EPJ manuscript No.
(will be inserted by the editor)

Analysis of large social datasets by community detection

Sergi Lozano, Jordi Duch and Àlex Arenas

Departament d'Enginyeria Informàtica i Matemàtiques. Universitat Rovira i Virgili. 43007 Tarragona,Spain

Received: date / Revised version: date

Abstract. Using a database of research projects of the European 6th Framework Programme, 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 determination makes easier the operation of crossing relational data (who is connected to whom) with particular information about each person or organization.

PACS. PACS-key discribing text of that key – PACS-key discribing text of that key

1 Introduction

Analytical techniques based on a structural approximation, have demonstrated their 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 applicability of these techniques to statistical measures of network properties. Additionally, it is quite common that these social data correspond to affiliation of people to organizations or projects (like movies or research projects, 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). In this case, 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.

In this contribution, we use a practical example using the FP6 database to present a procedure, based on the determination of the quantity and characteristics of communities inside the network, to ahead the problem presented above.

2 The original network and its community structure

Our starting point is a large dataset about membership in projects involved in the European Union FP6 (Sixth Framework Programme). From these data, we have built a network where nodes represent organizations and links indicate coincidence in, at least, one project. Like other similar networks obtained from social large datasets, its large size (about 3000 organizations) and high link density, makes hard to extract information about collaboration patterns between organizations in the FP6.

To determinate the community structure of the network, we have used a divisive algorithm based on the Extremal Optimization (EO) heuristics [3], which was proposed in [2]. This algorithm operates, basically, optimizing modularity Q (as defined in eq. 1) by the improvement of extremal local variables (the contribution of individual nodes to the summation). The Modularity Q is a widely accepted measure of the strength of a certain community structure. It was first presented in [1] as:

$$Q = \sum_r (e_{rr} - a_r^2) \tag{1}$$

Where e_{rr} are the fraction of links that connect two nodes inside the community r, and a_r the fraction of links that have one or both vertices inside of the community r.

3 Unraveling collaboration patterns

The main idea of our methodology is simple: Making a mesoscopic analysis by crossing communities composition with particular data of each organization. Three examples of organizations' information that can be useful in this particular case are nationality, type of activity and the market they belong to.

Nationality composition of communities gives information about collaboration between firms and research centers from different European countries and other geographical areas like America or Asia. Community shown in table 1 represents an example of collaboration between organizations from Europe and Asia.

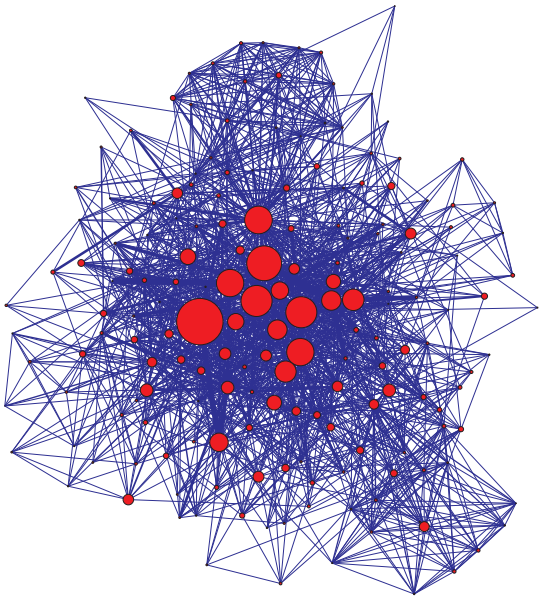

Fig. 1. Resulting community structure represented as a network. Nodes correspond to communities and links represent collaboration between members of the two connected communities. Diameter of nodes and width of links symbolize community size and number of crossed collaboration.

Table 1. Nationality profile of communities. An example of collaboration between european organizations (in blue) and others from Asia (in green).

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

By using data about organizations activity (research, government, business), we can obtain useful information about markets structure (who works with whom). We find an example in Table 2. Although both of the listed communities 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

Table 2. Two communities with different activity patterns. While community number 3 is composed by a rich variety of organizations dedicated to research, government and business, community number 138 includes only companies.

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

the industrial world and even, some of them, belong to the same firm group.

Finally, configuration of some other communities reveals alliances between complementary products and services providers. Community shown in table 3 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.

4 Conclusions and further possibilities

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

Table 3. A complete value chain in a unique community. Here we find a wide variety of organizations related, in some sense, to mobility and transportation.

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.

information about nationality, organization's type of activity and market.

This methodology could be applied to 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.E.J. Newman and M. Girvan, Phys. Rev. E **69**, (2004) 026113.
2. Jordi Duch and Àlex Arenas, Phys. Rev. E. **72**, (2005) 027104.
3. S. Boettcher and A. G. Percus, Phys. Rev. Lett. **86**, (2001) 5211-5214.