Twitter and e-Health

A case study of visualizing cancer networks on Twitter

Dhiraj Murthy, Alexander Gross, Scott Longwell

Social Network Innovation Lab, Department of Sociology
Bowdoin College
Brunswick, United States
dmurthy@bowdoin.edu

Abstract— **This paper seeks to understand health-related social networks on social media websites. The paper explores fundamental questions about social networks formed in the prominent social media website Twitter and demonstrates innovative new methods to conduct applied research in the health sciences using social media networks. The paper aims to address fundamental questions about health-related social networks in emergent social media regarding information flow and network structure. The paper uses a dataset built from Twitter data using a well-known American oncologist as a 'seed' and crawled the Twitter network three degrees out to form a total network of over 30 million nodes**

Keywords- Twitter; e-health; social media; cancer; visualization

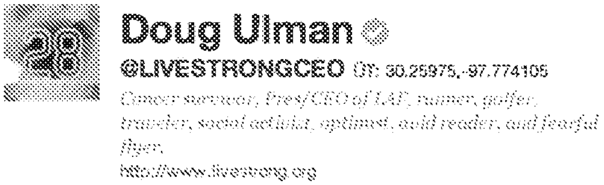
I. INTRODUCTION

Layfield [1] notes that the 'life sciences as an industry can be a slow moving machine, typically behind the adoption curve of anything new in terms of communications'. Of course, reasons of privacy, patient well-being, and differences in information literacy are partially attributable. In the case of Twitter (and social media more broadly), 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. According to data from the 2007 Health Information National Trend Survey (HINTS), 23% of respondents reported using a social networking site [2]. One reason for this is that these individuals form support networks with strangers who have the same chronic illness. This is not a phenomenon restricted to twitter, but rather as Orsini [3] 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. [2] 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. [4]. 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' [4]. Additionally, 15% of e-patients 'have posted comments, queries, or information about health or medical matters' [4]. Though the last percentage may seem small, it is statistically significant and this sharing of personal health information on social networking sites represents a starting, rather than ending point. From 2009 to 2010, social networking use among internet users aged 50-64 grew from 25% to 47% [5]. To put things in context, take the example of a hysterectomy and uterine prolapse surgery of a 70-year-old woman in Iowa which was tweeted real time through 300 tweets posted by a hospital official from a computer immediately outside of the operating room's sterile area [cited in 6]. The woman gave consent for the surgery to be tweeted so that her family could track the operation from the waiting room (and one family member tracked the procedure from her workplace) [7].

As the Iowa case highlights, a key difference of social media 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. As Licoppe [8] has shown, repeated mediated interactions foster telepresence, the perception of mediated communication as face-to-face communication, is felt. As McNab [9] puts it: 'Instant and borderless, it elevates electronic communication to near face-to-face'. She notes that 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' [9]. 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. Similarly, Twitter and social media like it present new opportunities for patient support networks. Hawn [10] describes the case of Rachel Baumgartel, 33, a diabetic who lives in Boulder, Colorado and sends tweets almost daily on 'what she had for breakfast, what her hemoglobin alC level is, or how much exercise she got on the elliptical equipment at the gym'. As Hawn notes, Baumgartel often receives reply tweets from followers, which encourage her to stick to her 'arduous health regimen'. Hawn finds that those who are chronically ill are successfully using social media, including Twitter, in this way.

The illnesses which tend to have the most active Twitter networks are either chronic or life-changing. Twitter social networks surrounding cancer are highly active and some

Twitter users insert the phrase 'cancer survivor' into their Twitter biographies (See Fig. 1).



Doug Ulman

@LIVESTRONGCEO ÜT: 30.25975,-97.774105

Cancer survivor, Pres/CEO of LAF, runner, golfer, traveler, social activist, optimist, avid reader, and fearful flyer.

http://www.livestrong.org

Figure 1. **Twitter profile** Illustrates biographical information including the phrase 'cancer survivor'

Survivors of cancer are shaped by their illness experience and 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 who would not necessarily the cross paths, 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]. The Vanderbilt-Ingram Cancer Center will use Twitter to locate clusters of people who are interested in lung cancer as well as lung cancer survivors and use these networks to inform these targeted individuals about the clinical trials they will be running.

II. Research Background

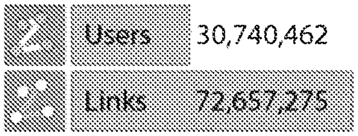
The study of health networks in social media is an emergent field. In the absence of a large body of accepted practices and methods, our preliminary study was critical to understanding the nature of health networks in social media, what types of information they contain, to develop reasonable expectations and methods as to what a researcher will be able to capture, and to anticipate potential roadblocks that may be encountered along the way. Twitter was chosen for our pilot study, as it is a prominent example of emergent social media communities[13]. 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 keywords including 'chemo', 'cancer survivor', and 'lymphoma'.* The second area of study has focused on the size, connectivity, and structure of specific social clusters within Twitter. These are clusters of users in the network focused around a particular topic or person *(e.g. cancer-related communities).*

III. Approach

Preliminary investigations into the structure of specialized cancer-related network clusters within Twitter began with a six-month pilot project in our lab (09/2010-03/2011). This project took as it goal to capture and visualize the structure of social networks focused around one individual "seed" user within Twitter. The "seed" user chosen is both an eminent oncologist and cancer researcher at a premier American Cancer research Center as well as a known and respected member of a variety of cancer-focused networks within a number of different social media platforms including Twitter. The investigations undertaken as part of the pilot project were developed with the specific intention of understanding information flow in health networks on Twitter as well as the structure of these networks. Part of the objective of this preliminary study was to identify what data was to be captured and to test the capabilities and requirements for storage of such data, but also to develop reasonable expectations and methods for future research interactions with such networks. Generally, the goal was to engage in a thorough testing of concepts so as to be able to anticipate potential roadblocks and issues that might be encountered in execution of the project detailed in this proposal.

IV. Process

We created a series of scripts 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 these networks 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 (See Fig. 2).



Users 30,740,462

Links 72,657,275

Figure 2. The Seed's network entities. The number of nodes and connections in the discovered network

Current estimates put the total number of Twitter users at about 175-200 million[14] 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 data proved difficult as the large size of the dataset could cause operations on the network to take long periods of time. All operations on the dataset had to be optimized to function as efficiently as possible. In addition, it was discovered that database optimization and formatting would be necessary to meaningfully use the data. Early visualization also posed various problems because most currently accepted tools for network visualization are not developed with such large networks in mind. The pilot initially used the Windows-based software package Pajek as a possible visualization tool for the captured data. Experiments with Pajek showed that it was largely insufficient to the needs of the project. We then turned to the java-based Cytoscape and achieved better and more consistent results for larger amounts of data.

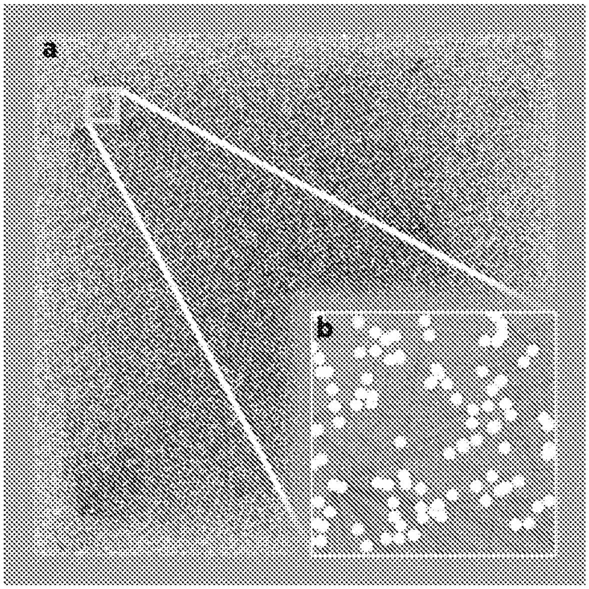


Figure 3. **Visualizing Large Networks** (a) This network graph contains more than 70,000 users and 90,000 connections, only 0.16% of the size of the complete distance-2 network around the Seed. (b) Up-close, node distinction improves, the it remains nearly impossible to distinguish which nodes are connected by which edges.

Though full network visualization is still too computationally expensive for the software, we have been able to test partial subnets with our current hardware resources. In order to begin to actually see pieces of the structure of the cancer-related network surrounding the Seed, we developed a methodology to categorize connections between nodes into one-way "links" and reciprocal "peer" connections (See Fig. 4).

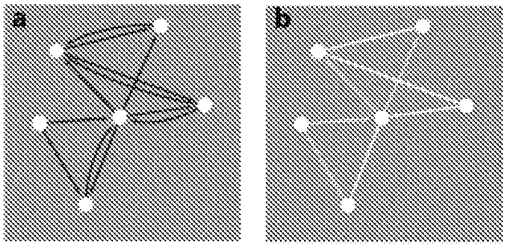


Figure 4. **Methodolgy for "peer" connections.** (a) In unfiltered networks all links are equally weighted and directional. (b) After applying a filter, reciprocal "peer" connections appear in green and directional "link" connections appear in purple

Using this methodology we were able to apply it to the existing network to investigate and visualize more focused and specific networks within the total network. A useful visualization that was created was a view of the degree 1 network of "peers" around the "seed", and then to additionally visualize the "peer" connections between any two nodes of that subset. This network consists of a much more reasonable 176 nodes connected on 2200 arcs. Seen here visualized in Cytoscape (see Fig. 5) using a force-directed layout, that nature of the layout causes the most highly connected nodes to exist near the center and nodes with close proximity to the seeds indicates that nodes are tightly meshed with the seed. In other words, they share a large number of common relationships with other the other nodes. Nodes that are only loosely related to the "seed" and shared few common relationships drift to the edges of the network. Investigation of the network shows this being the case with the Seed being tightly connected with nodes like @AmericanCancer (The American Cancer Society), @Cancerwise (The MD Anderson Cancer Center's official account), and @kevinmd (a popular health resource and consultant for USA Today) but less tightly connected with nodes like @ChuckGrassley (a United States Sentor), @NPRHealth (a generic health-related news feed), and @chirrps (A twitter-based search engine). This shows the value both of visualization as a tool for understanding networks but also the importance of focusing attention in large networks based utilizing consistent methodologies that allows for enhanced understanding.

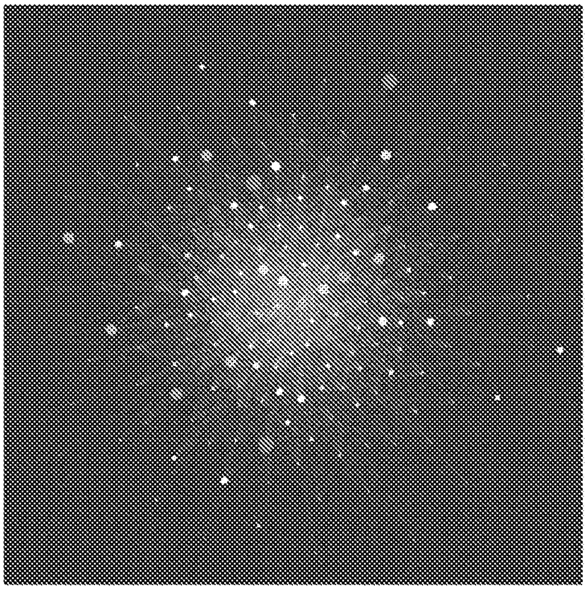


Figure 5. **Distance-2 network of "peers" around the Seed.** Network is layed out around the Seed using a force-directed layout. As a result, nodes near the center are highly connected with others nodes in the doctor's peer network. Nodes at the edges share only a few common peer connections with the doctor

V. Results

The results of this study not only reveal important concerns and issues that will be fundamental to understanding clustered health-related networks within social media. They also begin to validate the assumptions of the hypothesis that data analysis and visualization methods can begin to shed light on how health-related social networks in social media function. With up-close visualizations, we were able to see a cohesive, tightly intermeshed cancer network. The visualizations were also successful in highlighting issues, which will be key to the success of work done by future researchers which seek to visualize health networks on social media including, but not limited to Twitter. These issues are detailed below:

A. Space and Access Optimization

The first concern is structures for space and access to data. Though the physical memory space required for storing such data seems rather reasonable given the large number of entities they will store (less than 5 gigabytes for over 72 million arcs and 30 million nodes), it became clear that careful optimization will be required to allow fast access and operations on this data. Smart indexing on specific database fields has been shown to provide exponential speed increases for some operations. Structures developed for manipulating this data will need to be improved with efficiency and operational speed in mind.

B. Scale and Feasibility Concerns for Technology and Understanding

After initial experimentation, it seemed as if visual analysis might not always be a feasible approach for understanding networks above a given size. Both hardware and software limitations play some role in this as well as the limiting factors of human perception. The pilot project helped identify software and platforms which might be most appropriate to visualize large networks as well as providing a chance to test how the extent to which the capabilities of the software can be augmented with raw computing power. This knowledge will be helpful in choosing future research tools as well as for when the project begins to develop its own real time tools for network visualization.

C. Methodologies and Taxonomies for Enhanced Data Visualization

A key area for further investigation, brought to the fore though the difficulties in visualizing large networks and proved solvable through narrowing focus, was the need and potential the development of a set of methodologies for categorizing various classes of Twitter users as well as classes for categorizing connections between any two network nodes. The development of such taxonomies would allow for taking a look at more focused networks around a given user or set of users. As well as provide additional details about the networks themselves. Allowing new and different types of conclusions and hypotheses to be drawn in future social network analysis research. By using visualization tools to prototype and view the implementation of various connection taxonomies, we can ask whether these visualizations enhance or detract from the understanding of large social network clusters.

VI. CONCLUSION

The results of this study demonstrate that health-related networks on social media websites such as Twitter can be meaningfully visualized using current software platforms such as Cytoscape. The paper also demonstrates that large-scale health networks can be visualized in this way. Critically, the paper introduces the innovative methods we have developed to conduct applied e-health research in Twitter. These methods include data collection, weighting directional and bidirectional peer connections, and using seeds to understand complex health networks on Twitter. Another important purpose of this paper has been to highlight limitations which future e-health visualization research needs to address. The discovery of these limitations is a key conclusion of our study.

ACKNOWLEDGMENT

The authors acknowledge DJ Merrill and Andrew Currier for their invaluable help in scheduling and running the data collection software on the Bowdoin College High-Performance Computing Linux Grid.

REFERENCES

[1] C. Layfield, Modern Communications. Applied Clinical Trials, 2010. 19(5): p. 62-62.

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

[3] M. Orsini, 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.

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

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

[6] H.V. Krowchuk, Should Social Media be Used to Communicate With Patients? MCN The American Journal of Maternal/Child Nursing, 2010. 35(1): p. 6 -7.

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

[8] C. Licoppe, '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.

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

[10] C. Hawn, 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.

[11] B. Victorian, Nephrologists Using Social Media Connect with Far-Flung Colleagues, Health Care Consumers. Nephrology Times, 2010. 3(1): p. 1, 16-18.

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

[13] M. Naaman, J. Boase, and C.H. Lai, "Is it really about me? Message content in social awareness streams", in Proceedings of the 2010 ACM conference on Computer supported cooperative work. 2010, ACM: Savannah, Georgia, USA.

[14] M. Raby, Twitter on pace to reach...200 million users by 2011. TG Daily, 2010.