Community detection in a large social dataset of European Projects *

Sergi Lozano , Jordi Duch and Alex Arenas †

Abstract

We present a methodology to analyze large social data sets based on a new community detection algorithm. As a main advantage, we stress that community division makes easier the operation of crossing relational data (who is connected to whom) with particular information about each person or organization. As an example, we analyze the database of research projects of the European 6th Framework Programme.

1 Introduction

Link analysis, and other techniques based on a structural approximation, has demonstrated its utility for the analysis of all kind of data sets that can be represented as complex networks. Nevertheless, the increasing size of these available databases reduces the possibilities of these techniques to statistical measures of network properties. Additionally, it is quite common that these social data correspond to affiliation of people to an organization or a project (like a movie or a research paper, for instance). Network representation of this sort of data are extremely dense, making even more difficult to extract useful information from relational databases.

One example of these sort of large dataset, is about projects involve in the European Union Sixth Framework Programme (here referred, from now on, as FP6). More concretely, the available data consists on a list of members of all projects and some information about each particular organization. Since the main commitment of this Programme is to encourage collaboration between research organizations, any kind of information about collaboration patterns and dynamics that could be obtained is specially valuable.

One possible procedure to ahead the treatment of this and other similar databases, is determining the quantity and characteristics of communities inside the network. In the following sections, we look deeply on community detection and present a practical example using the FP6 projects database.

*Supported by Spanish Government Grant BFM 2003-08258-C02.

†Departament d'Enginyeria Informtica i Matemtiques. Universitat Rovira i Virgili.

2 Community structure and its determination in complex networks

The general notion of community structure in complex networks was first pointed out in the physics literature by Girvan and Newman [1], and refers to the fact that nodes in many real networks appear to group in subgraphs in which the density of internal connections is larger than the connections with the rest of nodes in the network.

In our particular case of the FP6 projects network, for example, this means that the more projects two organizations collaborate in, the more probable both of them belong to the same community. Consequently, the community structure of our network can reveal information about collaboration strategies, alliances and so on.

The problem of community detection is quite challenging and has been the subject of discussion in various disciplines. All existing methods intended to devise the community structure in complex networks, require a definition of community that imposes the limit up to which a group should be considered a community. However, the concept of community itself is qualitative: nodes must be more connected within its community than with the rest of the network. Some quantitative definitions that came from sociology have been used in recent studies [2], but in general, the physics community has widely accepted a recent measure for the strength of the community structure generated by an algorithm. This measure, introduced by Newman and Girvan [3], is based on the concept of modularity Q :

$$Q = \sum_r (e_{rr} - a_r^2) \tag{2.1}$$

where e_{rr} are the fraction of links that connect two nodes inside the community r, a_r the fraction of links that have one or both vertices inside of the community r, and the sum extends to all communities r in a given network.

3 The method

In this experiment, we have used a divisive algorithm based on the Extremal Optimization (EO) heuristics [4], for a detailed description of the method see [5]. This algorithm operates, basically, optimizing a global variable

as a result of co-evolutionary avalanches generated by the improvement of extremal local variables.

More concretely, the global variable to optimize is the modularity Q as defined in eq.(2.1). Thus, the definition of the local variables used in the extremal optimization problem should be related to the contribution of individual nodes i to the summation in eq.(2.1) for a given distribution into communities

$$q_i = \kappa_{r(i)} - k_i a_{r(i)} \tag{3.2}$$

where $\kappa_{r(i)}$ is the number of links that a node i belonging to a community r has with nodes into the same community, and k_i is the degree of node i. Note that $Q = \frac{1}{2L}\sum_i q_i$ where i refers to all nodes in the network given a certain partition into communities and L is the total number of links in the network. Eq.(3.2) provides a measure that depends on the node degree, and its normalization involve all the links in the network after summation. Re-scaling the local variable q_i by the degree of node i, we obtain a proper definition for the contribution of node i to Q, relative to its own degree and normalized in the interval [-1,1].

$$\lambda_i = \frac{q_i}{k_i} = \frac{\kappa_{r(i)}}{k_i} - a_{r(i)} \tag{3.3}$$

Keeping in mind this definition of λ_i we can compare the relative contribution of individual nodes to the community structure. We consider λ_i as the local variable involved in the extremal optimization process that characterizes an individual node. From now on we will refer to λ_i as the fitness of node i, using the common jargon in extremal optimization problems.

To find heuristically the optimal modularity value, the proposed algorithm follows these steps:

- First, the whole graph split in two random partitions having the same number of nodes each one. This splitting creates an initial communities division, where communities are understood as connected components in each partition.

- At each time step, the system self-organizes by moving the node with the lower fitness (extremal) from one partition to the other. In principle, each movement implies the recalculation of the fitness of many nodes because the right hand side of equation (3.3) involves the pseudo-global magnitude $a_{r(i)}$.

- The process is repeated until an "optimal state" with a maximum value of Q is reached. After that, all the links between both partitions are deleted and the previous step is proceed recursively with every resultant connected component.

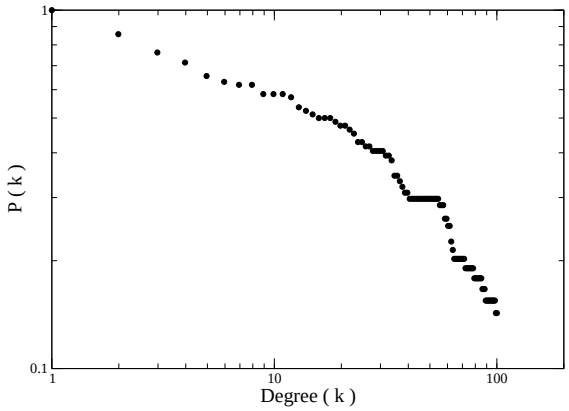

Figure 1: Cumulative degree distribution of the giant component of the FP6 network in log-log scale.

- At a certain moment, more subdivisions into communities necessarily decrease Q, and the process finishes. [1].

The cost of the algorithm is $O(n^2 log^2 n)$ where n is the number of nodes of the network, see [5]. Note that this process is not a bipartitioning of the graph as known in computer science [6], because: the number of nodes in each partition is dependent on the evolution process and not restricted to be the same at the end of the process; and more importantly, each partition could contain different connected components (communities) that when the partitions are disconnected result in several subgraphs. For an exhaustive comparative between the method proposed here and other existent methods in the literature see [7].

4 Application to unravel the structure of collaborations in FP6

We have run the presented community detection algorithm with data from the network of FP6. We focus our attention on organizations that collaborate al least in two projects, we have used only the giant component of the network (that is, the biggest one of the connected components of the network), leaving outside isolated groups. this giant component has 3030 nodes and 63964 links. The degree distribution is clearly far from

[1] The value of Q always refers to the whole network i.e. is the sum over all the communities. At a certain moment more subdivisions into communities will necessarily decrease Q because the limit of decomposition is a community per node whose value of Q is negative.

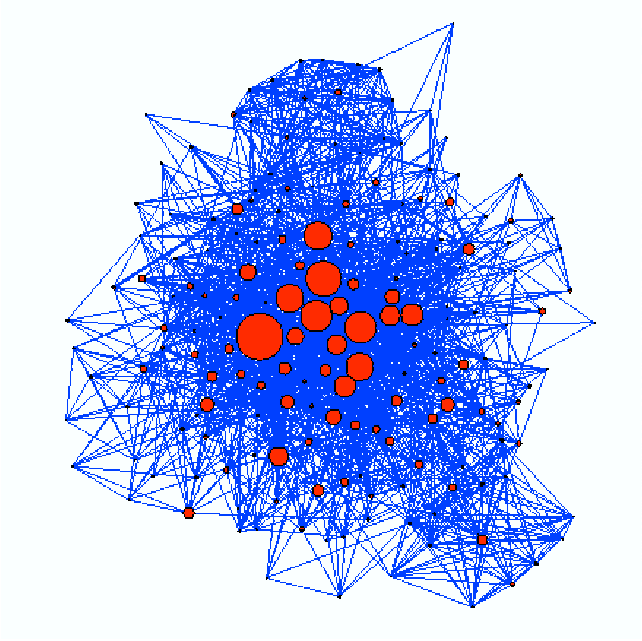

Figure 2: Community structure represented as a network. Nodes correspond to communities and link represent collaboration between members of the two connected communities. Diameter of nodes and width of links symbolize community size and number of crossed collaboration, respectively.

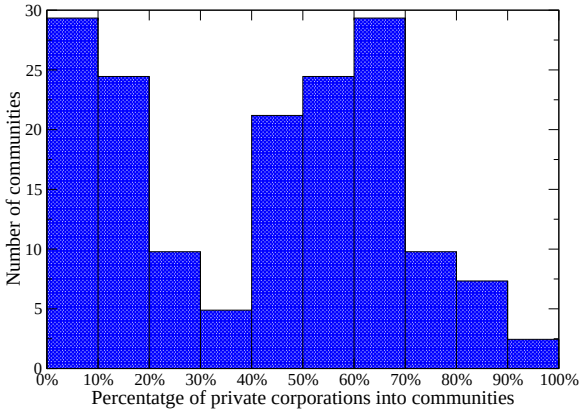

Figure 3: Number of communities as a function of the percentage of participation of private corporations in joint projects.

this of a scale-free network, see Figure 1.

The direct output of the algorithm is a list of organizations grouped in 163 communities. Figure 2 is a graphical representation of the obtained data. Each node corresponds to a community and a link between two nodes denotes that, at least, one member of each community have collaborated once. In addition, dimensions of nodes and links give information about community size and number of crossed collaboration, respectively. These indicator can reveal important information about, for example, collaboration strategies, since a small size of a community would imply that its members have not joint many projects, or that they have preferred having a little number of known partners.

The methodology we propose consist into crossing these communities composition profile with particular data of each organization, to obtain the structural information of the network.

One option is to employ information about organization's activity (research, government, business). We find an example in two communities listed in Table 1. Although both of them are composed by organizations working in electronics and telecommunications, they have completely different organization type patterns. The first one is a combination of public and private research centers, companies and regional governments organizations industries. On the contrary, members of the second one are exclusively from the industrial world and even, some of them, of the same firm groups. In Figure 3, we plot the number of communities that have joint projects with different levels of participation of private corporations. We observe that are approximately the same number of communities with none or low private participation as with large private participation.

From another point of view, configuration of some other communities reveals alliances between complementary products and services providers. Community shown in table 2 is a clear case of this assumption, we can find car builders, constructors of automobile parts, tech institutes, public organizations related with mobility and public transports, local governments and others.

Finally, nationality composition of communities could also be studied, crossing our community distribution with data about the country of each organization. Concretely, useful information about collaboration between firms and research centers from different European countries and other geographical areas like America or Asia. Community in table 3 represents an example of collaboration between organizations from Europe and Asia.

5 Conclusions

Summarizing, in this paper we have presented a methodology to analyze relational aspects of large

Community 3
Motorola Ltd
Institute for Infocomm Research
Siemens Aktiengesellschaft.
Thales Communications SA.
Kings College London.
Telecom Italia Learning Services SPA.
Universitaet Karlsruhe
Alcatel CIT
Swedish Institute of Computer Science AB
Nec Europe Ltd
European Telecommunications Standards Inst.
Telia Sonera AbPubl
The University of Surrey
France Telecom SA
Nokia Corporation
Siemens Mobile Communications SpA
Consorzio Ferrara Ricerche
Community 138
Hispasat
Alcatel Espacio S.A.
Ems Satcom Uk Ltd
Nera Broadband Satellite As
Shiron Satellite Communications Ltd
Sistemas y Redes Telematicas SL
Indra Espacio SA
Telemar
Telefonica Pesquisa e Des.do Brasil Ltda

Table 1: Two communities with a different activity patern. While community number 3 is composed by a rich variety of organizations dedicated to research, government and business, community number 138 includes exclusively companies

databases based on a new community detection algorithm. Beyond a purely macroscopic perspective of the whole network (provided by observables like network diameter or degree distribution), its division into communities facilitates the crossing of relational data (who is preferably linked to whom) with particular information about each node.

As an illustrative application in social sciences, we have built up a network from a database of research projects of the European 6th Framework Programme, calculated its community structure and analyzed the resulting data by crossing it with information about organization's type of activity, market and nationality. The results reveal different strategies of alliances in the European FP6 scenario. While some communities are clearly devoted to make a bridge between public research institutions in different countries and industries, other are formed exclusively by one of them unraveling pure scien-

Community 19
Centre Suisse d'ectronique et Microtechnique
EADS Deutschland Corporate Research Center
Lunds Universitet
Skoda Auto AS
Volkswagen Ag
Robert Bosch Gmbh
Technische Universitat Darmstadt
System Design and Research Association SRL
European Road Transport Telematics Organisation
Audi Aktiengesellschaft
Bayrische Motoren Werke Aktingesellschaft
Bmw Forschung und Technik Gmbh
Seat Centro Tecnico
Volvo Car Corporation
Blaupunkt Gmbh
Delphi Delco Electronics Europe Gmbh
Faurecia Sieges D'Automobile SA
Ibeo Automobile Sensor Gmbh
Siemens Vdo Automotive Sas
Fcs Simulator Systems
Federal Highway Research Institute
Essex County Council
Landeshaupstadt Hannover
Ministry Economics and Transport of Lower Saxony
Laboratory of Lighting Technology. Darmstadt Univ.

Table 2: An example of community centered in a unique market. Note that all organizations are related, in some sense, with automobiles.

Community 17
Finprory
Satama Interactive Oyj
Samsung Electronics Co Ltd
RWTH Aachen
Advanced Commu. Research and Development SA
Beijing University of Posts and Telecom
Danmarks Tekniske Universtet
Forschungszentrum Telekom Wien Betriebs GmbH
Nokia Corporation Oyj
Shanghai Inst. of Microsystem and Information Tech.
Tata Consultancy Services
VTT
Teliaorasis
University of Rome
Altemo Research Centre
Cefriel

Table 3: Nationality profile of communities is another important question. Community number 17 is an example of mixture of european and asian organizations

tific objectives or well defined product directed research. This analysis could be of interest to the councils devoted to provide worldwide research grants to propose target oriented programs.

The presented methodology opens the door to a deeper analysis of a wide variety of large data sets, not only that ones corresponding to affiliation networks (like the seen FP6 network), but also other sort of networks built up from massive data obtained electronically, like e-mail or WEB sites networks.

References

[1] M. Girvan and M.E.J. Newman. Proc. Natl. Acad. Sci., 99 (2002).

[2] F.Radicchi, C. Castellano, F. Cecconi, V. Loreto and D. Parisi. Proc. Natl. Acad. Sci., 101 (2004), 2658.

[3] M.E.J. Newman and M. Girvan. Phys. Rev. E, 69 (2004), 026113.

[4] S. Boettcher and A. G. Percus, Phys Rev Lett 86, 5211-5214 (2001).

[5] J. Duch and A. Arenas. Phys. Rev. E, 72 (2005), 027104.

[6] S. Boettcher and A.G. Percus. Phys Rev E, 64 (2001), 026114.

[7] L. Danon, A. Diaz-Guilera, J. Duch and A. Arenas, J. Stat. Mech. (2005) P09008