Netw Model Anal Health Inform Bioinforma (2013) 2:147–157
DOI 10.1007/s13721-013-0033-y

ORIGINAL ARTICLE

Characterizing development patterns of health-care social networks

Taridzo Chomutare · Eirik Årsand · Gunnar Hartvigsen

Received: 30 November 2012 / Revised: 16 April 2013 / Accepted: 27 April 2013 / Published online: 14 May 2013

Abstract The growing amount of data in health-care social media requires innovative new analysis methods, which are elementary to exploration of relationship dynamics, in a bid to understand the new roles social media plays in health care. In this work, we use network analysis to explore the temporal nature of two large diabetes social networks, with a view to enhancing our knowledge of the development of community structures and cohesion factors. We compare our findings with analysis of two other nonhealth-care social networks. Current results reveal how diabetes online communities are very dynamic, suggesting diabetes patients are usually actively engaged for periods of less than a year, typically immediately following diagnosis. Additionally, we observe shrinking of both diameter and density, as well as disassortative mixing. The presented empirical study informs future online intervention strategies for promoting health behavior and lifestyle changes among people with diabetes.

keywords Social networks · Social network analysis · Community detection · Temporal graphs · Diabetes

T. Chomutare (✉) · E. Årsand
Norwegian Centre for Integrated Care and Telemedicine, University Hospital of North Norway, 9038 Tromsø, Norway
e-mail: taridzo.chomutare@telemed.no

G. Hartvigsen
Department of Computer Science, University of Tromsø, Tromsø, Norway

1 Introduction

Diabetes is a growing global threat. The growth is overtly dramatic, particularly in the case of Type 2 diabetes which is associated with poor eating habits, sedentary lifestyles and the aging population. Lifestyle-related diseases can lead to poor quality of life for the individual and to high costs for the health-care system. New figures show 366 million people have diabetes worldwide, and the figure is projected to grow to 552 million by 2030 (IDF 2011). The threat transcends international economic boundaries, and Africa could see a 90 % increase by 2030.

Online social forums and networks are emerging as platforms for health-care interventions and convenient health-care information access and support tools (Burton et al. 2012). Present understanding of temporal development patterns and the factors that influence interaction in health-care online communities seem quite limited. In the presented work, we explore the development patterns of diabetes online communities and seek to understand the factors that characterize and influence community development in this domain. Network analysis [and community detection (Fortunato 2009)] is one of the most practical ways of facing the challenges of mining (Wegener et al. 2013) the growing data for meaningful information.

For our analyses, we designed networks of user interaction in two diabetes and two nonhealth-care forums and applied an existing community detection algorithm to time-partitioned datasets (snapshots). To better understand the periodical changes, we used similarity analysis based on Jaccard similarity index and cohesion analysis based on centrality measures and user attributes. Maintaining focus on diabetes social networks allows us to form generalizations applicable to the disease group.

We use real-world data from large forums to explore the temporal nature of interaction networks with a view to enhancing our knowledge on the (1) temporal patterns of the communities, (2) attributes that influence temporal community cohesion and (3) salient patterns that characterize the networks. Empirical observations based on real-world data are important for validating and informing existing general models.

2 Related literature

Previous research on temporal trends has included group evolution dynamics (Palla et al. 2007; Chakrabarti et al. 2006; Lin et al. 2008; Bródka et al. 2012). Other studies focused on content popularity and predicting social ties (Almansoori et al. 2012) and information flow (Yang and Leskovec 2011). Although many previous studies have substantially enhanced our understanding of group evolution dynamics, far less attention has been paid to healthcare networks.

Recent research in health care has studied development of community node attributes (Ma et al. 2010) and development of community growth phases (Durant et al. 2011), but in this study we focus on the development of community structures and the likely forces behind them.

In the former study, Ma et al. (2010) analyzed temporal weight changes over a 5-month period. The study reported positive correlations between the user neighborhood size and the weight changes in the user's neighborhood. Although the study was done for only a short period, and therefore difficult to say if the noted correlations are sustained, it nonetheless enhances our understanding of online influence and its propagation over time. The only drawback could be that no reference to known temporal models was made.

The latter study by Durant et al.(2011) analyzed data from six cancer forums and identified growth stages for the different online communities as well as topics that promote growth, using a new phase detection algorithm and a response function. The study concludes that treatment discussions rather than diagnosis discussions are more engaging to cancer patients and thus also promote growth. In this stud,y community developments are not considered.

A recent study by Bródka et al. (2012), although not health-related, proposed a group evolution discovery (GED) method for analyzing evolution of group structures or communities. The study provides a complete synthesis of temporal patterns of community structures to date. However, this method, as well as other approaches that presume group overlaps (Palla et al. 2005), seems to suffer a weakness when node sets from one period to the next have a consistently small number of elements in the intersection set, as is the case with some real-world networks. The algorithm consistently results in *formation* and *dissolution* patterns of evolution and requires additional supporting information to make more real-world sense.

We propose augmenting the framework by Bródka et al. with similarity measures to quantify the development patterns at both the network and community levels. In addition, we do cohesion analysis to reason about the development patterns in the context of diabetes and show that we gain new perspectives of the evolution even when the networks are extremely volatile. We also attempt to discover the unique development patterns that distinguishes health care from closely similar general social networks.

3 Methodology

An overview of the methodology is illustrated in Fig. 1, where networks are designed from user interaction data crawled from two diabetes forums. It is important to note that relationships in forum-like communities can be difficult to ascertain since there are no explicit relationships. An alternative is to form implied relationships from how the users interact with each other, forming bonds and ties from exchange of objects and through social discourse. To analyze the community structures, we applied an established community detection algorithm on the networks: the greedy optimization (GO) algorithm (Clauset et al. 2004). In addition, we formulated similarity and cohesion analysis using a blend of Java machine learning libraries and network visualization tools. We also explore some properties such as the temporal density and diameter to distinguish health-care development patterns from other general social networks.

3.1 Data extraction and modeling

Most social media data can be extracted using several web data extraction methods. In this instance, we developed a well-behaved python program to crawl and parse HTML data from two diabetes forums. The security risk for the users seemed quite small, but we nonetheless pseudonymized the data before the analysis. The crawling was done over a few days in December 2011 and January 2012 and pertains data ranging from 2006 to 2011.

The data contained threads and comments without any explicit relationships. We modeled a network from the thread creation and commenting cycles. The only major limitation of this approach is that we could not capture the private conversations and exchange of objects. However, most private networks blossom from public interaction, where the network has already been established and

recorded. Our network construction method is shown in pseudocode in Algorithm 1.

Algorithm 1 Pseudocode for constructing the networks

```
Input: All posts in the forum
Output: Network of thread creator and commenter
network ⇐ [commenter, creator, weight]
for all threads t in forum f do
  for all posts p in t do
    \*The poster of first post in the thread is the creator,
    the rest are comments*\
    if first post in thread then
      p.poster ⇐ creator
    else
      p.poster ⇐ commenter
    end if
    if not (commenter,creator) pair exists in network or
    commenter==creator then
      network.append(commenter,creator,weight=1)
    else
      \* Ignore duplicate pair in same thread, ignore self-
      loops*\
      continue
    end if
  end for
end for
for all (commenter,creator) pairs in network do
  \*Unique pairs in different threads increase weight*\
  weight ⇐ numberofunique(commenter, creator)pairs
  network.delete_duplicate_pairs
  network.update_weight(commenter,creator,weight)
end for
```

Algorithm 1 is the construction of a network based on the users as nodes. A person who creates a thread t is the *creator*, who is connected to a person who comments on the thread, *commenter*. The edges are directed, from a *commenter* to *creator*. It should be noted that an edge is immediately established at the first connection between a unique pair, and that no edges are created among the people that comment. These restrictions maintain a simplified abstraction for constructing the network. In addition, the *weight* of edges increase when a unique pair of nodes appear on a different thread.

Table 1 shows the characteristics of the forum data and the resulting static network. We can deduce from the number of nodes that only 27 and 23 % posted at least one item in $Diabetes_1$ and $Diabetes_2$, respectively. $Diabetes_1$ has data for 4 years between 2008 and 2011, while $Diabetes_2$ has data for 6 years between 2006 and 2011. It is easy to observe that the clustering coefficient and the average number of neighbors are both much higher in the diabetes networks.

For nonhealth data, we used the Slashdot dataset used by Gómez et al. (2008) and the Facebook dataset used by Viswanath et al. (2009), with three and four time periods, respectively. The former dataset is a network of threads and comments in a technology-related news website, and the latter network consists of Facebook wall posts for New Orleans networks. We chose the former dataset because it closely resembles our own design of the network based on forum thread creation and commenting. We selected the latter because it represents networks created from explicitly known relationships and thus contrasts with our own inference-based design.

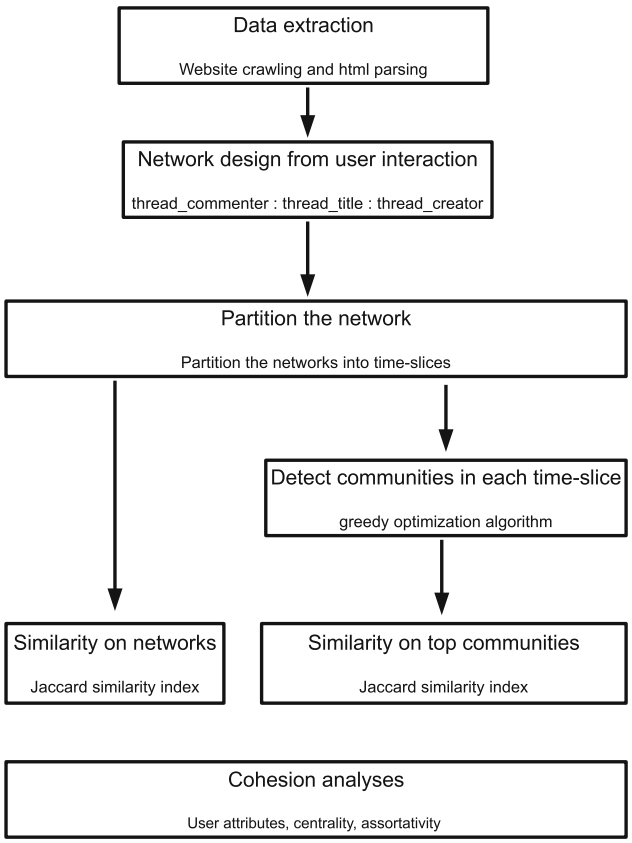


Fig. 1 The methodology for the study, summarizing the flow of the steps for each diabetes forum

3.2 Temporal analysis

Although there does not seem to be a unified framework, in most recent studies, temporal analyses have been based on partitioning of the networks into arbitrary time periods or snapshots. To start off, we present some of the major failing points for static analysis of health-care social networks in Fig. 2, as the network grows from period T_0 to T_2. In this instance, a static network is the absolute representation of the network from the beginning up to the cutoff period.

In the figure, static analysis of the network makes less and less sense as time progresses and the network changes because all nodes, including both new nodes and retired nodes, are treated as active. For example, Fig. 2c shows all the data from the beginning (with nodes 1, 2, 3 and 4) up to period T_2, and when looked at statically, without the distinguishing coloring, a lot of evolutionary details are

obscured. For instance, it may seem as if, at time T_2, the network is invigorated into a dense network when in fact node 1 and 3 are retired nodes.

In this work, we partitioned the datasets for each forum into periodical sub-datasets, to be able to isolate activity in specific time periods. Although we used annual time slices, we should highlight the problem of determining the optimal time partitions or slices. The next three subsections describe our methodology for temporal analysis to some detail.

3.2.1 Community detection

We used a well-studied community detection algorithm, the greedy optimization (GO) (Clauset et al. 2004) of the modularity score. This is based on Girvan-Newman algorithm and is based on hierarchical agglomeration. The algorithm is extremely fast and suitable for large networks. It has a complexity of $O(md \log n)$, where n is the number of vertices, m is the number of edges and d = the 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; see the pseudocode in Algorithm 2. It should be noted that we did not do a thorough evaluation of other methods, although our choice was influenced by our initial experiments in earlier work (Dias et al. 2012; Chomutare et al. 2013). We only needed a single well-studied method that we could apply uniformly to the datasets.

Algorithm 2 Pseudocode for the greedy optimization algorithm

```
Input: A network or graph G = (V, E)
Output: Clustering C of the graph
\*initial clustering C of single node clusters*\
C ⇐ singletons
matrix M
repeat
   find clusters C_i,C_j ∈ M with highest ↑ modularity
   merge C_i and C_j
   update matrix M
   \*has to stop if no improvements on modularity is
   possible*\
until |C| ≤ 0
```

3.2.2 Community similarity

Next, we compare the communities for the different years. In this context, communities are coherent subnetworks in the time-sliced network, that is, clusters of nodes with dense connections. We used the Jaccard similarity index to compare the networks and communities. Whereas the index has been used to compare the two datasets as a form of external validation, in this work we explore its use for analyzing two datasets from two periods, where the

Table 1 The basic characteristics of the forums and the modeled networks

	Diabetes_1	Diabetes_2	Slashdot	Facebook
Users	35,589	72,338	>1m	>1m
Nodes	9,679	16,404	51,083	46,952
CC	0.181	0.297	0.012	0.084
ND	11	9	17	21
CPL	3.6	3.3	5.3	6.1
AN	13	19	5	8

CC clustering coefficient, *ND* network diameter, *CPL* characteristic path length, *AN* average neighbors

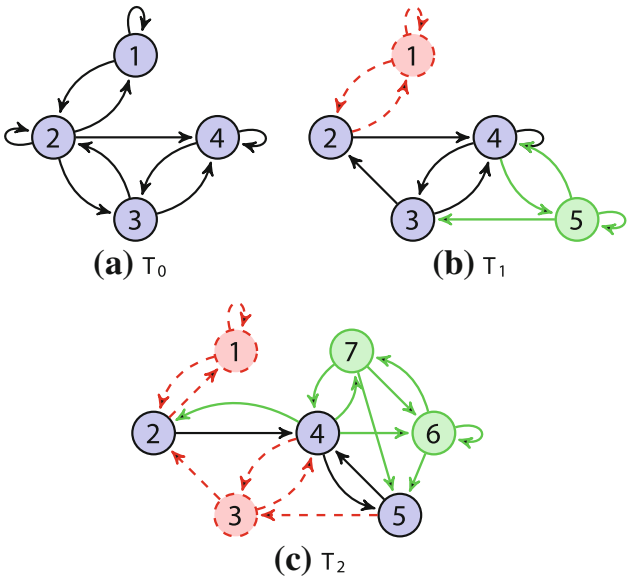


Fig. 2 Evolution of social ties through time t_0 to arbitrary future time t_2. The nodes and edges with *dotted red lines* are dissolved social ties. The nodes and edges with *solid green lines* are the new social ties in the period. The networks can *continue*, *shrink*, *grow*, *split*, *merge*, *dissolve* or *form* completely new ones (Bródka et al. 2012; Palla et al. 2007) (color figure online)

community C_a can have n nodes $\{x_1, \ldots, x_n\}$ at time T_0 and the community C_b is the similar community at time T_1 with m nodes $\{x_i, \ldots, y_m\}$, where x_i can be a subset of C_a. The aim of the analysis is twofold:

– first, for quantifying the similarity at the network level (see Fig. 3), declaring the first (or preceding) network as the benchmark:

$$J(C_a, C_b) = \frac{|C_a \cap C_b|}{|C_a \cup C_b|} \quad (1)$$

where C_a is the benchmark network at time T_0, and C_b is the network at an arbitrary future time T_n;
– second, for quantifying the similarity for the communities in each time period, where each network

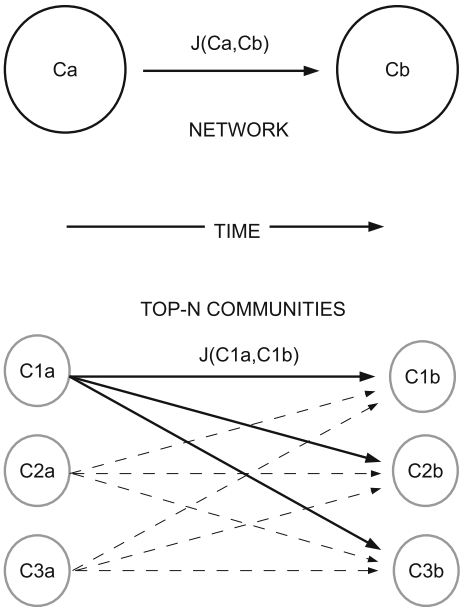


Fig. 3 The *top large circles* resemble the network and the *bottom circles* represent the top communities in one period to the next, as they are compared using the Jaccard index, illustrated by both the *solid* and *dotted lines*

connection is annotated with a community identifier, for identification in the subsequent period. Communities in each period are compared with communities in the subsequent period. In this case, each of the top three communities is compared to each of the top three in the next period (see Fig. 3).

These comparisons allow us to gauge the stability of the communities from one period to the next period or periods, providing both a course and detailed overview of how the communities form or dissolve over time.

3.2.3 Community cohesion heuristics

Finally, we analyzed community cohesion to understand the bonding factors. Several types of attributes were available for this analysis: (1) years-since-diagnosis, (2) type-of-diabetes, (3) HbA_{1c}, (4) age and (5) gender.

We further looked at (1) degree assortativity, (2) diameter (3) density and (4) average degree. Assortativity as described by Newman (2002) (also called homophily in the literature (McPherson et al. 2001) is the tendency for similar or dissimilar nodes to connect to each other. Degree assortativity describes the extent to which nodes of similar degree cluster together. For example, people with many connections in popular social networks tend to connect to other people with many connections. In this study, we used a Java machine learning library that implements the assortativity formula, which is merely the Pearson correlation coefficient; (see Equation 2 and Equation 21 in Newman 2002).

$$r = \frac{\sum_{xy} xy(e_{xy} - a_x b_y)}{\sigma_a \sigma_b} \tag{2}$$

where $a_x = \sum_y e_{xy}$ and $b_y = \sum_x e_{xy}$, and e_{xy} represent the fraction of edges between the vertices x and y, and σ_a and σ_b are standard deviations.

Finally, we studied diameter and density, which are key to understanding networks because they describe the interconnectedness of the nodes and can be a plausible basis for distinguishing network characteristics. We also focused on understanding the changes in the average degree over time.

4 Results and discussion

The core of the results of the major steps we took to analyze the health-care networks: (1) community detection, (2) community similarity, and (3) community cohesion analyses are illustrated in Fig. 4, where Fig. 4a shows a zoomed-in section of one of the detected communities at time T_0. At the future time T_1, we expected the new nodes (users) to attach to existing community structures (Fig. 4b). This expectation is based on the fundamental goals of social media for chronic illnesses; maintaining membership and sustainable support structures for new and existing users. However, what we observe is that completely new structures are formed, with only a few of the original nodes (Fig. 4c). This finding is consistent in both health-care forums and has far reaching consequences. One of the consequences is that earlier reports in the literature of increased or decreased engagement in health-care social networks could be incidental and connected with the number of active new users, rather than sustained engagement by old users. Another consequence is that diabetes social networks have to reconsider their models for patient engagement for a more sustained participation.

4.1 Community detection

Community detection over the time periods is shown in Table 2, where the basic characteristics are summarized. One interesting factor to observe is that there seems to be some distinction among the different datasets. For instance, we see Facebook has a high number of clusters (C, F_4 row) and small average cluster sizes (CS, F_4 row), but the opposite is true for both diabetes datasets, where the number of communities (C, F_1 and F_2 rows) are comparatively smaller and the average cluster sizes (CS, F_1 and F_2 rows) are higher. This observation could be partially

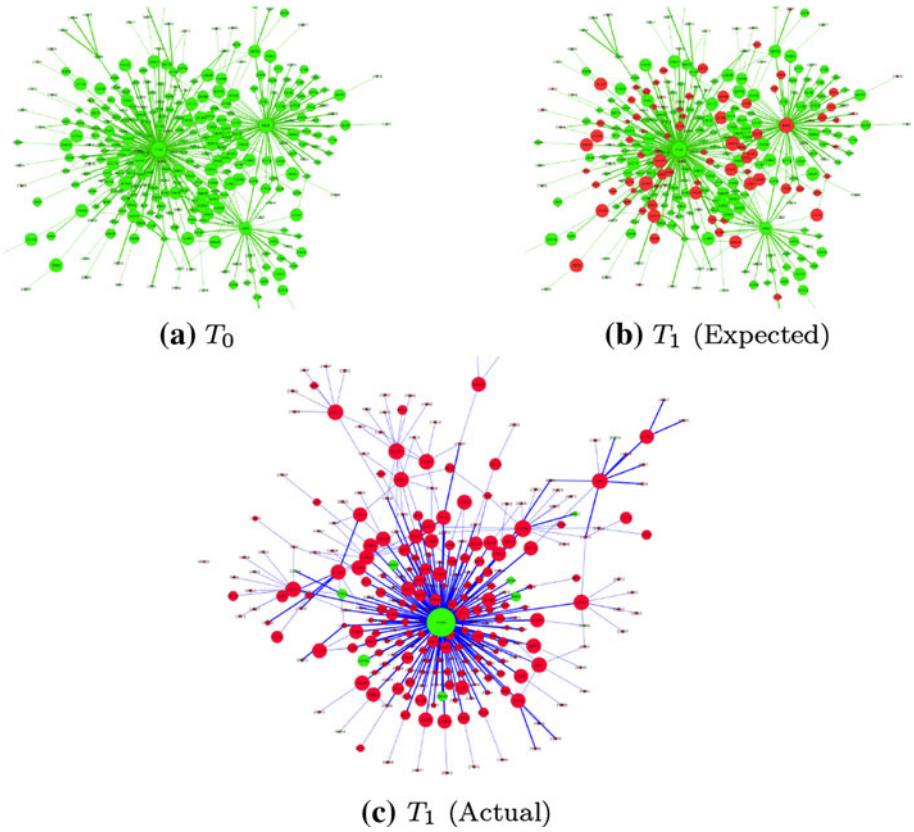


Fig. 4 Community evolution from time t_0 to arbitrary time t_1. *Green* original nodes at the starting period t_0, *red* new nodes appearing at time t_1. The illustration is a zoomed-in visualization of one of the discovered communities in 2009 and 2010 from $Diabetes_1$ (color figure online)

explained by the larger average number of neighbors, perhaps reflecting the need for patients with similar problems to cluster together. Further investigation is required before we can fully understand how patients organize themselves to get support or empathy from peers.

Another interesting observation from Table 2 is the differences in modularity. The nonhealth networks (M, F_3 and F_4 rows) have a much higher modularity than the diabetes networks (M, F_1 and F_2 rows). A plausible partial explanation may be the fact that the nonhealth networks are more dense, hence also easier to divide into communities.

4.2 Community similarity

The network similarity results at the bottom of Table 2 reveal something quite unexpected: there is little similarity between the time-sliced networks over the period under review. In this instance, similarity measures the node composition in the network or communities. The highest recorded Jaccard similarity index for $Diabetes_1$ at network level is 0.15 from 2009 to 2010. The details of top community similarities are shown in Table 3, where the similarity values are much smaller. It should be noted that Facebook similarity numbers are much higher, suggestive of a more stable and persistent friendship network.

This finding is surprising for several reasons. First, this implies users are only active for short periods of time. User networks are volatile and do not survive the year. However, the year-on-year activity levels as shown by the number of edges and nodes seem to grow steadily in health-care forums as shown in Fig. 5a, b, before an eventual decline or fluctuation. Although the number of active edges and nodes grow, there seems to be a corresponding number of edges and nodes that quit or retire. Since the node composition similarity from one period to the other is very low, we can reasonably deduce that the increase in activity is a result of new active nodes.

All the networks (Fig. 5c–f) exhibited a scale-free nature as shown in the cumulative degree distribution ($D(k) \approx k^{-2}$). This is important to highlight as a common property across the studied networks. This observation implies that only a few nodes have a very high degree, while most nodes have a very small degree, resulting in a long-tailed or power-law degree distribution.

Second, the volatility of the networks feels dramatic, because we expected users to remain engaged in online

Table 2 Community detection of the 4- and 6-year periods for the first forum F_1 and the second forum F_2, respectively

	Development periods					
	1	2	3	4	5	6
C						
F_1	24	59	86	110	-	-
F_2	20	17	84	123	159	172
F_3	83	116	85	-	-	-
F_4	565	1,831	2,101	2,124	-	-
CS						
F_1	75	53	39	34	-	-
F_2	36	106	37	73	59	96
F_3	223	142	159	-	-	-
F_4	16	11	20	10	-	-
M						
F_1	0.334	0.252	0.31	0.31	-	-
F_2	0.297	0.227	0.219	0.230	0.235	0.23
F_3	0.581	0.573	0.589	-	-	-
F_4	0.660	0.597	0.624	0.861	-	-
NS						
F_1	-	0.11	0.15	0.14	-	-
F_2	-	0.09	0.12	0.14	0.14	0.05
F_3	-	0.14	0.13	-	-	-
F_4	-	0.41	0.39	0.41	-	-

F_3 and F_4 are Slashdot and Facebook datasets, respectively

C clusters, *CS* average cluster size, *M* modularity, *NS* network similarity

Table 3 Similarity of communities from one period to the next in $Diabetes_1$ (F_1 in Table 2)

	Cluster 1	Cluster 2	Cluster 3
2008/9			
Cluster 1	0.06	0.04	0.06
Cluster 2	0.02	0.02	0.01
Cluster 3	0.02	0.01	0.03
2009/2010			
Cluster 1	0.02	0.03	0.01
Cluster 2	0.02	0.08	0.02
Cluster 3	0.12	0.03	0.02
2010/2011			
Cluster 1	0.01	0.11	0.02
Cluster 2	0.05	0.01	0.04
Cluster 3	0.03	0.02	0.04

communities for longer periods since diabetes is a chronic disease, as well as to help other new users. Even though these results initially seem unexpected, they are supported by further attribute analysis. There are about 80 % active newly diagnosed patients (less than 2 years) in any year, suggesting that once the new patients get to grips with their diagnosis, they sever all network ties with other patients. Only a few highly motivated patients remain active over longer periods.

4.3 Community cohesion heuristics

In the succeeding subsections, we explore some of the attributes that are highly relevant to analysing cohesion in diabetes networks. There are potentially quite many attributes that we could discuss, but we highlight just a few based on expert opinion and what the forum users could have provided in their public profiles. In the process, we also explore the discussion and debate around use of personal health information, and its availability for information processing.

4.3.1 Years-since-diagnosis

Years-since-diagnosis was an obvious cohesion factor because almost 80 % of the registered users have been diagnosed less than 2 years ago in any of the periods. This is indicative of how online communities have become the preferred source of support for newly diagnosed patients. While some newly diagnosed patients also supported other new patients, the majority only acted as information hubs. The authoritative role (Kleinberg 1999) was assumed by patients with 2–10 years' experience after diagnosis as can be seen in Fig. 6a, a zoomed-in figure representative of the communities. From the figure, we can observe that a huge majority of the nodes are the newly diagnosed users (green), which are connected to central figures who have more experience with diabetes (red,black).

4.3.2 Type-of-diabetes

Type-of-diabetes is an intuitive attribute for cohesion because the main types of diabetes (Type 1 and Type 2) have several lifestyle and behavioral goals in common: mostly blood glucose management, dietary and physical activity goals. Only about 5%-10% of patients with diabetes have type 1 diabetes, and this likely obfuscate some community patterns unique to type 1 diabetes.

4.3.3 HbA_{1c}

HbA_{1c} is a measure (in percentage) of long-term blood glucose levels and is an important outcome for people with diabetes. It is not certain why forum users did not disclose their HbA_{1c}. Only 3 and 5 % of the users of $Diabetes_1$ and $Diabetes_2$, respectively, disclosed their values. The majority of the disclosed HbA_{1c} values were between 6 and 9, which is considered an acceptable range for people with

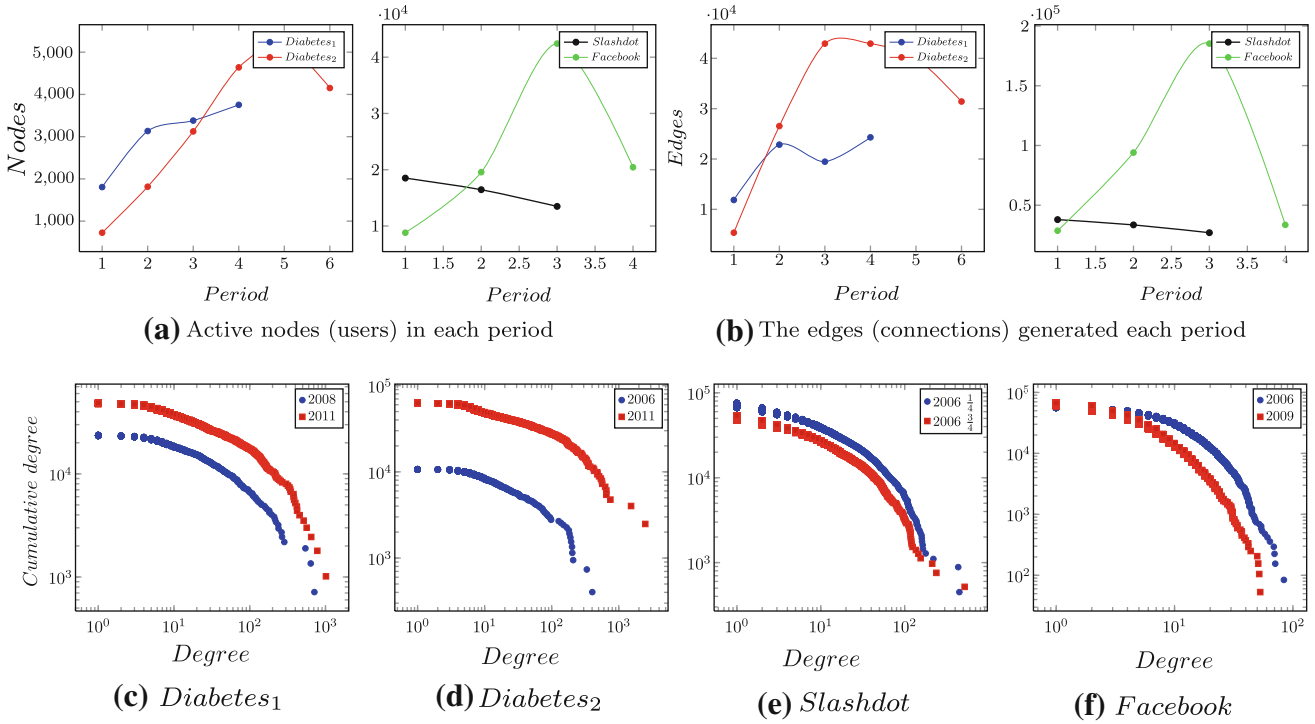


Fig. 5 Characteristics of the temporal networks. **a**, **b** *i*llustrates the levels of activity in the different time periods. **c**–**f** Shows the cumulative degree distribution against the degree for the four forums for their respective first and last periods only (color figure online)

diabetes (values around 7 are more desirable). This suggests that people who manage their disease well are more likely to disclose their HbA_{1c}.

4.3.4 Age and Gender

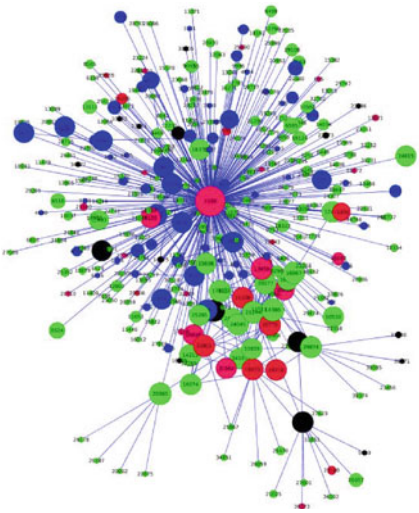
Age was disclosed by only 5 and 10 % in $Diabetes_1$ and $Diabetes_2$, respectively. Therefore, it is difficult to explore its impact. The fact that users do not want to provide their age may be suggestive of its irrelevance. On the other hand, just over half disclosed their gender.

4.3.5 Assortativity and other network attributes

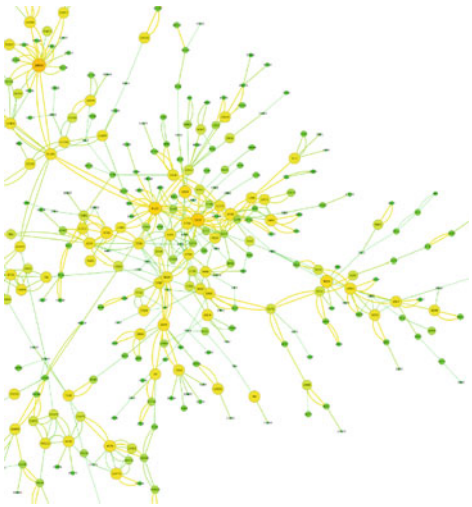
In this section, we form generalizations about the salient characteristics that distinguish health-care datasets from other social networks. We base the analysis on the trends of (1) assortativity, (2) network diameter, (3) network density and (4) average degree over the studied period.

We can observe from Fig. 7 that there is significantly higher assortative mixing in the Facebook network, while there is mostly disassortative mixing in the health-care datasets. This result is not surprising because Facebook networks are more decentralized, and it is easy for users with high degree to be connected to other users with high degree, as can be seen in Fig. 6b. We further observe from Fig. 6c that health-care network structure, as sampled here, has a far more centralized star topology. This means users with very high degree connect to several users with very low degree, hence the mostly negative assortative mixing. Perhaps this reflects the very core of diabetes forums, where a few experienced and knowledgeable users tend to support a large number of newly diagnosed users as seen in Fig. 6a. Information dissemination becomes vital as it is placed in the hands of a few central nodes that have a very short path to a large number of nodes. These findings can be contrasted to Newman (2002) results that indicate most social networks exhibit assortative traits.

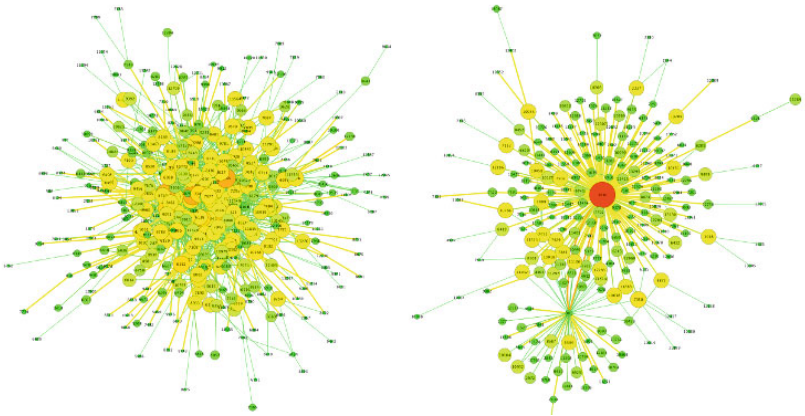
It follows from the network structure argument in the preceding paragraph that the diameter for health-care datasets is much lower than nonhealth-care datasets. In terms of the temporal patterns, it seems from Fig. 7 that the network diameter falls with time in health-care datasets, while it actually increases in the other datasets. The density of the nonhealth-care datasets was extremely low. On the other hand, the density in the health-care datasets, while also very low but higher than in nonhealth-care, exhibited a diminishing trend over time, from 0.010 in period one for $Diabetes_1$ to 0.002 in the last period. Although some recent studies like (Leskovec et al. 2005) have shown that density increases and diameter shrinks over time for most networks, our results suggest both density and diameter shrink



(a) Cluster zoom-in for attribute years-since-diagnosis, where green=0-1 year after diagnosis, red=2-10 years after diagnosis, and black= more 10 years after diagnosis. (blue means no data was provided by the user)



(b) This is a zoom of one of the detected communities in the *Facebook* network, showing a *decentralized structure*. The nodes are colored and sized by in-degree.



(c) This is a zoom of some of the detected communities in $Diabetes_1$, showing a distinct *star topology*. The nodes are colored and sized by in-degree.

Fig. 6 Some of the visualizations of the communities found in the networks using the GO community detection algorithm at different zoom levels in different time slices (color figure online)

in health-care forum data. This finding may be partially explained by the tendency to attach to the central and more experienced node rather than for novices to interconnect among themselves, resulting in shrinking of both density and diameter.

The average degree is higher in diabetes networks than the nonhealth-care networks, and perhaps this reflects the information needs of many newly diagnosed users as they try to get to grips with diabetes, while they communicate with only a few experienced users.

4.4 Limitations

A potential limitation of the presented analysis is intrinsic to the nature of the data. Although is it unlikely for the datasets we used, the data are susceptible to tampering. For example, it has been noted that some website administrators fabricate initial data to project success and popularity. This is partially because it is difficult to attract new users in the early phases of the communities. Additionally, it should be noted that although all the obtained data are in the public domain, some users divulge their personal health information without fully contemplating the privacy implications.

Again, we could have examined several temporal models and clustering or community detection algorithms to increase validity our results. Although this could certainly have given us more insight, the algorithm we used is well accepted and deals quite well with large datasets and has been the subject of much study. Taken together, these limitations are not critical enough to invalidate our analysis or results.

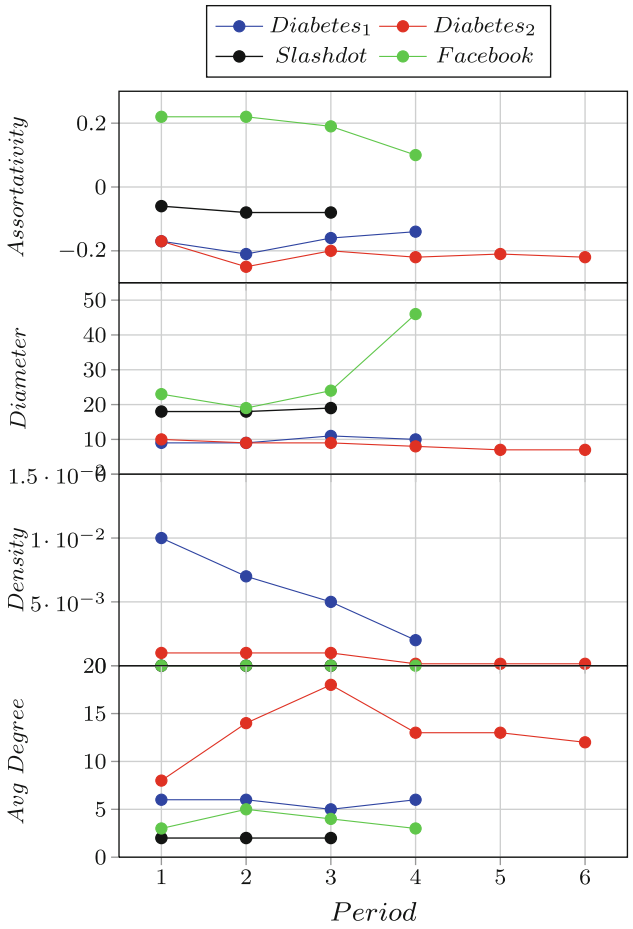


Fig. 7 Comparison of the temporal networks in terms of the average degree, network diameter and degree assortativity. It is interesting to note that the diabetes networks are always on the same side of the spectrum (color figure online)

5 Conclusion

In this work, we applied a novel ensemble of methods for developing a framework to better understand the unique patterns of development of diabetes social networks. We have shown how existing methods may fail to meaningfully describe extreme development patterns where communities constantly dissolve and form. Extending existing temporal models with quantifiable similarity measures and reasoning about community cohesion seemed to reveal potentially hidden details in the real-world networks.

More important, the empirical findings in this study provide a new understanding of social engagement in diabetes social networks. The most surprising finding to emerge from this study is that diabetes communities are very dynamic and short lived, implying that users engage only for short periods and do not sustain any noteworthy networks or communities. Perhaps, the lack of will to invest themselves in online communities is reflected in their reluctance to disclose personal data. Finally, we observed the shrinking diameter and density and the disassortative mixing in the diabetes networks.

Current work informs future intervention strategies for promoting health behavior and lifestyle changes among people with diabetes, but further research is required before much of the implications of the discovered patterns are more clearly understood.

Acknowledgments This work was supported in part by the Research Program for Telemedicine, Helse Nord RHF, Norway, and the Centre for Research-based Innovation, Tromsø Telemedicine Laboratory (TTL); Norwegian Research Council Grant No. 174934.

References

Almansoori W, Gao S, Jarada T, Elsheikh A, Murshed A, Jida J, Alhajj R, Rokne J (2012) Link prediction and classification in social networks and its application in healthcare and systems biology. Netw Model Anal Health Inf Bioinf 1:27–36

Bródka P, Saganowski S, Kazienko P (2012) Ged: the method for group evolution discovery in social networks. Soc Netw Anal Min, pp 1–14

Burton S, Tanner K, Giraud-Carrier C (2012) Leveraging social networks for anytime-anyplace health information. Netw Model Anal Health Inf Bioinf 1:173–181

Chakrabarti D, Kumar R, Tomkins A (2006) Evolutionary clustering. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '06. ACM, New York, pp 554–560

Chomutare T, Arsand E, Fernandez-Luque L, Lauritzen J, Hartvigsen G (2013) Inferring community structure in healthcare forums: an empirical study. Methods Inf Med 52(2)

Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70(6):066111

Dias A, Chomutare T, Botsis T (2012) Exploring the community structure of a diabetes forum. Stud Health Technol Inf 180:833–7

Durant KT, McCray AT, Safran C (2011) Identifying temporal changes and topics that promote growth within online communities: a prospective study of six online cancer forums. IJCMAM 2(2):1–22

Fortunato S (2009) Community detection in graphs. CoRR abs/0906.0612

Gómez V, Kaltenbrunner A, López V (2008) Statistical analysis of the social network and discussion threads in Slashdot. In: Proceedings of international World Wide Web conference, pp 645–654

IDF (2011) Idf diabetes atlas. Int Diabetes Fed 1(5)

Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46:604–632

Leskovec J, Kleinberg J, Faloutsos C (2005) Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining KDD '05. ACM, New York, pp 177–187

Lin YR, Chi Y, Zhu S, Sundaram H, Tseng BL (2008) Facetnet: a framework for analyzing communities and their evolutions in dynamic networks. In: Proceedings of the 17th international conference on World Wide Web. WWW '08. ACM, New York, pp 685–694

Ma X, Chen G, Xiao J (2010) Analysis of an online health social network. In: Proceedings of the 1st ACM international health informatics symposium IHI '10. ACM, New York, pp 297–306

McPherson M, Lovin L, Cook J (2001) Birds of a feather: homophily in social networks. Annu Rev Soc 27(1):415–44

Newman MEJ (2002) Mixing patterns in networks. Cornell University Library

Palla G, Derényi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043):814–818

Palla G, lászló Barabási A, Vicsek T, Hungary B (2007) Quantifying social group evolution. Nature 446:2007

Viswanath B, Mislove A, Cha M, Gummadi KP (2009) On the evolution of user interaction in facebook. In: Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09)

Wegener D, Rossi S, Buffa F, Delorenzi M, Rping S (2013) Towards an environment for data mining based analysis processes in bioinformatics and personalized medicine. Netw Model Anal Health Inf Bioinf 2(1):29–44

Yang J, Leskovec J (2011) Patterns of temporal variation in online media. In: Proceedings of the fourth ACM international conference on Web search and data mining WSDM '11. ACM, New York, pp 177–186