, and V. Snásel, Editors. 2010, Springer: New York; London. p. 3-26.

[42] Kwak, H., et al., *What is Twitter, a social network or a news media?*, *e 19th international conference on World wide web*. 2010, ACM: Raleigh, North Carolina, USA.

[43] Java, A., et al., *Why we twitter: understanding microblogging usage and communities*, in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. 2007, ACM: San Jose, California.

[44] Cha, M., et al. *Measuring User Influence in Twitter: The Million Follower Fallacy*. in *Fourth International AAAI Conference on Weblogs and Social Media*. 2010. George Washington University.

[45] Bakshy, E., et al., *Everyone's an influencer: quantifying influence on twitter*, in *Proceedings of the fourth ACM international conference on Web search and data mining*. 2011, ACM: Hong Kong, China.

[46] Marwick, A.E. and d. boyd, *I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience.* New Media & Society, 2010.

[47] Gruzd, A., B. Wellman, and Y. Takhteyev, *Imagining Twitter as an Imagined Community.* American Behavioral Scientist, Forthcoming(Special issue on Imagined Communities).

[48] Cortes, C. and V. Vapnik, *Support-vector networks.* Machine Learning, 1995. **20**(3): p. 273-297.

[49] Witten, I. and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. 2005: Morgan Kaufmann.

[50] Jansen, B.J., et al., *Twitter power: Tweets as electronic word of mouth.* Journal of the American Society for Information Science and Technology, 2009. **60**(11): p. 2169-2188.

[51] Diakopoulos, N.A. and D.A. Shamma, *Characterizing debate performance via aggregated twitter sentiment*, in *28th international conference on Human factors in computing systems*. 2010, ACM: Atlanta, Georgia, USA.

[52] Culotta, A., *Towards detecting influenza epidemics by analyzing Twitter messages*, in *Proceedings of the First Workshop on Social Media Analytics*. 2010, ACM: Washington D.C., District of Columbia.

[53] Luhmann, N., *Trust and Power*. 1979: Wiley.

[54] D. Lewis, A.W., *Trust as a social reality.* Social Forces, 1985. **63**(4): p. 967-985.

[55] Gambetta, D., *Trust: Making and Breaking Cooperative Relations*. 1990: Blackwell.

[56] Golbeck, J., *Computing and Applying Trust in Web-based Social Networks*, in *Computer Science*. 2005, University of Maryland - College Park.

[57] M. Richardson, R.A., P. Domingos, *Trust Management for the Semantic Web*, in *International Semantic Web*. 2003.

[58] R. Guha, R.K., P. Raghavan, A. Tomkins, *Propagation of Trust and Distrust*, in *13th Annual International World Wide Web Conference*. 2004: New York, NY.

[59] Ye, S. and F. Wu, *Measuring Message Propagation and Social Influence on Twitter.com* in *International Conference on Social Informatics* 2010.

[60] Golder, S. and S. Yardi, *Structural Predictors of Tie Formation in Twitter: Transitivity and Mutuality*, in *IEEE International Conference in Social Computing*. 2010.

[61] Phelan, O., K. McCarthy, and B. Smyth, *Using Twitter to Recommend Real-Time Topical News*, in *ACM Conference on Recommender Systems*. 2009.

[62] Petrović, S., M. Osborne, and V. Lavrenko, *Streaming first story detection with application to Twitter*, in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics: Los Angeles, California.

[63] Grier, C., et al., *@spam: the underground on 140 characters or less*, in *Proceedings of the 17th ACM conference on Computer and communications security*. 2010, ACM: Chicago, Illinois, USA.

[64] Prasad, S.K., et al., *Can You Judge a Man by His Friends? – Enhancing Spammer Detection on the Twitter Microblogging Platform Using Friends and Followers*, in *Information Systems, Technology and Management*. 2010, Springer Berlin Heidelberg. p. 210-220.

[65] I. Witten, E.F., *Data Mining: Practical Machine Learning Tools and Techniques*. 2005: Morgan Kaufmann.

[66] Vapnik, C.C.a.V., *Support-vector networks.* Machine Learning, 1995. **20**(3): p. 273-297.

[67] L. Bilge, E.K., C. Kruegel, M. Balduzzi, *Finding Malicious Domains Using Passive DNS Analysis* in *18th Annual Network and Distributed System Security Symposium (NDSS)*. 2011.