Inferring Community Structure in Healthcare Forums

An Empirical Study

T. Chomutare1; E. Årsand1,3; L. Fernandez-Luque3; J. Lauritzen3; G. Hartvigsen1
1University hospital of North Norway, Norwegian Center for Integrated Care and Telemedicine, Tromsø, Norway;
2Northern Research Institute, Tromsø, Norway;
3University of Tromsø, Department of Computer Science, Tromsø, Norway

Keywords
Diabetes, social networks, homophily, assortativity, community detection

Summary

Background: Detecting community structures in complex networks is a problem interesting to several domains. In healthcare, discovering communities may enhance the quality of web offerings for people with chronic diseases. Understanding the social dynamics and community attachments is key to predicting and influencing interaction and information flow to the right patients.

Objectives: The goal of the study is to empirically assess the extent to which we can infer meaningful community structures from implicit networks of peer interaction in online healthcare forums.

Methods: We used datasets from five online diabetes forums to design networks based on peer-interactions. A quality function based on user interaction similarity was used to assess the quality of the discovered communities to complement existing homophily measures.

Results: Results show that we can infer meaningful communities by observing forum interactions. Closely similar users tended to co-appear in the top communities, suggesting the discovered communities are intuitive. The number of years since diagnosis was a significant factor for cohesiveness in some diabetes communities.

Conclusion: Network analysis is a tool that can be useful in studying implicit networks that form in healthcare forums. Current analysis informs further work on predicting and influencing interaction, information flow and user interests that could be useful for personalizing medical social media.

1. Introduction

In this work, we examine whether interactions in healthcare forums can be represented as networks. We examine diabetes networks because it is a disease case that involves several stages of illnesses, and is a heterogenic disease with many sub-factors. Diabetes is also a highly relevant disease to focus on from a patient healthcare and economic perspective. In 2003, 194 million people globally were estimated to have a form of diabetes, which is predicted to increase to 333 million in 2025, which constitute an increase of 72% [1]. The problem is seen not only in Europe and America, but on a global scale. WHO estimates that, 5–10% of the national healthcare budget in western countries is used on diabetes, which will increase with the increasing number of diabetes patients globally.

In some instances, patients must rely on peers for emotional, empathetic and practical support. Internet forums are one of the most popular social media for self-help. A major distinguishing feature between interaction in these forums (message boards) and interaction in other social media is that the forums normally do not have explicit relationships, and unlike most social networking websites, relationships in forums are implied. These relationships are encoded in large datasets of forum threads-and-comments dynamics, and in this work we use network analysis to decipher these relationships. Community detection [2] are a group of network analysis [3] methods that hold a potential to reveal characteristics that help us identify important peers [4], predict community attachments and influence information flows and temporal patterns [5, 6].

1.1 Related Literature

The idea of discovering communities from forum interactions is not new; researchers have long been fascinated by the prospect [7, 8]. Different methods have been discussed extensively in the literature, but work by L’Huillier et al. [8] enhanced our understanding of how forum discussions can be analyzed and connected using network analysis and text mining. While the work focused on terrorism and a few hundred users, it nonetheless sheds some light. In healthcare, the review by Dunn and Westbrook [9] provides a complete synthesis to date of the relevant network analysis concepts for small scale healthcare networks. Network analysis has been used extensively in bioinformatics [10] and

Correspondence to:
Taridzo Chomutare
Norwegian Center for Integrated Care and Telemedicine
University Hospital of North Norway
9038 Tromsø
Norway
E-mail: taridzo.chomutare@telemed.no

doi: 10.3414/ME12-02-0003
received: March 1, 2012
accepted: August 28, 2012
prepublished: February 8, 2013

For personal or educational use only. No other uses without permission. All rights reserved.
varied other biomedical subdomains [11–13], and much of the seminal work has been reviewed in [14, 15]. The reviews show how early exploratory work developed quite rapidly to influential as network analysis quickly gained acceptance, especially in epidemiology and sexual health. More recently, Christakis and Fowler used network analysis to claim that contagion or social influence affects spread of obesity [16] and smoking behavior [17], as well as happiness [18] in large social networks. Although their findings are quite interesting, the conclusions might have been much more convincing if they had considered other factors such as homophily; critiques have questioned the validity of their analyses [19, 20]. Another interesting application by Ma et al. [13] analyzed a healthcare forum for weight changes and the influence it had on the weight of people in relationship circles, and Burton [21] extended the analysis to video social network (YouTube) interactions for public health. Cancer forums have also been analyzed to identify temporal patterns and influential topics that promote community growth [6], while the structure of co-occurring symptoms in cancer patients has been analyzed using bipartite graphs [22].

Although there seems to be a substantial body of literature, existing accounts do not provide a unified quality assessment framework for the discovered communities. Previous studies have consistently shown that assessing whether the discovered communities are meaningful or good is application-specific [23, 24], and there have been some heuristic evaluations [25, 26]. We propose a user interaction similarity measure for assessing the quality of the discovered communities in healthcare forums. The specific aims of the study are to i) represent interactions in diabetes forums as networks, ii) detect and explore community structures in the networks, and iii) propose a quality measure characterizing the discovered communities.

2. Methods

The procedure for the study involves acquiring datasets of forum activity, cleaning the data and modeling a network. Then we detect communities within the networks and assess their structure using the proposed quality measure, and homophily measures. We use two well-known community detection algorithms that deal very well with large datasets; Greedy Optimization (GO) [27] and the Affinity Propagation (AP) [28] algorithms. GO is based on the Girvan-Newman algorithm; hierarchical agglomeration. The algorithm is extremely fast and suitable for large networks such as we used for this study. With a complexity of $O(md \log n)$, where $n =$ vertices, $m =$ edges and $d =$ depth of the dendrogram, the algorithm is based on modularity maximization, where the number of edges within a community are preferred to edges between communities. The AP algorithm is based on message-passing, where an initial data set is chosen at random and then refined in iterations. Obviously, the success of this algorithm is contingent on a good initial selection. The algorithm has a complexity of $O(N^2)$, where $N =$ size of the network (vertices).

2.1 Data Acquisition and Network Modeling

Although clean data is not always readily available for research, there are several techniques that can be used to crawl publicly available data on the Internet. We developed a web data extraction program in Python to crawl five diabetes forums on the Internet, comprising Spanish and English forums, and two of them were dedicated to juvenile diabetes. The forums have a total in excess of 140,000 registered users and over 1.6 million posts. Although the risk was minimal, we pseudonymized the forum data by one-way hashing the usernames and aliases, and removing other identifying data in the forums and user profiles, to guard against any potential exposure of personal health information [29].

An edge is created between a user (node) who creates a topic (edge) and a user who comments on the topic. We kept the edges directed (from commenter to thread creator) to allow us Authority-Hub and PageRank analysis [30] and weighted (weight is the number of edges with the

---

**Figure 1** A network of thread creation and comments is developing over time, where at Time $t_0$ (a) the initial network is established, and it develops progressively through Time $t_1$ (b) to a blossoming cluster at Time $t_2$ (c).
same commenter-creator pair), and we removed all self-loops, that is, where topic creators comment on their own topics, thereby creating a link to themselves. As an illustration, from Figure 1, we can observe how a typical forum network could develop over time. For instance in Figure 1a, a network is established at time $t_0$ when user B comments on topic L1 that was created by user A. At an arbitrary future time $t_1$, the topic creator posts a comment on his/her own topic, and a new user C posts a comment on topic L1 as well. With time, networks and sub-networks emerge as user D and E join the conversations.

### 2.2 Community Detection and Attribute Analysis

To analyze the network structure and topology, we used well-known community detection algorithms. We observed structure variables such as the diameter and characteristic path length at the network level [31] and centrality measures [32] at the node level. Even though many users are not willing to share most of their personal health information, we can still salvage knowledge from how they interact. As a guideline, we provide Table 1 that shows the attributes that we managed to collect by crawling the forums and user profiles for publicly disclosed information. The absence of complete data for most attributes makes the analysis quite difficult because we do not know if the available data are representative.

### 2.3 Quality Assessment Heuristics

There are quite many cluster validation methods that have been proposed [23], but there is no single framework that works in all situations, and most of them focus on the structure and interconnectedness (density) of the clusters. In this study, we relied on the modularity to validate the structure of the networks. A far greater challenge beyond the structure validation is determining how good the discovered clusters are.

We propose a quality function based on the homophily concept [33], where we assume the general hypothesis that similar users tend to cluster together. In some texts, homophily may also be referred to as assortative mixing or assortativity. Studies have suggested varied roles of homophily in adoption of health behavior in online social networks [34], and how contagion effects alone cannot always explain spread of health behavior [20, 35]. In uncontrolled empirical observations, the distinction between the effect of homophily, contagion or confounding variables on community clustering or adoption of health behavior is rather difficult to measure.

Measures of homophily work well when additional data about user attributes are available. In some instances, additional data about the user are not available, and Akcora et al. [36] showed how friendship networks can be used to analyze homophily where there is little information about the users. Akcora’s work assumes that we have explicit information about user friendships, groups and communities.

To solve the problem where we do not have explicit user friendship networks or cannot rely on additional user information, we propose a measure based on forum activity; a previously unconsidered factor. We show how the measure correlates with homophily measures based on degree or known attributes such as years-since-diagnosis or diabetes-type, and propose that the new measure can be used to assess networks where additional data about the users is scant, implicit or unavailable. The following are the three steps in the quality assessment procedure:

1. **Construct a set of neighbors from the top community.** For this analysis, the output is the neighborhood set of all nodes $n_i \in G$ (that is, immediate connections of $n_i$) where $G$ is the top community (a complete subgraph) in the network.
2. **Construct a set of neighbors from user interaction similarity analysis.** Just as in Step 1, we construct a set of neighbors for the nodes based on interaction similarity. We used a Java machine learning framework capable of dealing with large datasets to test several algorithms for computing similarity. The best algorithm for each situation is normally subjective and found by trial and error, and we chose the City Block and Tanimoto similarity coefficients because they yielded the best similarity values. We designed forum activity (without the network information) as binary (and not interval) data to represent users’ preferences. For example, a user is logged as preferring (1) a topic if the user com-

### Table 1  The percentage of users who disclosed personal data in the forums, where $x =$ data either unavailable or could not be extracted

<table>
<thead>
<tr>
<th>Location</th>
<th>$F_1$</th>
<th>$F_2$</th>
<th>$F_3$</th>
<th>$F_4$</th>
<th>$F_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Registered users</td>
<td>35589</td>
<td>72338</td>
<td>10527</td>
<td>1681</td>
<td>21968</td>
</tr>
<tr>
<td>Years from diagnosis</td>
<td>29183 (82%)</td>
<td>5787 (8%)</td>
<td>x</td>
<td>x</td>
<td>6590 (30%)</td>
</tr>
<tr>
<td>Diabetes type</td>
<td>28471 (80%)</td>
<td>54977 (76%)</td>
<td>x</td>
<td>x</td>
<td>6590 (30%)</td>
</tr>
<tr>
<td>Gender</td>
<td>18506 (52%)</td>
<td>15191 (21%)</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Age</td>
<td>1779 (5%)</td>
<td>7234 (10%)</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>HbA1c</td>
<td>1068 (3%)</td>
<td>3617 (5%)</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

**Figure 2**  The quality function based on user interaction similarity, where $N_c$ is the neighborhood sets from the discovered top community and $N_i$, for the user interaction similarity.

$$QF(G) = \sum_{i \in V} \frac{N_c \cap N_i}{\text{ang}(N_c \cup N_i)}$$
3. Results

Results suggest that it is feasible, from forum data, to model networks that contain meaningful sub-communities. We present basic network statistics before we delve into the details of the main findings. The statistics could help us understand the datasets, and may be useful for benchmarking.

3.1 Nature of the Diabetes Networks

The results in Table 2 show the basic characteristics of the five forums we studied, and reveal several interesting points. The high network diameter means that the small-world concept [30, 37] does not seem to apply in these kinds of networks. It is difficult to say why, but this may be partially explained by the fact that healthcare forum networks are formed by users with very little else in common, other than the disease. As a result, community nodes connect mostly with the central nodes, resulting in the distinctly dominant star topology with maximum centralization [9]. However, perhaps with additional data about other user activity and exchange of private messages and forum objects, the structure might look different; further analyses are required.

The average number of neighbors is not as high as could be expected from conventional types of social networks.

![Figure 3](image_url)

**Figure 3** The clustering coefficient distribution for two forums $F_1$ (3a) and $F_2$ (3b). The bold line in Fig. 3b is the Power-law distribution ($y = ax^b$) with a correlation of 0.863 to the network’s distribution.

### Table 2  Basic network characteristics from the 5 datasets and the community detection results. AP = Affinity Propagation algorithm, and GO = Greedy Optimization algorithm. The control column contains data from a previously published similar study (12).

<table>
<thead>
<tr>
<th></th>
<th>Control (12)</th>
<th>$F_1$</th>
<th>$F_2$</th>
<th>$F_3$</th>
<th>$F_4$</th>
<th>$F_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of nodes</td>
<td>7569</td>
<td>9679</td>
<td>16404</td>
<td>5553</td>
<td>470</td>
<td>2948</td>
</tr>
<tr>
<td>Number of edges</td>
<td>103592</td>
<td>109695</td>
<td>467677</td>
<td>767413</td>
<td>42924</td>
<td>75027</td>
</tr>
<tr>
<td>Users who posted (%)</td>
<td>–</td>
<td>27</td>
<td>23</td>
<td>53</td>
<td>28</td>
<td>13</td>
</tr>
<tr>
<td>Clustering coefficient</td>
<td>0.173</td>
<td>0.181</td>
<td>0.297</td>
<td>0.232</td>
<td>0.154</td>
<td>0.233</td>
</tr>
<tr>
<td>Network diameter</td>
<td>12</td>
<td>11</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>Characteristic path length</td>
<td>3.320</td>
<td>3.6</td>
<td>3.3</td>
<td>2.9</td>
<td>3.2</td>
<td>3.9</td>
</tr>
<tr>
<td>Average no. of neighbors</td>
<td>–</td>
<td>13</td>
<td>19</td>
<td>62</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Number of clusters (AP/GO)</td>
<td>–</td>
<td>360 (171)</td>
<td>2600 (242)</td>
<td>1290 (115)</td>
<td>193 (23)</td>
<td>670 (22)</td>
</tr>
<tr>
<td>Average cluster size (AP/GO)</td>
<td>–</td>
<td>27 (57)</td>
<td>6 (68)</td>
<td>4 (48)</td>
<td>2 (20)</td>
<td>4 (134)</td>
</tr>
<tr>
<td>Maximum cluster size (AP/GO)</td>
<td>–</td>
<td>1976 (3911)</td>
<td>2163 (8450)</td>
<td>470 (2597)</td>
<td>239 (131)</td>
<td>1351 (1372)</td>
</tr>
<tr>
<td>Modularity (AP/GO)</td>
<td>–</td>
<td>0.07 (0.34)</td>
<td>0.14 (0.35)</td>
<td>0.02 (0.20)</td>
<td>0.40 (0.28)</td>
<td>0.35 (0.35)</td>
</tr>
</tbody>
</table>
such as Facebook. The figures also show very low user participation rates in all the forums, suggesting high levels of activity among the few users who do participate actively.

### 3.2 Community Structure and Visualization

The networks exhibited Power-law degree distribution because of the few users who are very active and therefore have a disproportionately higher degree distribution than the rest of the users, resulting in a long tailed degree distribution. Figure 3 shows the clustering coefficient distribution for forums F1 and F2.

To explore the structure further, we used user attribute data from forum F1, where 82% of the users provided information about when they were first diagnosed with diabetes. We used the data to visualize the community structures as shown in Figure 4. The figure illustrates the dominant star topology, and shows how the central nodes in the network have at least two years’ experience with diabetes. The majority (78%) of the users who provided the data have been diagnosed less than two years ago (green). Patients with between two and ten years seem the most active in supporting the newly diagnosed patients. There is a comparatively smaller number of patients with more than ten years’ experience. Another interesting perspective is how detailed analysis can provide deeper insight into the actual communication links and their direction, as shown in Figure 5.

In practical terms, Figure 5 shows how easy it is to isolate users who may be struggling with the disease. For instance, a strong edge between a node with high in-
degree and another with low in-degree means the latter initiates most ties with the former, suggesting a needy user or someone highly motivated and involved in self-care. A central node with very low in-degree suggests a very hyperactive user who initiates ties with several users. Although these analyses reveal interesting aspects that suggest the discovered communities are in fact valid, we cannot say further about the quality of the communities without additional considerations.

3.3 Quality of the Discovered Communities

The quality measure is quite exploratory and is predicated on the concept of homophily [33], where similar users must belong to the same communities [38], and having closely similar users attaching themselves to different communities is counterintuitive.

From the results in Table 3, the degree-based homophily measures (measured between −1 and +1) as described in [39] are negative for all the networks, suggesting dissortativity. Some previous studies such as the one done by Newman [39] have shown that most social networks have high degree homophily because nodes with high degrees tend to be connected to other nodes with high degree. Contrary to these previous findings, our results suggest node degree is not a significant factor for cohesion in diabetes forums. A plausible explanation could be the long tailed degree distribution and the highly centralised topology, perhaps reflecting the very nature of online disease support.

Homophily based on degree seems inadequate to describe the nature of the studied diabetes social networks, and the proposed new measure that is based on forum interaction and activity could be complementary. The measure seems like a better approximation of the homophily measures based on observed user attributes (year-since-diagnosis and diabetes-type). Therefore the new measure could be useful to assess disease networks where additional user attribute data is scant or unavailable.

4. Discussion

Our analyses have shown that the topic creation-comment cycles produce meaningful implicit networks. These networks have been shown to correlate with user attributes such as the years-since-diagnosis. We further demonstrated that the networks contain community structures of good quality. There are a number of implications that these results have on our understanding of forum discourse in the medical web.

The use of directed edges allowed us to examine the hubs and the authorities in the communities. Identifying important people or leaders is critical for sustaining social interaction and dialogue. Leaders could be users with very high in-degree, meaning they are an authority and the rest of the users are likely to respect their opinion or advice. A more elegant measure, PageRank, allows us to examine the nodes to which the authorities connect. Having information about whom the authorities interact with the most can be useful for determining trust in the network. Comparing Forum1’s top 10 sets of in-degree and PageRank nodes, we found they had six nodes (60%) in common. Two of the six were female, three male and one undisclosed. They all had more than four years since diagnosis and the majority had type 2 diabetes, but one of the females had type 1 diabetes. The female seemed to derive her high rankings from her intimate knowledge and personal experience with insulin pump technology, and she reports a significant reduction in HbA1c following use of the pump.

 Authorities and trusted users can act as mentors for newly diagnosed patients, and this is quite a critical role since we discovered that about 78% of the users in forum F1 had fewer than two years with the disease. In contrast, users with high out-degree are literary information hubs. They help disseminate information because they are constantly probing for discussion, thus also sustaining dialogue in the forums.

Similarly, we have also shown how patients who may be struggling with the disease can be identified by examining the out-degree. However, a high out-degree can also indicate patients that are actively involved in their own care and perhaps struggling patients are not even reflected in the network. Text mining and natural language processing could be used to further examine forum dialogue in order to characterize the nature of the high out-degree nodes.

By examining the actual content of a few threads, we identified a number of “needy” users in the juvenile diabetes forum (F2) in Figure 5, but that was partially because the users were worried parents who constantly sought help for their young children (mostly Type 1 diabetes). The same forum has also showed the highest participation rate, where at least 53% of registered users participated in dialogue at least once.

4.1 Limitations

The primary limitation in this study is intrinsic to the nature of the data. Since the data we used does not contain information about private messages among the users, there might have been some private dia-

<table>
<thead>
<tr>
<th>TC (CB)</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greedy Optimization (user interaction)</td>
<td>0.6 (0.8)</td>
<td>0.6 (0.9)</td>
<td>0.5 (0.7)</td>
<td>0.6 (0.7)</td>
<td>0.7 (0.9)</td>
</tr>
<tr>
<td>Affinity Propagation (user interaction)</td>
<td>0.5 (0.6)</td>
<td>0.5 (0.7)</td>
<td>0.5 (0.6)</td>
<td>0.4 (0.6)</td>
<td>0.7 (0.8)</td>
</tr>
<tr>
<td>Homophily (degree)</td>
<td>−0.16</td>
<td>−0.17</td>
<td>−0.18</td>
<td>−0.14</td>
<td>−0.20</td>
</tr>
<tr>
<td>Homophily (years-since-diagnosis)</td>
<td>0.20</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Homophily (diabetes-type)</td>
<td>0.42</td>
<td>0.35</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
logue that could alter the network density, for example. It can be argued that this is not a serious threat, because in most instances, the users first contribute to open conversations where the network is instantly discovered.

Another limitation lies with the availability of additional user attribute data. This has limited our analysis to only two attributes that had sufficient data. This is a typical problem in analysis of healthcare data, and the measure we proposed embodies this core limitation.

A potentially more serious threat, not unique to our analysis, is that some forum data may be phony or forged, and missing or erroneous data has been shown to affect critical measures [40]. Forum administrators starting out a new forum may generate phony users and data to project success and popularity in order to entice new registrants. It is likely the vice permeates the whole Internet, but it not easy to determine its prevalence.

5. Conclusion

This study has shown how implicit networks can be modeled from forum interaction. Further, we showed that the resulting networks actually do make sense and correlate with some attribute-based homophily measures. Current empirical observations strengthen the basic science that supports future analysis and reasoning about diabetes patients as users of the medical web. Further studies are required before the nature of these healthcare online networks is more clearly understood; to foster adoption of healthy lifestyles among people with diabetes and other chronic illnesses.

Acknowledgments

Work supported in part by the Research Program for Telemedicine, Helse Nord RHF, Norway, and Centre for Research-based Innovation, Tromsø Telemedicine Lab. (TTL), Norwegian Research Council Grant No. 174934.

References

35. Arai S, Muchnik L, Sundararajan A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proceedi-


