

## OPTIMIZATION OF THE INFORMATIONAL FLOW IN A SOCIAL NETWORK – A PROTEIN NETWORK-BASED APPROACH

RĂZVAN BOCU<sup>(1)</sup> AND DORIN BOCU<sup>(2)</sup>

**ABSTRACT.** This paper introduces an integrated inter-personal communication quality assurance method that is based on the study of protein networks. The method assesses each individual in the social network on two dimensions, considering the analysis of the biological network as a model. The procedure allows for an informative and comprehensive analysis of the social networks to be conducted, at various levels of complexity.

### 1. INTRODUCTION

The aim of this paper is to present a novel two-dimensional social data analysis method that is based on the experience that we have accumulated during the previous phases of our research activity in relation to networks of proteins [1]. These are very large networks that can be compared to the social networks from a structural perspective. Moreover, it can be stated that social networks are smaller as compared to the protein networks, considering their size [2]. Therefore, any optimal approach in relation to protein networks is bound to ensure an adequate analysis of the social networks, both considering the execution times and the accuracy of the results that are obtained. The core of the procedure is based on the computation of each individual's importance, with the help of betweenness centrality, and on the determination of the communities (clusters of individuals).

The quality of the interpersonal informational channels greatly affects the operation of the social network seen as a whole [3]. In this respect, this paper demonstrates that in the case of problematic social networks, malfunction is related to the most important individuals in the network and, as a consequence, the normal operation of the overall structure is greatly disturbed.

---

Received by the editors: March 25, 2011.

2010 *Mathematics Subject Classification.* 91C20.

1998 *CR Categories and Descriptors.* H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *Information networks*.

*Key words and phrases.* Interactome networks, protein-protein interactions, protein communities, social network, protein importance, analysis method.

An accurate understanding of the structure and importance of social networks requires the usage of efficient analysis techniques [4]. The efficiency of this procedure is demonstrated through a case study that takes into account a large social structure.

**1.1. Remarks Regarding the Social Networks and the Importance of Individuals.** Betweenness is a centrality measure that is based on the shortest path computation, and is widely used in the complex networks analysis [1]. It deals with one of the main problems in network analysis that supposes the precise assessment of the importance (or the centrality) of a particular vertex (or an edge) in a network, at the scale of the whole network [5].

Let us also recall that we have extensively used the concept of centrality in a series of research endeavours that belong to the area of bioinformatics through the study of protein networks, considering a computer scientist's perspective. Thus, it can be stated that the use of centrality is an important instrument for the proper analysis of various networked structures that determine the nowadays world. The myriad of social networks that co-ordinate all human interactions are among the most important ones. The following section describes the analysis method through a case study that analyzes a medium-sized social network that is represented by the supporters of an Irish football team, which is constituted by 13,738 supporters and 693,075 social links. The group of supporters is divided into informal clusters that are autonomous. The history of this social group records certain periods of time when various problems arose, such as unsatisfactory attendance at the stadium. We took over the task to analyze the causes of these problems through the analysis method that is described. First, we chose a few supporters that appear to occupy the position of community liaison. They have all been analyzed following the directions of the method. One of the supporters, let us conventionally call him *S1*, constitutes the main subject of the case study that assesses the effectiveness of the analysis method. Following, the presentation contains references to concepts like the Dijkstra-adapted betweenness computation algorithm and the flag-based community detection algorithm [1]. They represent contributions that have been produced by the protein network-related part of our research, and due to typographic space constraints they are not presented in this paper. Nevertheless, they are properly referentiated, and the interested reader can look up for additional information in relation to them.

## 2. THE ANALYSIS METHOD

The supporter *S1* has been analyzed through the following steps, which actually conform to the structure of the social data analysis method conceived by the research presented in this paper:

- Considering the social data set, the Dijkstra-based adapted algorithm [1] was run in order to determine the absolute importance of the supporter at the scale of the whole social network.
- Following, the functionally related clusters of supporters have been determined using the flag-based community detection algorithm [1].
- Consequently, the sub-communities that express (contain) the supporter  $S1$  have been isolated.
- Taking into account the centrality scores of the supporters that are components of the sub-communities determined, their overall importance has been calculated.
- For the next step, real-world information about the individual in question has been used, in order to precisely link the topological features of the  $S1$ -determined sub-communities to the information about their role at the scale of the whole group of supporters.

**2.1. Description of the Analysis Process.** The practical analysis that was performed as part of the research pathway presented strictly followed the procedure described in the previous section. Let us note that betweenness has been normalized relative to all the practical assessments and analyses that were performed. As a consequence, the betweenness is defined in this context as a function  $C_B : GOS \rightarrow [0, 1]$ . Here,  $GOS$  is an acronym that comes from *Group Of Supporters*, and designates the set of all the individuals that are contained in the social network. In other words, this convention means that, in a particular case that supposes all the social links pass through a certain supporter, its betweenness value can be maximum 1. Following, the output generated by each step of the analysis process is presented.

**2.1.1. Betweenness Computation.** The social data set has been processed considering the Dijkstra-adapted sequential betweenness algorithm as the main routine. In order to ensure perfect accuracy of the results, the output was compared to the results produced by the parallel versions of Brandes algorithm and Dijkstra-adapted algorithm [1]. The supporters have been classified in three categories, based on their betweenness values. Thus, supporters that feature a betweenness that is less or equal than 0.3 are considered to be characterized by a low importance, supporters that feature a betweenness that is less or equal than 0.6 but greater than 0.3 are considered to be characterized by a medium importance, while supporters whose betweenness is greater than 0.6 are referred to as important and having a significant influence on the social network as a whole.

The supporter  $S1$  is ranked as an *important entity*, with a betweenness value of 0.843. The result is produced by the Dijkstra-adapted betweenness

algorithm, and is confirmed by the two parallel betweenness computation algorithms, Brandes and Dijkstra-adapted.

*2.1.2. Community Detection.* The community structure of the social network has been determined using the flag-based community detection algorithm [1]. Let us recall that modularity is used throughout this research in order to assess the quality of the determined community structure. Consequently, it is used as a stop criterion for the community detection algorithms that were implemented. Considering the suggestions found in the literature and this research's results, it is possible to ensure the algorithm is stopped when the social network is partitioned into meaningful communities, situation that corresponds to a modularity value a little bit greater than 0.8, in the case of social networks, which is another important similarity as compared to protein networks. This threshold is lowered to 0.4 in the case of other practically-useful networks, such as road networks or telecommunications networks. Thus, the algorithm does not generate unnecessary iterations, and the accuracy of the community detection is assured.

Furthermore, the overall importance of sub-communities that contain the supporter  $S1$  has been assessed. The results proved that he belongs only to globally central sub-communities (with betweenness values in the range [0.679..0.842]), which significantly influence the informational flow in the social network.

The community isolation step generated suggestions about several other supporters bearing key roles in the overall group. It is interesting to note that the empirical data confirms the importance of the individual supporters.

As a consequence, the analysis method is not only useful for the study of individual persons, it also suggests other persons that co-operate in order to fulfill the same social function. It is equally important to note that, while it is interesting to confirm the existence and the role of *already proven* social liaison individuals, it is especially useful to detect *possible candidates* that may have a significant influence at the scale of the overall social group, as this is bound to allow for a faster and more effective identification and correction of the issue that may affect the communication and the informational flows in the social network.

*2.1.3. Mapping Experiments' Output to Real World Data.* The last phase of our research involved checking the accuracy of the results that have been obtained through an empirical approach. Thus, the role of the supporter  $S1$  in the overall group of supporters has been carefully investigated over a period of time. Consequently, it has been discovered that any failure of  $S1$  to fulfill his normal community liaison tasks provoked a significant disorganization of the group he is a member of. As an example, when some health problems

kept  $S1$  away from the usual daily activities, most of the supporters that belong to the same group gave up attending the matches of their favourite team. Furthermore, this has generated problems at the scale of the whole social network, as the group whose leader  $S1$  is acts as a globally central one. Thus, the informational flow between some other groups has been affected.

**2.2. Conclusive Remarks on the Case Study.** The main aim of the case study that has been described was to demonstrate the usefulness and effectiveness of the social data analysis method. Thus, it can be stated that the goal has been reached, as the outcome of the case study can be summarized as follows:

- The supporter  $S1$  is an important entity at the scale of the entire social network (group of supporters).
- The social communities it is a member of are globally central and, thus, any problem that prevents him from acting normally significantly affects the informational flow social network wide.
- The link that exists between issues that may prevent  $S1$  from interacting normally with the rest of the group and informational flow problems at the scale of the social network has been established with the aid of empirical data.
- The social data analysis method is also able to detect related important members of the social network, apart from the one that is the central entity of the analysis. Moreover, it is important to note that the output of the analysis offers suggestions about individuals that seem to regulate the same social network, but that have not yet been accepted as *de facto* holders of a key role social network wide.

The same analysis method has been applied in relation to other important members of the social network and found to successfully produce relevant results considering all instances. Furthermore, it can be stated that any member of any social network can be analyzed as per this method, provided the following requirements are met: the individual has to be properly defined in a social data set, and a proper empirical investigation on the actual social network has to be performed, in order to allow for the precise social role of the individual to be established.

### 3. CONCLUSIONS AND FUTURE DEVELOPMENTS

The social data analysis method is able to process social networks in a comprehensive and accurate way considering the specific information provided by both individual members' importance assessment and communities detection components. The resulting analysis method allows us to explore social

networks in a more informative way than is possible by just making use of traditional analysis techniques. It allows us to distinguish between central and peripheral hubs of highly connecting community members, revealing individuals that form the backbone of the social network. The fact that we observe an enrichment of members that influence the informational links in this group and also their highest betweenness centrality values indicates the central role of these individuals.

It is important to note that while the method is suitable for the assessment of individual members, it is also extremely useful for discovering other important social network members that have not yet been recognised as such. As a consequence, relevant sociologists' efforts can significantly benefit from using the analysis procedure.

As an additional important remark, it can be stated that the successful re-utilization of the analysis technique both for protein networks and social networks, already suggests that it should be suitable for any other networked structure that exists in the contemporary world.

The next stages of our research will involve further optimizations of the algorithms that form the backbone of the analysis method. Additionally, we intend to analyze even more biological and social data sets and, possibly, expand its usage to other types of networked data.

#### REFERENCES

- [1] R. Bocu and S. Tabirca, *The Flag-based Algorithm - A Novel Greedy Method that Optimizes Protein Communities Detection*: International Journal of Computers, Communications and Control, 6(1):33-44, 2011.
- [2] W.W. Zachary, *An information flow model for conflict and fission in small groups*: Journal of Anthropological Research 33, 452-473, 1977.
- [3] M. Sales-Pardo, R. Guimera, A.A. Moreira and L.A.N Amaral, *Extracting the hierarchical organization of complex systems*: PNAS September 25, 2007 vol. 104 no. 39 15224-15229, 2007.
- [4] C. Song, S. Havlin and H.A. Makse, *Self-similarity of complex networks*: Nature 433, 392-395, 2007.
- [5] L.C. Freeman, *A set of measures of centrality based on betweenness*: Sociometry, Vol. 40, 35-41, 1977.

<sup>(1)</sup>TRANSILVANIA UNIVERSITY OF BRASOV, DEPARTMENT OF COMPUTER SCIENCE, 50 IULIU MANIU STREET, BRASOV, ROMANIA  
*E-mail address: razvan.bocu@unitbv.ro*

<sup>(2)</sup>TRANSILVANIA UNIVERSITY OF BRASOV, DEPARTMENT OF COMPUTER SCIENCE, 50 IULIU MANIU STREET, BRASOV, ROMANIA  
*E-mail address: d.bocu@unitbv.ro*