Social Connection Dynamics in a Health Promotion Network

Eric Fernandes de Mello Araújo, Michel Klein and Aart van Halteren

Abstract The influence of social connections on human behaviour has been demonstrated in many occasions. This paper presents the analysis of the dynamic properties of longitudinal (335 days) community data (n=3,375 participants) from an online health promotion program. The community data is unique as it describes how the network has evolved since its inception and because the information exchanged through the network was predominantly about the achievements of participants in the program and therefore influencing behavior through social comparison. The analyses show that the largest component of the community network has characteristics of a small world network. The analyses also show that connections are formed according to a strong attachment preference according to the gender, and a weaker homophily for Body Mass Index. The presented analysis can serve as basis for creating novel interventions that influence physical activity behavior through social connections.

1 Introduction

Social Network Analysis (SNA) is a broad research area, with applications in many different disciplines, incorporating aspects of sociology, social psychology and anthropology [19]. SNA is useful for studying nodes' influences within a network, and how behaviours, opinions or sentiments are spread in social networks [3, 6]. The nodes with an important position can be used to find points of interventions to stop or to enhance the process under study [1, 2, 9, 11, 21].

However, many of the contributions in this field are based on static networks, without taking the time dimension into account. The dynamics of the network can reveal more about how the network evolves over time [5, 22].

Eric Fernandes de Mello Araújo (e-mail: e.araujo@vu.nl)✉ · Michel Klein (e-mail: michel.klein@vu.nl)✉
VU Universiteit, Amsterdam, The Netherlands

Aart van Halteren (e-mail: aart.van.halteren@philips.com)✉
Philips Research, Eindhoven, The Netherlands

H. Cherifi et al. (eds.), *Complex Networks & Their Applications V*,
Studies in Computational Intelligence 693,
DOI 10.1007/978-3-319-50901-3_61

In this paper, we investigate the dynamic properties of longitudinal (336 days) community data (n=3,375 participants) from an online health promotion program. This data set presents a network of people that share their physical activities and see others' activity levels. It is a data set specifically focusing on health promotion, in contrast with other research which is mostly using online social networks for general purposes, such as Facebook, Twitter, etc. [8, 15].

To build this data set, the participants wore an activity monitor device that tracks their physical activity level (PAL). They also had access to an online system where they could befriend other participants in order to share and see each others' PAL. The data sample used in this work was collected from 28/04/2010 until 30/03/2011. The analysis of the characteristics of this social network in a health promotion context provides a basis for answering the following questions:

1. How does the largest component of this specific social network develop over time?
2. Does this social network demonstrate the homophily phenomenon (concerning gender and BMI)?
3. Can we use the dynamic analysis of the network to determine influential nodes?

The paper is organized as follows. Section 2 discusses the dynamic aspects of social networks, and presents the concepts explored here. Section 3 explains the analysis performed, metrics used and the selection process. Section 4 shows the results of the analyses. Finally, Section 5 concludes the paper with a discussion of the consequences and the possible applications of the findings.

2 Dynamical Social Network Analysis

The dynamic aspects of social networks can be analyzed in two ways: (1) looking at the changes *inside* the network (changes in the nodes' attributes as opinions, beliefs, etc.), or (2) looking at the changes of the network itself (the topology of the network, the nodes' degrees, etc.). Dynamical networks are considered here as social networks where the topology changes over time due to new connections or new subjects inside the network.

Static measures of nodes' degrees, centrality, shortest paths, etc. of one fixed snapshot of the data are not sufficient to understand real networks that evolve over time. How new connections are made in or removed from the social network can to some extent be explained by these two phenomenons: homophily and preferential attachment ('more becomes more') [4, 14]. These concepts will be explored further in this work.

The dataset that we use is also used in [10]. In their work, the authors explore the internal states of the nodes and the correlations between the characteristics of the nodes for a shorter period (14 weeks). In [13], the same data set is the basis for a study on the differences between people inside and outside a community, showing how the community aspect plays a role in changing the physical activity level during an intervention. The current work is dedicated to the topological and structural aspects of the network and its connections over time.

3 Methods

This section explains the data collection and the data processing. The aim is to provide a clear understanding of how the data was collected, how the subset was selected and how the analysis was done.

3.1 Data Set and Data Selection

The data set is the result of an online physical activity promotion program, where the participants wore an activity monitor that tracks their physical activity level (PAL). The devices were synchronized with an online system, which also provided the possibility for them to join a community through connection requests. The participants could also participate in a health promotion program, and those who decided to do that were tagged in our data set with a 'start plan date'. The data used in this work spans 336 days, from 28/04/2010 until 30/03/2011.

As the decision to join the community was optional for the participants, around 10% of them decided to join the social network to exchange their information about the PAL tracked by their devices. In total there are almost 5,000 nodes that opted to join the online community at some moment during the experiment.

Due to changes in the system, some cleaning was necessary to keep the data set reliable for the analyses performed. From the originally 5,000 nodes and around 28,000 edges, we filtered nodes and edges according to the following characteristics:

a) Nodes without 'start plan date' were removed;
b) Nodes were included according to the date of their started plan;
c) Nodes that dropped out the experiment (tagged with a value for 'dropout date') were taken off at the day when they quit the network and the program;
d) Nodes without a value for BMI (Body Mass Index), gender and nodes in which all information was missing were taken out;
e) Edges without 'start date' value were removed;
f) Edges connected to excluded nodes were removed.

From a total of 28,418 edges, 3,802 edges didn't have information about the date of connection, because some requests for connections in the network were not approved from the receiving peer. As these edges are represented in two directions, 1,901 unique edges were discarded. From the 24,616 edges left, 12,047 are duplicated edges, i.e., node A connects to B, but the edge (B,A) already exists. As all connections are bidirectional, this is redundant data. So we have, in the end, a total of 12,569 edges representing connections that were formed during the experiment.

The data set originally contained 4,989 nodes. Of those, 1,614 nodes were not eligible because they do not have values for all the attributes needed for the analysis (i.e., gender, BMI and start plan date). The selected data set has 3,375 nodes left.

The nodes are only included in the network in the period between the start plan and the drop out date (for those that dropped out). After the node leaves the network, all its connections are deleted also. The impacts of the cleaning process are irrelevant, because the nodes and edges removed didn't participated in the program as demanded.

3.2 Social Network Analysis

The network measures that are calculated are [19]: (1) degree distribution; (2) average degree; (3) closeness centrality; (4) eigenvector centrality; (5) betweenness centrality; and (6) average shortest path . These aspects were analyzed for each day of the experiment.

Formula 1 shows the calculation for the **combined centrality**, a combination of the betweenness and closeness values:

$$Comb_C(i) = \frac{C_C(i) + C_B(i)}{2} \tag{1}$$

$C_C(i)$ and $C_B(i)$ are the closeness and betweenness centralities, respectively. This formula doesn't consider the balance between the two centrality measurements, and might be improved for future analysis. For our analysis it is correct to say that the Closeness centrality will influence more than the betweenness for having higher values in general.

Homophily is the tendency of nodes to create strong connections with others that are alike, have the same opinions, or share similar characteristics [14]. The homophily principle can be studied in two ways: the *social* homophily and the *value* homophily [12, 20]. In this work, the *social* aspects (gender and BMI) are studied in depth, while the *value* aspects are left out of the analysis.

The homophily according to gender was calculated using the gender of the nodes' edges. These edges were categorized as follows:

Edge MM (EMM): a connection between two male nodes;
Edge MF (EMF): a connection between a male node and a female node;
Edge FF (EFF): a connection between two female nodes.

As the three categories are disjoint, the total number of edges equals to $EMM + EMF + EFF$. The homophily for female gender and male gender are given by equations 2 and 3, respectively.

$$Homophily_F = \frac{EFF}{EFF + EMF} \tag{2}$$

$$Homophily_M = \frac{EMM}{EMM + EMF} \tag{3}$$

To calculate homophily for the BMI, we considered nodes with BMI in the same range as equals. Two different thresholds were used: 5.0 and 6.5, which are the respective ranges for the group of Normal and Overweight BMI in the categorization according to [18].

The ratio between the nodes' edges with a small difference in BMI and the total number of edges yields the percentage that follows the homophily principle for the BMI. The equations follow the same principles of equations 2 and 3.

The **ego-network density** for the nodes is used to find important nodes. The density is calculated in two steps. First, the ego-network of all the nodes (including the observed node) is created using 1-step neighborhood. After this step, the density of the ego-network was calculated as: Ego-density $= \frac{|E|}{n(n-1)}$, where $|E|$ is the number of edges in this subgraph, and n is the number nodes.

4 Results

This section presents the results obtained from the social network analysis. The section is organized according to the questions from Section 1:

1. How does the largest component of this social network develop over time?
2. Does this social network demonstrate the homophily phenomenon (for gender and BMI) ?
3. Can we use the dynamic analysis of the network to determine influential nodes?

4.1 Nodes, edges and degree distribution

On day 98 of the experiment the number of nodes in the graph is stabilized at 2,996. The number of nodes in the largest component increases until the end of the experiment, due to new connections established among the nodes.

For the edges there is also a point of stabilization in the new connections around day 100. From that day onward there is a very small increase in the number of connections (around 8.2%). Most of the edges are in the largest component, as it is expected in a network that follows the Small World Network model.

The graph follows a Power-law distribution for the degrees of the nodes for all time steps. Figure 1 shows the degree distribution for the days 1, 100 and 336 in a log-log scale (for illustration[1]). The lower graphics show the coefficients for the linear regression of the correlation between the degree of the nodes and the number of nodes with certain degree.

As shown in the lower graphic, the p value is always significant for our data set, and the R-squared is close to 1, showing that the model explains very well the data, mainly after day 100.

The ‘more becomes more’ principle is the assumption that nodes with higher degree have a higher chance of receiving more connections over time [16]. Figure 2 shows how the degrees of the nodes with the fewest connections (the ‘poorest’, right) and nodes with the most connections (‘richest’, left) evolve over time. More investigation is needed to claim that the preferential attachment is observed here, but the information about the rich and poor nodes suggests that it could be present in our data set.

4.2 Largest component and other components

The ‘largest component’ is the biggest connected component among all components of any graph. Figure 3 shows the percentage of the nodes of the graph that are part of the largest component for all time steps in two different scenarios. In the first scenario, all nodes are included in the graph. As can be observed, the average number of nodes in the largest component is 65% after day 296 for the entire graph. The increase in the percentage follows the inclusion of new edges after time 100 (when the number of nodes is stable).

As there are many nodes with degree 0 (isolated nodes), for the second scenario, the nodes with degree 0 were excluded from the graph. In this scenario the percentage of nodes in the largest component goes up to 80%.

[1] The other days and other animations can be seen at `http://www.cs.vu.nl/~efo600/cn2016/`

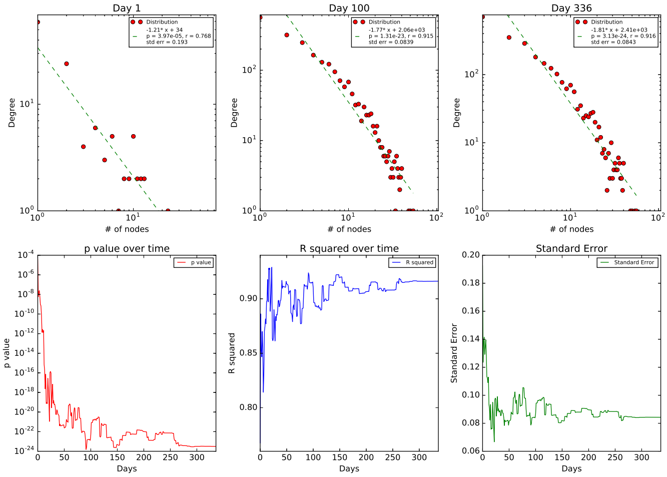


Fig. 1: Degree distribution in days 1, 100 and 336 (top) and *p* value for slope, R squared and standard error (bottom)

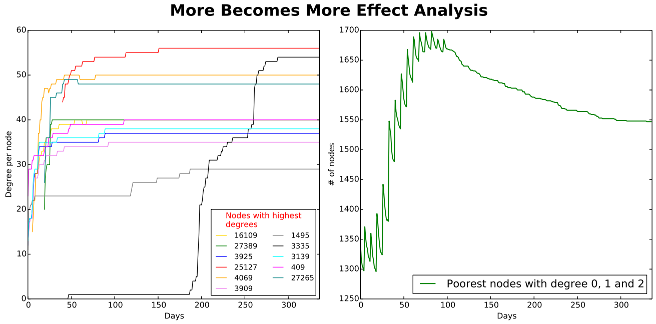


Fig. 2: Degree of the richest nodes (left) and number of poor nodes, with $degree \leq 2$ (right)

Figure 4 shows the evolution of the connected components over time. The upper graphic shows the number of components over time. As edges are inserted, many components are joined, explaining the decrease from around 1,200 connected components to almost 600 in the end. The red line shows the number of components bigger than 1, i.e., non isolated nodes. This number goes from 39 on day 1 up to 164 in the last day of the experiment. The number of isolated nodes goes from 1,193 in day 1 down to 492, what explains the high number of components, even after the largest component gathered more than 60% of the nodes of the network.

The correlation between the size of the components and the number of components with a specific size (frequency of occurrence) is shown in the middle part of Figure 4 in three graphics, for days 1, 165 and 335. The correlation is significant for all time steps. The three lower graphics show the *p* value, R squared and standard error for the regression done in all the time steps of the data set. It can be seen that the fit parameter goes from approximately 65% to less than 40% in the end of

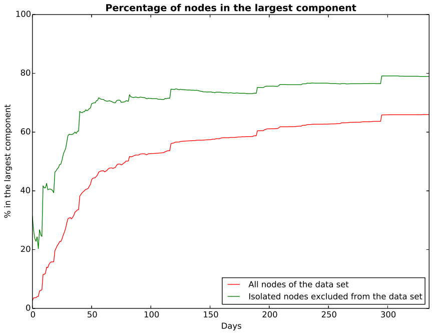


Fig. 3: Percentage of the nodes in the largest component. All nodes (lower red line) and nodes with degree larger than 1 (higher green line)

the experiment. This can be explained by the changes in the largest component, and the joining of previously separated components.

4.3 Centrality measurements

As the largest component has most of the nodes and edges, it is also interesting to explore the centrality measurements for this component. The following metrics were analyzed: (1) betweenness centrality, (2) closeness centrality, (3) eigenvector centrality, (4) average shortest path.

The **betweenness centrality** indicates how important a node is for the transfer of information or any kind of spreadable element inside a network. Nodes with higher betweenness have more shortest paths passing through themselves, and therefore can enhance their role in the network. The **closeness centrality** is the proximity of a node to the rest of the network, and it is calculated by the inverse of the sum of the shortest distances between each node and all other nodes in the network. The **eigenvector centrality** is calculated based on the centrality of its neighbors.

The average centrality for all the nodes (betweenness, closeness and eigenvector) is shown in Figure 5. The first three graphics on the left show all time steps, while the first three graphics on the right provide a zoomed-in version between day 50 and 336.

The lower graphic in Figure 5 shows the average shortest path. The average shortest path for our data set stabilizes around 6.5, a low value as suggested by the theory in [17].

The combined centrality is useful in finding important nodes that combine a good betweenness centrality and closeness centrality. Figure 6 shows the combined centrality for all the nodes with degree higher than 1.

It is possible to highlight the list of nodes with higher centrality (the most potentially influential nodes in the network). Figure 6 shows the most central nodes measures of betweenness, closeness and the combined centrality. As shown in Figure 6, nodes 68593 and 3335 are very important for this data set, as they present the highest values for these measurements.

4.4 Homophily

To investigate homophily according to gender and BMI, the edges were evaluated to determine whether the nodes they connect belong to the same category. The results for the gender analysis follow the equations 2 and 3. The data set has 51.4% of the nodes of gender male, and 48.6%

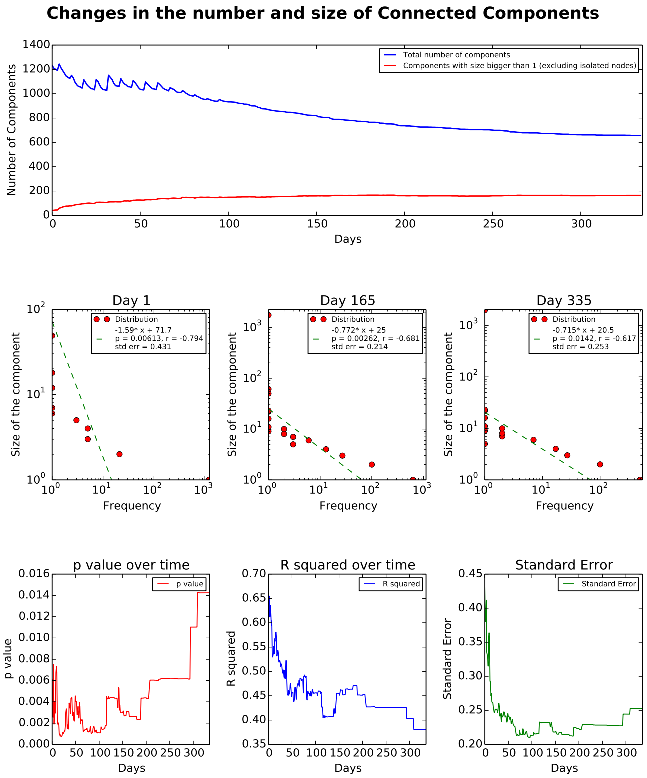


Fig. 4: Components analysis. Number of components in the graph over the time (upper), the correlation between the size of the component and the frequency of the size (days 1, 165 and 335) (middle) and the parameters from the linear regression for all time steps (lower)

female. Regarding the BMI of the population, 0.8% are underweight, 33.6% are normal, 34.8% are overweight and 30.8% are obese [18].

Figure 7 (left) shows the homophily according to the BMI of the nodes. Two ranges were tested for the nodes: 5.0 and 6.5. For the range of 5.0, the ratio of edges with nodes within the same range is around 50% after day 100, while for range 6.5 this value is increased to around 59%. For both ranges, more than half of the connections are within nodes with close BMI.

Figure 7 (right) shows the homophily according to gender. Three calculations were made: (a) edges connecting male-male nodes, (b) edges connecting female-female nodes and (c) edges connecting same gender nodes (male-male plus female-female edges). In this data set, the homophily for women holds for between 50% and 60% of the edges. That means that women connect around half of the time with other women.

For men we observe that more than 60% of the connections are to nodes of the other gender, female. The fact that women have more connections among themselves is know by other studies on gender and relationships [7]. However, the figure also shows that homophily is not present for the male-male connection (i.e., new connections of men are more often with women). When taking both categories together, there is homophily on gender: above 60% of the edges connecting people of the same gender.

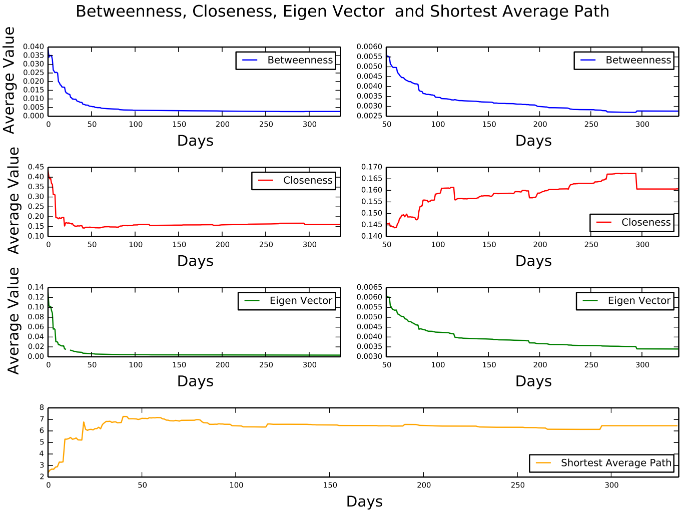


Fig. 5: Mean of all centrality measures for all nodes at each time step (six graphics on top) and average shortest path (bottom)

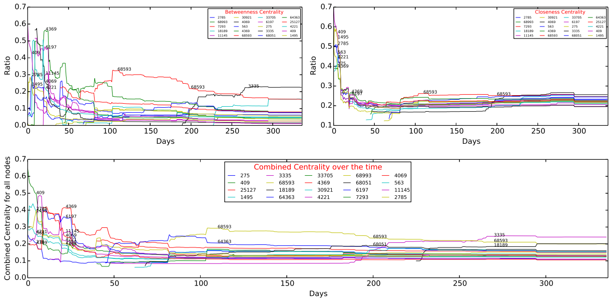


Fig. 6: Centrality for most central nodes. Betweenness (upper left) and closeness (upper right) for the 20 nodes with the highest combined centrality. Lower graphic shows the combined centrality measures

4.5 Identifying influential participants

The dynamic nature of the network is clearly visible from the analyses presented in the previous sections. In previous work we have shown that the more successful participants in the program are, the smaller is the density of their ego-network [10]. This section demonstrates that the set of most influential participants dynamically changes over time. We identify influential participants by comparing properties such as betweenness centrality, closeness centrality, eigenvector centrality, ego-network density and average shortest path.

Figure 8 shows the relation between the node degree of each participant and their ego-network density for the first and last day of the experiment. In this graph we're interested in nodes that have a low density yet a growing degree, as they can be bridges on spreading of emotions, for instance. These are the participants in the top-left quadrant of the graph. Despite the fact that this is just a

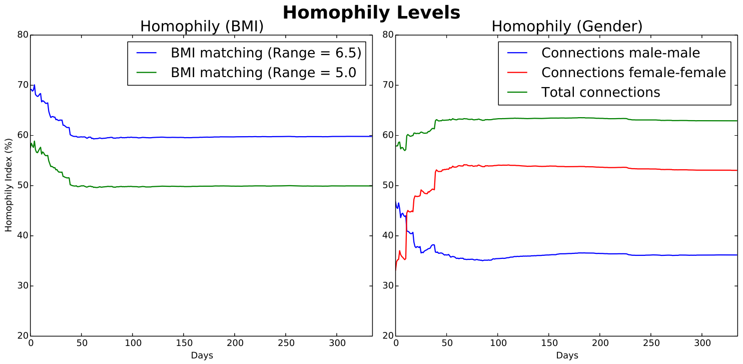
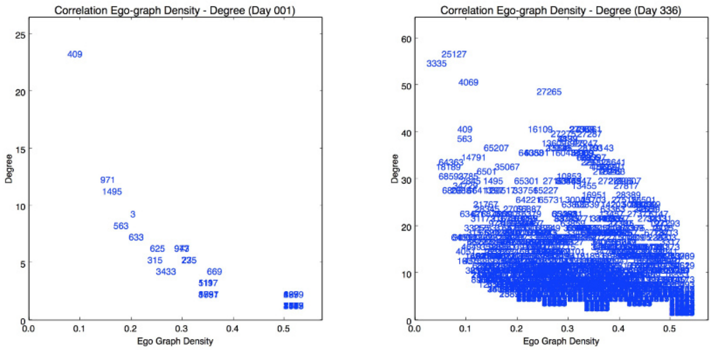


Fig. 7: Homophily according to the BMI (left) and gender (right) of the nodes



(a) First Day

(b) Last Day

Fig. 8: The dynamic relation between ego network density and nodes' degrees

snapshot, the changes over time provided by the combination of each day's relation can give a better picture of what is happening inside a network.

We plotted graphs for all days of the dynamic network which revealed that the set of nodes that emerges in the top-left quadrant are frequently changing. During the experiment, four leader nodes were in evidence considering the ratio between the degree of the nodes and the ego-network density. Node 409 (from day 1 to 12), node 3069 (from day 13 to 40), node 25127 (from day 41 to 254) and node 3335 (from day 255 to 336).

5 Conclusions

In this paper, we have investigated the *dynamic* properties of a longitudinal study of a networked community participating in an online health promotion program. It turned out that studying the dynamics gives additional insights in characteristics of the network. For example, it is shown that the number of components in the network is decreasing while the size of the components is increasing at the same time. The components themselves follow a Power-law distribution at all time steps: there are a few components with many nodes, and a lot of components with only a few nodes. It is also shown that characteristics like betweenness, closeness, eigen vector and average shortest path at the start of the network are very different from the values after 356 days; however it turned out that already after 50 to 100 days most measurements were relatively stable.

The dynamical data set also allowed us to evaluate whether two well-known phenomena of evolving networks are present: homophily and preferential attachment. Our analysis showed that homophily takes place on the aspect BMI and gender; the latter especially for female-female connections. Apart from the possible preferential attachment, more investigation is needed to affirm that it is present in this data set.

Finally, the combination of degree measurements and the density of the ego-network was presented, and we aim to use it to identify people that are potentially influential in their network in further work. Interestingly, the set of people who are influential according to this metric changes during the evolution of the network, even after the moment that the nodes of network have stabilized. This suggest that continuous monitoring the evolution of a network is important to identify such people.

We believe our discoveries and methods can form the basis for automated (health) interventions that exploit the social network for changing behaviours of individuals, and possibly lead us to future discoveries about leadership, spreading of emotions or any other application related to the network's topology and dynamics.

Acknowledgements E.F.M. Araujo's funding is provide by the Brazilian Science without Borders Program, through the Coordination for the Improvement of Higher Education Personnel, CAPES (reference 13538-13-6).

References

[1] Acemoglu, D., Ozdaglar, A.: Opinion dynamics and learning in social networks. Dynamic Games and Applications **1**(1), 3–49 (2011)

[2] Acemoglu, D., Ozdaglar, A., ParandehGheibi, A.: Spread of (mis) information in social networks. Games and Economic Behavior **70**, 194–227 (2010)

[3] Araújo, E.F.M., Tran, A.V.T.T., Mollee, J.S., Klein, M.C.A.: Analysis and evaluation of social contagion of physical activity in a group of young adults. In: ACM International Conference Proceeding Series, vol. 07-09-Ocob (2015)

[4] Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(October), 509–512 (1999)

[5] Blankendaal, R., Parinussa, S., Treur, J.: A temporal-causal modelling approach to integrated contagion and network change in social networks. In: Proceedings of the 22nd European Conference on Artificial Intelligence, ECAI16 (2016)

[6] Christakis, N.a., Fowler, J.H.: The spread of obesity in a large social network over 32 years. The New England journal of medicine **357**(4), 370–9 (2007)

[7] Duck, S., Wright, P.H.: Reexamining gender differences in same-gender friendships: A close look at two kinds of data. Sex Roles **28**(11-12), 709–727 (1993)

[8] Ellison, N.B., Steinfield, C., Lampe, C.: The benefits of facebook ”friends:” Social capital and college students' use of online social network sites. Journal of Computer-Mediated Communication **12**(4), 1143–1168 (2007)

[9] Eubank, S., Guclu, H., Kumar, V.S., Marathe, M.V., Srinivasan, A., Toroczkai, Z., Wang, N.: Modelling disease outbreaks in realistic urban social networks. Nature **429**(6988), 180–184 (2004)

[10] Groenewegen, M., Stoyanov, D., Deichmann, D., van Halteren, A.: Connecting with active people matters: the influence of an online community on physical activity behavior. In: International Conference on Social Informatics, pp. 96–109. Springer (2012)

[11] Kempe, D., Kleinberg, J., Tardos, É.: Influential Nodes in a Diffusion Model for Social Networks. Automata, Languages and Programming **3580**, 1127–1138 (2005)

[12] Lazarsfeld, P.F., Merton, R.K.: Friendship as a Social Process: A Substantive and Methodological analysis. Freedom and Control in Modern Society **18**, 18–66 (1954)

[13] Manzoor, A., Mollee, J.S., Araújo, E.F., van Halteren, A.T., Klein, M.C.A.: Online sharing of physical activity: does it accelerate the impact of a health promotion program? In: Socialcom 2016 (2016)

[14] Mcpherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology **27**(1), 415–444 (2001)

[15] Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement - IMC '07 pp. 29–42 (2007)

[16] Newman, M.E.J.: The structure and function of complex networks. Siam Review **45**(2), 167–256 (2003)

[17] Newman, M.E.J., Watts, D.J.: Scaling and percolation in the small-world network model. Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics **60**(6 Pt B), 7332–7342 (1999)

[18] Organization, W.H., et al.: Global database on body mass index: an interactive surveillance tool for monitoring nutrition transition. World Health Organization: Geneva (2012)

[19] Scott, J.: Social Network Analysis. Sage (2012)

[20] Tsvetovat, M., Kouznetsov, A.: Social Network Analysis for Startups: Finding connections on the social web. " O'Reilly Media, Inc." (2011)

[21] Valente, T.W.: Network models of the diffusion of innovations, vol. 2. Hampton Press (NJ) (1995)

[22] Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**(6684), 440–2 (1998)