# Topological properties of the core group in online communities

M. R. Martinez-Torres, M. C. Diaz-Fernandez
Business and Management department
University of Seville
Seville, Spain
rmtorres@us.es, cardiaz@us.es

S. L. Toral
E. S. Ingenieros
University of Seville
Seville, Spain
storal@us.es

***Abstract*— Online communities are self-organized networks with a clear core-periphery structure, where most of the contributions are due to a small group of users known as the core group. Although several methods for detecting the core-periphery structure have been proposed, all of them assume a pre-defined structure of the core group. However, online communities exhibit a wide variety of organizations and shapes, and the patterns of behavior of the core group is continuously changing. This paper investigates the relationship between the global parameters of the community and those of the core group. Findings reveal that the behavior of the core group determines the global structure of the community and therefore, the identification of the core group should consider the global characteristics of the network in which they are contributing.**

***Keywords—core group; social network analysis, topological properties, Open source software, online communities***

## I. INTRODUCTION

Online communities have aroused a great interest in the research community due to their ability for promoting collective intelligence and innovation models [1], [2], [3]. However, there is still some ambiguity in the definition of online communities, depending on the people or subject considered [4]. It is commonly accepted that a virtual community is integrated by a group of people who may or not meet one another face-to-face and that exchanges words and ideas through the computers and networks mediation [5]. From a social perspective, Internet-based technologies facilitate the construction of personal relationships, creating weak links among geographically disperse individuals who regularly participate in the online community [6]. Online communities have also being considered as virtual organizations, which is a form of cooperation between companies, institutions and/or individuals sharing common interests and aims [7].

Although online communities represent a social phenomenon, they also exhibit a social structure. Actually, online communities are self-organized structures of people, with their own hierarchy that emerges as the community evolves [8], [9]. In this line, communities can be modeled as a graph, with the nodes representing community members and the arcs representing their interactions [10]. Such representation has led to numerous insights within online communities field, borrowing some ideas from natural and information sciences as well as from complex network models [11]. A core-periphery structure has been identified in many online communities [12]. This structure is based on the participation inequality, typical of many online communities, and the Legitimate Peripheral Participation, that has been described as one of the basic processes that support the sustainability of communities. Participation inequality means that only a small fraction of the community members is responsible of the majority of contributions [13], leading to a well known network model known as scale-free networks, where the contributions of users follow a power law distribution [14]. The Legitimate Peripheral Participation is the process by which newcomers can become full members or even experts by learning from more competent practitioners [15]. Therefore, the core group of the community needs to be continuously renewed with new members as not all of them are going to be part of the community during its lifetime. The identification of the core group is actually an open issue in the research about online communities. This paper is focused on the identification of the core group of online communities using several topological properties of nodes. The aim consists of identifying which topological properties can be the best predictors of the core group depending on the global characteristics of the network. This information if of great interest to monitor the evolution of communities and to check how the Legitimate Peripheral Participation is working and nurturing the core group.

The rest of the paper is structured as follows: next section reviews the previous works about online communities and their core-periphery structure. Section III introduces the case study based on a open source software community and the methodology, including the topological characteristics of nodes to be considered. Section IV described the obtained results and, finally, section V concludes the paper.

## II. RELATED WORK

The open access to online communities attracts numerous members to them, although with different degree of involvement. As a result, online communities self-organize attending to three basic categories of members [16], [17]. The core group represent the most prolific group of users responsible of guiding and coordinating the development of the community. They are usually involved with the community during a long period of time making significant contributions. Moderators and experts are included in this group. The second category is the active members group, that make regular contributions but are not engaged with the development of the community. Finally, the last category is given by the group of peripheral members, that only make irregular and occasional contributions, and that are only engaged with the community during short and sporadic periods of time.

The relevance of the role played by core members in online communities have been analyzed in several previous studies [2], [7]. Findings emphasize the mediation activity that must be developed by the core group, especially in the case of active members [6]. Core members are not only responsible of the majority of contributions but they also have to promote the participation among other group members [18]. Several approaches have been proposed to examine the core-periphery structure in a network such as block models [19], k-core organization [20], the connectivity of information and short paths through a network [21], and communities overlapping [22].

Despite of the previous mentioned approaches, the most popular notion of core-periphery structure in networks was developed by Borgatti and Everett [19], who proposed algorithms for detecting the core-periphery structure in weighted, undirected graphs. Their approach to the core-periphery structure is based on comparing a network to a block model that consists of a fully-connected core and a periphery that has no internal edges but is fully connected to the core. Latterly, other researchers had partially criticized Borgatti and Everett´s core-periphery structure method. For instance, their approach assume a pre-defined structure of the core group (fully-connected) and it don't consider the possibility of multiple cores. Less restrictive approaches entail identifying densely-connected core nodes and sparsely-connected periphery nodes [12]. In contrast to the Borgatti and Everett approach, the nodes in a core are also reasonably well-connected to those in the periphery allowing their new method of computing core-periphery structure to identify multiple cores in a network considering different possible core structures.

The main limitations of previous methods is that all of them assume certain patterns of connections between the core group an between the core and peripheral members. However, the reality is that online community members connections exhibit a wide variety of shapes and densities. Although they typically follow a power law distribution, the exponent of the power law can be very different from one network to another. Our assumption is that the topological properties of the core group vary depending on the global characteristics of the network they belong to. This study is therefore focused on the identification of the local topological properties of the core group in order to check if they are influenced by the features of the whole network. For this purpose, the first required step is obtaining the list of the core group. This is actually a challenging step: the previous methods only provides an biased estimation of the core group, and they don't consider the characteristics of the network in their estimation. The only possible solution is obtaining a ground truth, which means making use of experts to analyze online communities and to obtain the real core group. As this is a time consuming task, this study is focused on a small open source software community, mush easier to be manually analyzed.

## III. Case Study and Methodology

The case study is based on open source software communities, which constitute a successful example of the application of online communities to the development of software using collective intelligence. This section introduces open source communities and the methodology based on social network analysis.

### A. *Open Source Software communities*

Open Source Software (OSS) projects represent a model of software creation based on the contributions of developers and users geographically dispersed but connected together through shared values and the Internet [23]. As a difference to proprietary software, the source code is available and can be adapted, modified or improved by the public [24]. Many previous studies have been focused on the structure of OSS communities [25], [7]. They don't have a flat structure and different degrees of involvement are allowed. For instance, developers need users to inform their practices, and users need developers to implement their requests. More specifically, it has been concluded that OSS communities follow a core-periphery structure, with a core group of active developers in the inner layer of the structure and less active users as long as we move to the outer layers [26]. The identification of the core group of developers is a major issue for the survival of the community, as they have a direct incidence in its successful development [27]. This study is focused on the Debian Project, which is an association of individuals who have made common cause to create a free operating system called Debian GNU/Linux, or simply Debian for short [28]. More specifically, the study analyzes Debian port to ARM mailing lists, which can be publicly accessible at https://lists.debian.org/debian-arm/. This website include people interactions since 1999, and they are organized as threads of discussion. Mailing lists are useful to put in contact information seekers and information providers, and they are a very useful resource for those who need to adapt Debian to a specific processor.

### B. *Modelling communities as Social Networks*

Online communities can be modeled as social networks by representing them as a graph, where nodes identify users posting messages and arcs represent their interactions. When modeling communities, it is important to decide the meaning of interactions. For instance, in the case of the threads of discussion, threads are initiated by one user posting a message, usually a question, and then subsequent messages answering this initial question are aggregated below. The decision in this case is how to establish the links among nodes. In general, it is cognitively more complex to answer a thread of discussion than to answer a single message. Answering a thread of discussion usually requires to read all the previous content in the thread to write a coherent answer [29]. Following this idea, an author posting to a thread was tied to all the authors who have previously posted to the same thread when constructing the social network. As a result, a directed and valued graph was obtained. The direction of the arc is given by the flow of information between two authors, and the value of the arc represent the number of times an author is answering messages in the same or in a different thread.

### *C. Topological features of networks*

Several local and global topological characteristics of social networks can be extracted once they are modeled as a graph. Local characteristics refers to the individual properties of nodes within the network, while global characteristics refers to those metrics related to the network as a whole. In this paper, local characteristics were used as predictors of the core group of developers and global characteristics were used to determine to what extent the core group has an incidence over the successful development of the community. Table 1 and Table 2 details the considered local and global features of online communities, respectively.

| Measure | Description |
|---|---|
| Out-degree | Number of arcs sent by a node |
| Betweenness centrality | Measure of centrality given by the intermediary role developed by a node. |
| Closenesss centrality | Measure of centrality given by the distance to the rest of nodes of the network |
| Modularity | It is based on edge betweenness. Edges connecting communities will have high edge betweenness |
| Hub | Heavily linked nodes in a network. They are defined as the principal eigenvector of $AA^T$, where A is the adjacency matrix of the graph |

**Table 1. Local topological characteristics of nodes.**

| Measure | Description |
|---|---|
| Size | Number of nodes of the network |
| Density | Number of arcs of the network |
| Average degree | Average value of the out-degree of nodes in the network |
| ASP | Average shortest path |
| Diameter | Length of the longest path in the network |
| Modularity | Community structure detection based on edge betweenness |
| $\alpha$ coefficient | Exponent of the power law distribution typical of scale free networks |

**Table 2. Global topological characteristics of the network.**

Local features include the out-degree of nodes, which represent their activity, two different measures of centrality, based on their intermediary role and their distance to the rest of the community, and two measures related to the connectivity of nodes. Modularity based on edge betweenness is a community structure detection method based on the idea that many networks consist of subnetworks which are densely connected themselves but sparsely connected to other subnetworks, as stated in some related previous works. Finally, hubs represent those nodes of particularly high out-degree but connected to those other nodes with particularly high in-degree. The notion of hubs takes up the idea of a core group connected with the periphery, as stated by Rombach et al. (2014).

The global characteristics of the network consider several general issues like size, density or the average degree of nodes. The average shortest path and the diameter of the network provide information about the cohesion of the network. Finally, modularity and the $\alpha$ coefficient refers to the community structure. The $\alpha$ coefficient was estimated according to the goodness-of-fit based method described in [30].

## IV. RESULTS

The Debian port to ARM mailing lists were extracted and modeled as a graph using a specific crawler developed in R. Each month of the last 13 years was considered as a separate network, leading to a total of 156 separate networks. As an example, Figure 1 shows the network corresponding to December, 2012. Each node represents an author that posted a message during this month, and the size of nodes was represented proportional to its out-degree. The out-degree is one of the local topological properties considered for the identification of the core of the community.

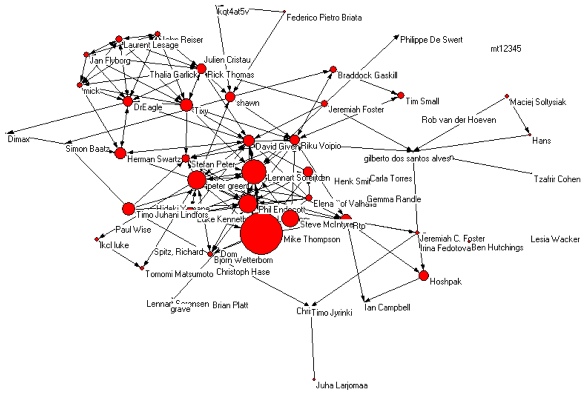


**Figure 1. Mailing list network of Debian-ARM corresponding to December, 2012.**

A panel of experts analyzed each of them, establishing those nodes that can be considered as the core group attending to their level of participation as well as to the content of posted messages. This annotation represent the ground truth that was used as the dependent variable in this study. The independent variable are the local properties of nodes of Table 1. All of them have been calculated using the package 'igraph' from R.

A binary logistic regression was then performed for each of the 156 networks, obtaining for each case the best predictor of the condition of being part of the core group (ones in the dependent variable). As the core group usually represent a small percentage of the community, the dependent variable has a high number of zeros (non core members). This kind of problems where the dependent variable have a high number of zeros are known as zero inflated problems, and they can be solved considering a negative binomial distribution [31].

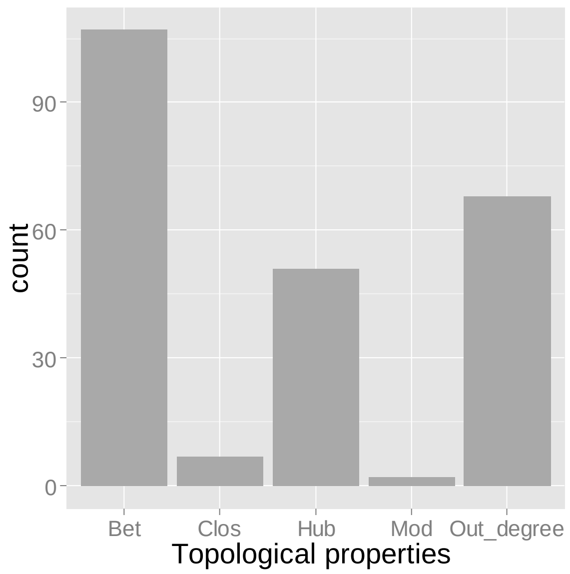


**Figure 2. Topological properties of the core group.**

Figure 2 shows the best predictor of the core group for the 156 networks considered. For each logistic regression, the classification rate (between core and non core members) of each independent variable is calculated, and the one achieving the best prediction results is considered as the best predictor of the core group. There are cases where two or even three of the independent variables achieve the same classification results. All of them are then considered as best predictors. The average classification rate was 0.87. Obtained results show that three of the independent variables of Table 1 concentrate most of the best predictions: Betweenness centrality, hub and out-degree. It is interesting to notice that closeness centrality and modularity are not good predictors. This fact can be explained because obtained networks are dense networks where distances between nodes are short and where it is not easy to detect sub communities. However, as it can be visualized in Figure 1, the topology of the network shows significant differences in the out-degree of nodes. Additionally, these networks tend to behave as scale free networks, which are characterized by the presence of hubs. Finally, betweenness is a different metric of centrality based on the intermediary role of nodes.

These three variables were used to cluster the initial group of 156 networks in three subgroups. The 'Bet' group is integrated by those networks where the best predictor was the betweenness centrality. The 'Hub' and 'Out-degree' group contains those networks were the best predictors were hub and out-degree, respectively.

| | **'Bet'** | **'Hub'** | **'Out-degree'** |
|---|---|---|---|
| Size | 50.07 | 57.66 | 56.00 |
| Density | 843.08 | 1001.16 | 954.53 |
| Av. degree | 33.07 | 32.66 | 33.07 |
| ASP | 2.24 | 2.45 | 2.24 |
| Diameter | 5.00 | 6.50 | 5.00 |
| Modularity | 0.28 | 0.38 | 0.25 |
| $\alpha$ coefficient | 1.90 | 2.13 | 1.77 |

**Table 3. Global characteristics of the three subgroup of networks.**

Table 3 details the global characteristics of the three subgroups of networks. The main difference is appreciated for the value of the $\alpha$ coefficient. The property of being a hub works better as a core predictor when the value of $\alpha$ is high, which means a core group concentrating a high number of interactions and a network with a clear scale-free behavior. This result is also supported by a higher value of modularity and density. The 'hub' group of networks have a dense core group also connected with peripheral nodes. In contrast, the out-degree networks are just the opposite: networks with a light scale-free behavior, and less densely connected. Finally, the 'bet' group has intermediate properties respect to the other two groups. The rest of global characteristics of networks doesn't show clear differences in the three groups of networks, so they don't have incidence over the behavior of the core group. In summary, the structure of networks and the patterns of activity of the core group are related to each other. OSS communities are self organized networks and the core group activity determines not only the structure of the network, but also drive the successful development of the underlying project, as stated in [27]. The 'bet' and 'out-degree' group of networks are in line with the Borgatti and Everett´s core-periphery method, while the 'hub' group consider the idea proposed in [12] about a core connected to the periphery. Consequently, the identification of the core group should consider the global structure under which they are developing its activity.

## V. CONCLUSIONS

This paper have analyzed the structure of online communities and their relationships with the core group of developers. For this purpose, a ground truth of core developers was manually obtained for the proposed case study, and then used as the dependent variable. Obtained results show that the best predictors of the core group also identify several global characteristics of the networks they belong to. As a result, methods for identifying the core group should not only consider one pattern of behavior but different patterns depending on the global structure of the general network.

## ACKNOWLEDGMENT

This work was supported by the Consejería de Economía, Innovación, Ciencia y Empleo under the Research Project with reference P12-SEJ-328 and by the Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad under the Research Project with reference ECO2013-43856-R.

## REFERENCES

[1] E. von Hippel, G. von Krogh, “Open source software and the ‘‘private-collective” innovation model: issues for organization science”, Organization Science, Vol. 14, no. 2, pp. 209–223, 2003.

[2] F. Rullani and S. Haefliger, “The periphery on stage: The intra-organizational dynamics in online communities of creation”, Research Policy, vol. 42, pp. 941– 953, 2013.

[3] M. Sarma and A. Lam, “Knowledge creation and innovation in the virtual community? Exploring structure, values and identity in Hacker Groups, Paper to be presented at the 35th DRUID Celebration Conference, Barcelona, Spain, June, 17th-19th, 2013

[4] J. Preece, “Sociability and usability in online communities: determining and measuring success”, Behabiour & Information Technology, vol. 20, no. 5, pp. 347-356, 2001.

[5] H. Rheingold, “A slice of life in my virtual community. In L.M. Harasim (ed.), Global Networks: Computers and Intenational Communication (Cambridge, MA: MIT Press), pp. 57-80, 1994.

[6] S.L. Toral, M.R. Martínez-Torres and F. Barrero, “Analysis of virtual communities supporting OSS projects using social network analysis”, Information and Software Technology, vol. 52, pp. 296–303, 2010.

[7] M. R. Martinez-Torres, M. C. Diaz-Fernandez, "Current issues and research trends on open-source software communities", Technology Analysis & Strategic Management, Vol. 26, Iss. 1, pp. 55-68, 2014.

[8] H. Kautz, B. Selman, M. Shah, “Referral Web: combining social networks and collaborative filtering”, Communications of ACM, vol. 40, no. 3, pp. 27–36, 1997.

[9] P. Raghavan, “Social networks: from the web to the enterprise”, IEEE Internet Computing, vol. 6, no. 1,pp. 91–94, 2002.

[10] M. R. Martinez-Torres, " Application of evolutionary computation techniques for the identification of innovators in open innovation communities", Expert Systems With Applications, Vol. 40, Iss. 7, pp. 2503-2510, 2013.

[11] A.-L. Barabási, Taming complexity, Nature Phys., vol.1 pp. 68–70, 2005.

[12] M.P. Rombach, M.A. Porter, J.H. Fowler and P.J. Mucha, “Core-periphery structure in networks”, Slam J. Appl. Math, vol. 74, no. 1, pp. 187-190, 2014.

[13] G. Kuk, "Strategic Interaction and Knowledge Sharing in the KDE Developer Mailing List", Management Science, Vol. 52, Iss. 7, pp. 1031–1042, 2006.

[14] A. L. Barabási, The Physics of the Web, Physics World, Vol. 14, Iss. 7, pp. 33-38, 2001.

[15] C. Kimble, P. Hildreth, and P. Wright, Communities of practice: Going virtual. In Hildreth, Paul M. and Kimble, Chris, editors, Knowledge Networks: Innovation through Communities of Practice, Idea Group Publishing, 220–234, 2000.

[16] A. Mockus, T. Fielding, and D. Herbsleb, Two Case Studies of Open Source Software Development: Apache and Mozilla, ACM Trans. Software Eng. and Methodology, Vol. 11, no. 3, pp. 309-346, 2002.

[17] J. Xu, Y. Gao, S. Christley, and G. Madey, A Topological Analysis of the Open Source Software Development Community, Proceedings of the 38th Annual Hawaii International Conference on System Sciences, HICSS '05, 188-198, 2005.

[18] M. R. Martinez-Torres, "Analysis of activity in open-source communities using social network analysis techniques", Asian Journal of Technology Innovation, Vol. 22, Iss. 1, pp. 114-130, 2014.

[19] S. P. Borgatti and M. G. Everett, Models of core/periphery structures, Social Networks, Vol. 21, pp. 375–395, 1999.

[20] D. F. Gleich, PageRank, software package, available online at http://www.mathworks.com/matlabcentral/fileexchange/11613-pagerank, 2006.

[21] K. T. Poole, Voteview, software and data sets on political leanings of congressional representatives, available online at http://voteview.com (2014).

[22] D. A. Smith and D. R. White, Structure and dynamics of the global economy: Network analysis of international trade, Social Forces, Vol. 70, pp. 857–893, 1992.

[23] J. Herbsleb, A. Mockus, "An empirical study of speed and communication in globally distributed software development", IEEE Transactions on Software Engineering, Vol. 29, pp. 481–494, 2003.

[24] S. L. Toral, M. R. Martinez-Torres, F. Barrero, " Modelling Mailing List Behaviour in Open Source Projects: the Case of ARM Embedded Linux", Journal of Universal Computer Science, Vol. 15, Iss. 3, pp. 648-664, 2009

[25] G. von Krogh, E. von Hippel, "The promise of research on open source software", Management Science, Vol. 52, pp. 975–983, 2006.

[26] M. R. Martinez-Torres, "A genetic search of patterns of behaviour in OSS communities", Expert Systems With Applications, Vol. 39, Iss. 18, pp. 13182-13192, 2012.

[27] S. L. Toral, M. R. Martinez-Torres, F. Barrero, F. Cortes, "An empirical study of the driving forces behind online communities", Internet Research, Vol. 19, Iss. 4, pp. 378-392, 2009

[28] J. Mateos-Garcia & W. E. Steinmueller, “The institutions of open source software: Examining the Debian community”, Information Economics and Policy, Vol. 20, pp. 333–344, 2008.

[29] N. Knock, “Compensatory adaptation to a lean medium: An action research investigation of electronic communication in process involvement groups”, IEEE Trans. on Professional Communication, Vol. 44, no. 4, pp. 267–285, 2001.

[30] A. Clauset, C. R. Shalizi, M. E. J. Newman, "Power-law distributions in empirical data", SIAM Review, Vol. 51, pp. 661-703, 2007.

[31] J. Hinde, and C. Demetrio, "Overdispersion: Models and Estimation", Computational Statistics and Data Analysis, Vol. 27, pp. 151-170, 1998.