Eur. Phys. J. Plus (2012) **127**: 20
DOI 10.1140/epjp/i2012-12020-4

THE EUROPEAN PHYSICAL JOURNAL PLUS

Regular Article

Do regions matter in ALICE?

Social relationships and data exchanges in the Grid

E.D. Widmer[1], G. Viry[1,a], F. Carminati[3], and G. Galli-Carminati[2]

[1] Dpt. of Sociology, University of Geneva, Geneva, Switzerland
[2] Units of Mental Development Psychiatry, Department of Mental Health and Psychiatry, University Hospitals of Geneva Geneva, Switzerland
[3] CERN, Geneva, Switzerland

Received: 27 June 2011 / Revised: 19 December 2011
Published online: 16 February 2012 –

Abstract. This study aims at investigating the impact of regional affiliations of centres on the organisation of collaborations within the Distributed Computing ALICE infrastructure, based on social networks methods. A self-administered questionnaire was sent to all centre managers about support, email interactions and wished collaborations in the infrastructure. Several additional measures, stemming from technical observations were collected, such as bandwidth, data transfers and Internet Round Trip Time (RTT) were also included. Information for 50 centres were considered (about 70% response rate). Empirical analysis shows that despite the centralisation on CERN, the network is highly organised by regions. The results are discussed in the light of policy and efficiency issues.

1 Introduction

This study aims at measuring the impact of regions on the structure of interactions between the centres of the Distributed Computing ALICE infrastructure, based on social network methods [1]. These centres are part of the Worldwide Large hadron collider Computing Grid (WLCG [2]). They form a large computational network where data and workload are exchanged, but also a large and complex social network with $\sim 3\,000$ possible links. The operation experience of this Grid for ten years has shown that the coordination and collaborations between the different centres are as important as the material conditions to ensure the proper functioning of this complex system extended over different time zones and continents.

In a previous article [3], we found that the centres of the ALICE Distributed Computing Infrastructure derive most of their support from CERN and that various types of interactions were centralised on CERN. However, we also found that there were signs of local organisation of data exchange and collaborations. In this paper, we systematically assess the impact of regional influences on collaborations. The research issue considered is the extent to which ALICE is organised with reference to regional anchorages that may be related to cultural, historical, political or network connectivity issues.

2 Data

Information about collaborations were collected using a self-administered questionnaire which was filled by the technical manager of each ALICE centre in the Fall 2009 and the beginning of 2010. The questionnaire was sent via e-mail to the 73 centres of the ALICE Grid. Answers for 50 centres were received and considered in the empirical analysis (68% response rate). Technical managers had to estimate by yes/no questions which centres of the Grid provided their centre with significant help in its work at least once a week (support). They also had to estimate which centres were in e-mail contact at least once a week with their own (interactions), and with which centres their centre would like to have more interactions in its work (wished collaborations). These indicators refer to self-reported exchanges. Ties were binary (dichotomous) (*e.g.*, support from one centre to another exists or does not exist) and directed (*e.g.*, support goes from one centre to another, so that support between two given centres can be absent, mutual or unreciprocated).

[a] e-mail: gil.viry@unige.ch

Table 1. Network indices (with and without CERN).

	Help	E-mail contacts	Wished collaborations	Bandwidth	Data transfers	RTT capacity
With CERN						
Density (%)	4,3	4,9	3,6	19,2	11,7	19,3
Betweenness centralisation (%)	36.2	31.3	26.3	5.7	8.3	21.8
Indegree centralisation (provided) (%)	72.6	59.5	42.1	40.1	15.6	41.1
Outdegree centralisation (providing) (%)	33.1	17.9	13	31.8	38.5	70.2
Number of cliques (min. size 3, sym. max)	24	19	23	48	40	71
Without CERN						
Density (%)	2.2	3.4	2.5	17.8	10.5	17.6
Betweenness centralisation (%)	0.5	3.3	2.4	6.6	7.3	23.4
Indegree centralisation (provided) (%)	12.7	11.5	6	14.6	14.6	35.2
Outdegree centralisation (providing) (%)	10.5	9.3	12.4	40.1	40.1	71.4
Number of cliques (min. size 3, sym. max)	10	9	6	48	28	69

Additional information were gathered by technical observation such as the theoretical capacity of the network linking the centres (bandwidth), the actual quantity of data exchanged and the Internet Round Trip Time (RTT). These indicators refer to observed exchanges, in contrast with self-reported exchanges provided by responses of technical managers in the self-administered on-line questionnaire. Threshold values were fixed to determine only the substantial interactions between the centres. Two centres were considered to be linked when values were included in the two last deciles (20% upper values). Regarding bandwidth and the actual quantity of data exchanged from one centre to the other, this threshold corresponded to an amount of at least 84.7 and 0.064 MB/s, respectively. RTT links were considered high when the Internet Round Trip Time from one centre to the other was below 20.87 ms.

Overall, this study focuses on support, interactions, wished collaborations, theoretical capacity of the links between any two centres (bandwidth), quantity of data exchanged, and Internet Round Trip Time (RTT). We investigate the extent to which those interconnections are structured by regions.

Centres are dispersed in various regions. Overall, the sample includes 32 centres in Europe, 6 in Asia and 8 in Russia. In order to estimate the impact of regions, we could not include Africa (one centre), South America (one centre) and North America (two centres) due to the very limited number of centres pertaining to each of these regions.

3 Measurements

Network analysis focuses on relations between actors rather than on actors' attributes. It aims to identify and interpret the pattern or structure of ties linking interdependent persons or entities. A variety of measures may be used to characterise relational structures [1,4,5]. Some focus on the cohesion of network members, while others aim to assess the degree of inequality in power or resources between network members. In the present study, we use density of the full network and of regions as a measure of overall and regional cohesion. Density in a directed network is equal to the number of existing arcs (directed ties) divided by the total number of possible arcs. In order to estimate the inequality of prominence of centres in the network, two measures of network centralisation were computed, which capture different conceptual dimensions of centrality of actors within the network. In-degree and out-degree centralisations express the variability (or inequality) among centres in the number of ties pointing to and going from a specific centre. For instance, a network characterised by a small number of centres receiving many direct ties, and a large number of centres receiving few ties, has a strong in-degree centralisation. These measures provide information on the local dimension of centralisation, as a centre may have many connections within a rather isolated subgroup of centres. Quite distinctly, betweenness centralisation measures the variability among centres in the proportion of interactions in the network captured by any centre. The network is said to be centralised if a small number of centres lie between all other centres' chains of relationships. Degree and betweenness centralisation are expressed as a percentage, where 100% is the centralisation of a star network. The theoretical "star network" where there is one centre connected to all the other centres in the network and no other ties is the most centralised network. Finally, the number of cliques

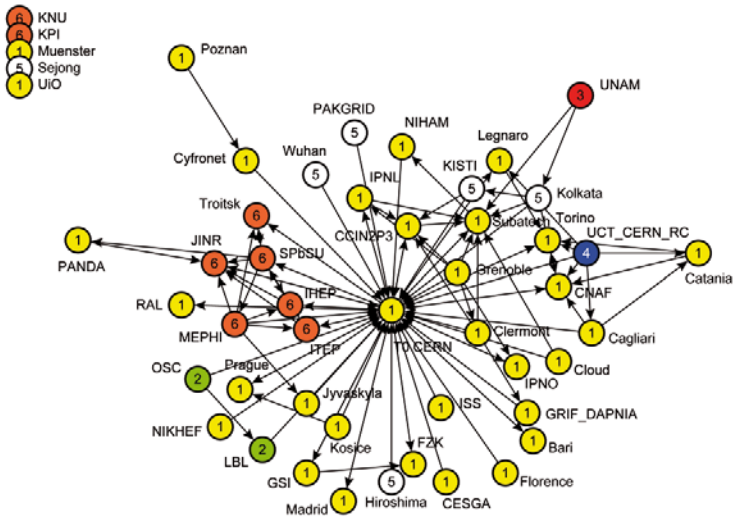

Fig. 1. Help provided by regions.

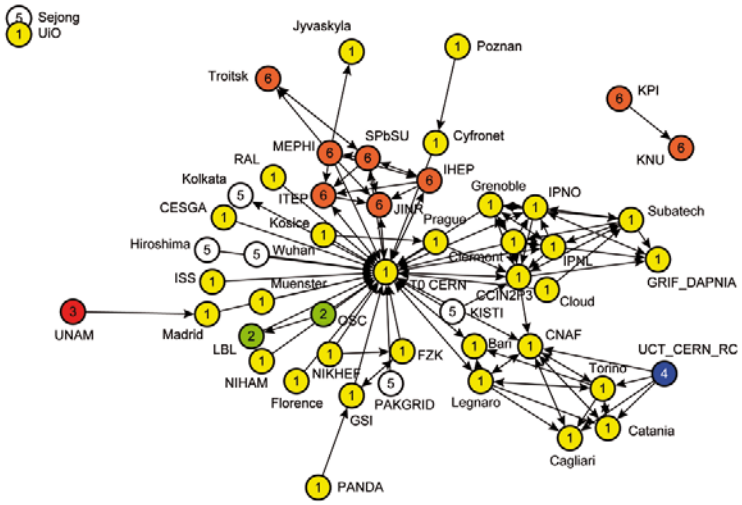

Fig. 2. Exchanges of e-mails by regions.

measures the extent to which the network is structured around multiple clusters of centres or, conversely, is composed of a small number of big clusters. Formally, a clique is a subset of a network including the maximum number of centres that have all possible ties present among themselves. For the present analyses, a clique includes at least three centres. All theses measures are referenced in social network methods [1,4,5].

Using the software UCINET [6], we compute these parameters on the overall network, with and without CERN in it. This two-step procedure was set up in order to control for the impact of CERN on the overall structure, which was acknowledged in a previous publication [3]. It is expected that, since CERN is the largest laboratory, both source of the experimental data and of most of the software used on the WLCG Grid, it has a large influence on the overall structure of the network. Because CERN is by its institutional role lead to be central, it is necessary to estimate regional influences with and without it included in the network. Table 1 presents various indices measuring the cohesion and the centralisation of the network.

4 Results

Results from the overall network, with CERN included, show a highly centralised network and a low level of density for all interactions. When CERN is excluded from the network, the centralisation becomes much weaker. The wished collaborations have a lower level of centralisation, thus showing that centre managers value the development of less centralised interactions within ALICE. For observed exchanges, there is also less centralisation. Getting CERN out of the network does not make such a great difference as in the case of self-reported exchanges.

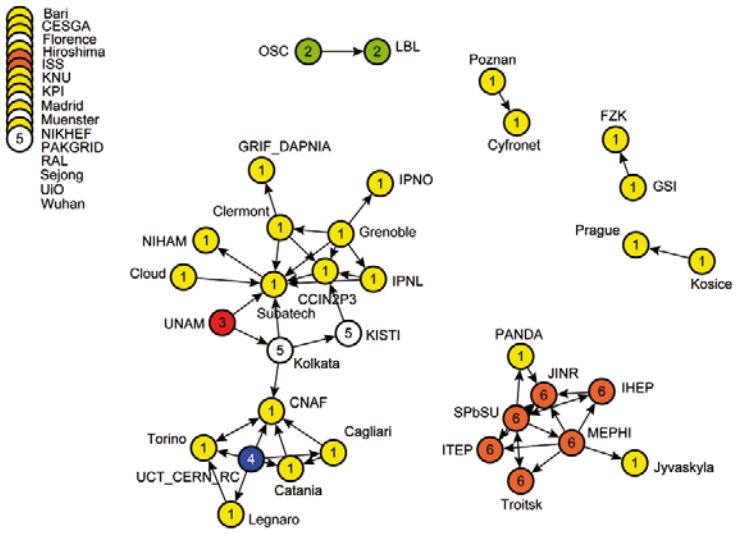

Fig. 3. Help provided by regions without CERN.

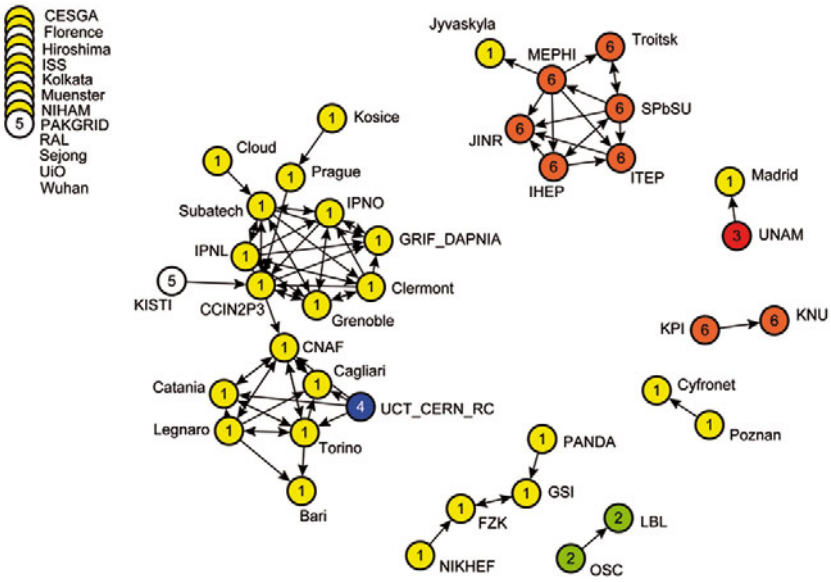

Fig. 4. Exchanges of e-mails by regions without CERN.

Results from the visual inspection of various graphs confirm that interactions, both self-reported and observed, depend to a large extent on regional memberships. Figure 1 shows the help provided. Figure 2 report responses given to the question about contacts by emails. Figures 3 and 4 display the same graphs without CERN. The different regions are illustrated with the following colours: centres from Europe (yellow, $n = 32$), North America (green, $n = 2$), Asia (white, $n = 6$), Russia (orange, $n = 8$), Africa (blue, $n = 1$) and South-America (red, $n = 1$). Both networks show a high level of centralisation on CERN as well as an organisation by regions. Figures 5 and 6 refer to the wishes of centre managers for collaborations, with and without CERN. In that case, the network is fairly less centralised on CERN and the structuring by regions is less straightforward compared with current social interactions.

Were the observed interactions among centres also significantly associated with their regional affiliations? As fig. 7 shows, the theoretical capacity is higher among the centres of the same region than among centres of distinct regions. Also, there are more exchanges among centres belonging to the same regions than among centres of distinct regions (fig. 8), and the Internet round trip time (fig. 9) is in their case lower.

Were those graphical results confirmed by computational analysis? We first present the density within the main regions where most ALICE centres are located. Those regions are Europe, Russia and Asia. Indices for North America, South America and Africa were not computed as the number of countries associated with each regions is very small or even, in some cases, equal to 1. The density within Europe was computed with and without CERN in it.

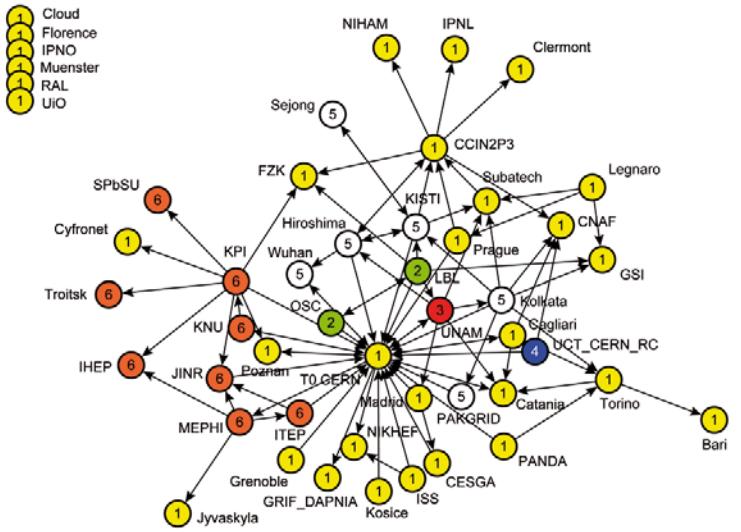

Fig. 5. Wished interactions provided by regions.

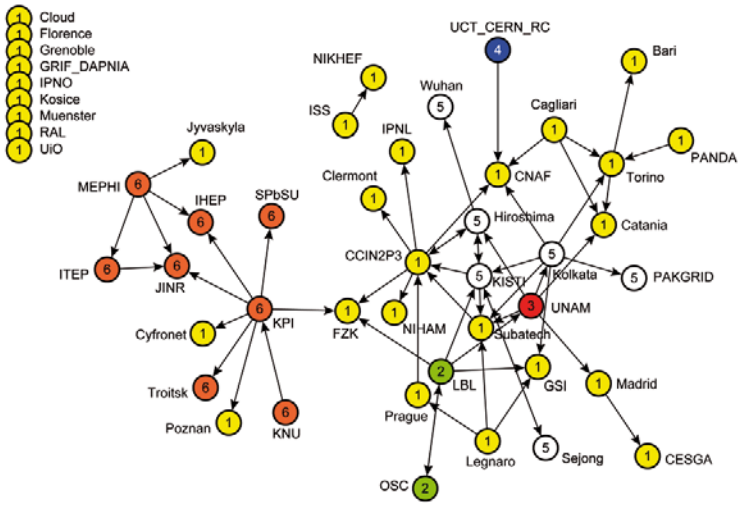

Fig. 6. Wished interactions provided by regions without CERN.

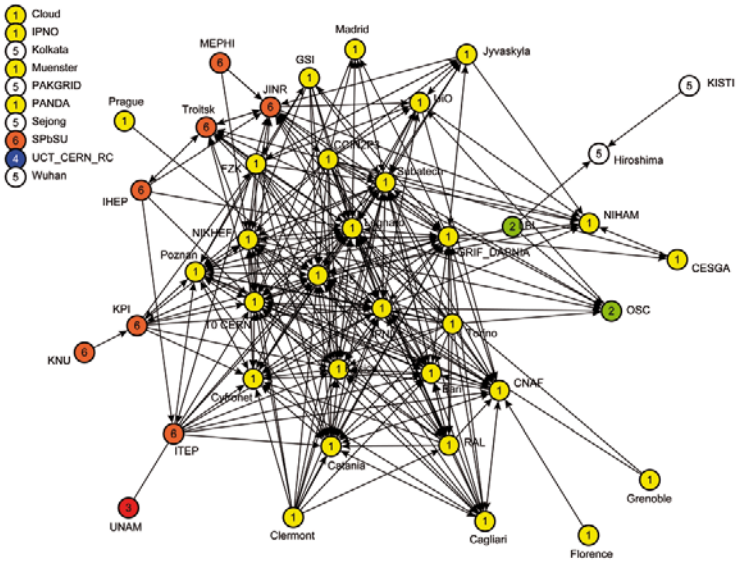

Fig. 7. Bandwidth by regions.

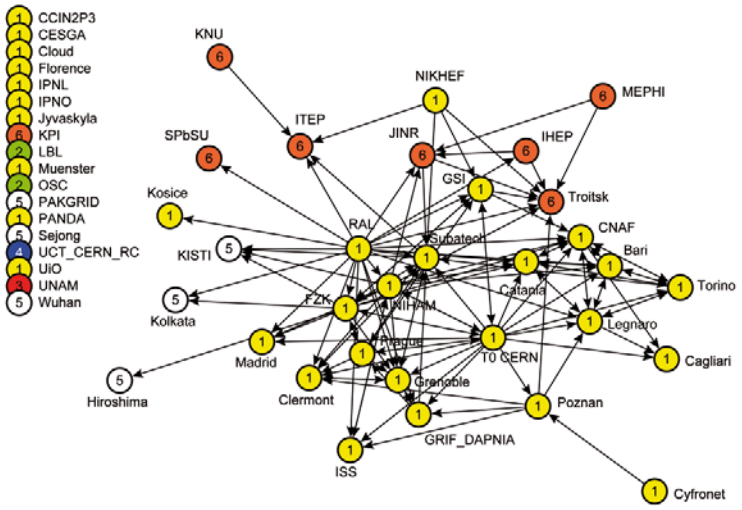

Fig. 8. Data transfer by regions.

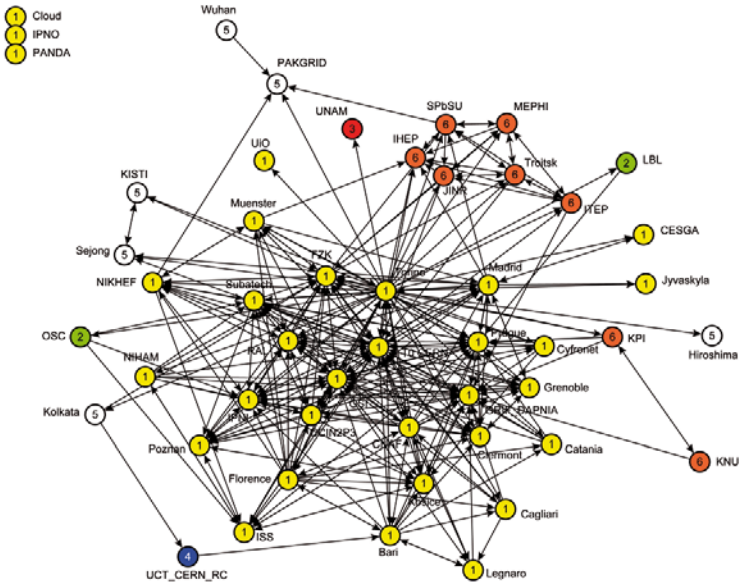

Fig. 9. RTT Capacity by regions.

Table 2 shows that, overall, the density on almost all self-reported exchanges is much higher within each region than in the full network. For instance, the density of contacts is twice higher in Europe than in the full network. Similar patterns of collaboration within regions were found for support exchange, in particular in Russia. Interestingly, the observed exchanges are also frequently organised at the regional level. The only exception concerns wished collaboration, where there is a clear distinction between centres located in Europe and centres in other regions. Centres in Europe are not especially seeking collaboration with other European centres (average similar to the overall mean). Contrastingly, centres from Russia, and especially Asia, wish to increase their interactions at the regional level. The pattern of density of interactions among Asian centres is somewhat different from that in other regions. The density on self-reported and objective exchanges is low in comparison with other regions or even the full network, while the wishes of centre managers for collaborations are very high.

We also estimated the extent to which those results were statistically significant. In order to know whether the patterns revealed in table 2 are due to chance or reveal structures, we first built a hypothetical network perfectly structured around regions, *i.e.*, in which a tie exists if and only if the centres are of the same region. We then ran a set of permutation models, based on QAP (Quadratic Assignment procedures) [7]. The Quadratic assignment methods compute correlations between the entries of two square matrices and assess the frequency of random and true correlations between them, in order to test whether or not these dimensions are correlated beyond chance. The algorithm proceeds in two steps. In the first step, it computes a Pearson's correlation coefficient between corresponding

Table 2. Density within regions.

	Help	E-mail contacts	Wished collaborations	Bandwidth	Data transfers	RTT capacity
Europe (with CERN)	6.1	8.4	3.2	40.4	22.2	38.3
Europe (without CERN)	2.5	6	1.9	38.4	19.5	34.9
Asia	3.3	0	23.3	4.3	0	11.1
Russia	23.2	26.8	16.1	22.2	16.2	54.2
Total (with CERN)	4.3	4.9	3.6	19.2	11.7	19.3
Total (without CERN)	2.2	3.4	2.5	17.8	10.5	17.6

Table 3. QAP testing of regional influences (Pearson's correlation coefficient).

	Help	E-mail contacts	Wished collaborations	Bandwidth	Data transfers	RTT capacity
With CERN	0.12**	0.17**	0.05	0.41**	0.27**	0.43**
Without CERN	0.09**	0.18**	0.06*	0.40**	0.24**	0.41**
$^{**}p < .01$ $^{*}p < .05$						

cells of the two data matrices (corresponding to the observed and hypothetical network data). In the second step, it randomly permutes rows and columns (synchronously) of one matrix (the observed matrix, if the distinction is relevant) and recomputes the correlation. The second step is carried out hundreds of times (for the present study 5000 times) in order to compute the proportion of times that a random correlation is larger than or equal to the observed correlation calculated in step 1. A low proportion ($< .05$) results in the rejection of the null hypothesis of independence (a network with permuted centres could have a correlation with the hypothetical network at least as high as the observed network) and suggests a strong relationship between the matrices that is unlikely to have occurred by chance [6]. Here again, Pearson's correlation coefficients were computed with and without CERN.

Pearson correlations between each empirical matrix and the theoretical matrix constructed on the assumption that ties only exist within regions are reported in table 3. It also provides a p-value which computes the proportion of times in which the permuted empirical matrix shows a correlation with the theoretical matrix equal or higher than the original empirical matrix. The table reveals significant effects of regions for all indices considered in this research, except for wished collaborations when CERN was included in the network. With the exception of this latter measure, belonging to the same region makes the density of interactions among centres significantly increase. The correlations are even higher for observed exchanges than for self-reported ones.

5 Discussion

The distribution of links within the ALICE project is definitely non-random. This does not come as a surprise, as the Grid is a highly structured network with specific functionalities and connections. However, this fact implies that the relations found in this paper are significant and can be interpreted as the effect of structuring factors. The patterns emerging from this study make a contribution to the understanding of the underlying structure of the ALICE Grid within the Worldwide LHC Grid structure. Access to data and resources is actually ubiquitous within the ALICE Grid. Members of the ALICE Collaboration can submit work requests from any point in the world, which are executed by any centre available with the only constraint of data locality for reconstruction jobs, while simulation workload is assigned only on the basis of CPU availability.

If we look at the "physical" layer (RTT and bandwidth) we can "see" some regionalisation in figs. 7 and 9. This is already a non-trivial observation because of the presence of large supranational initiatives like Géant in Europe establishing high-speed international networking.

Coming to the role of CERN, from table 1 we note that it does not play a particularly important role in the structure of the physical layer, as the different indicators for RTT and bandwidth show relatively little variation with and without it. Only the substantial reduction of the in-degree centralisation in the RTT suggests that CERN may play the role of a "hub" in the ALICE physical network. This role is confirmed by the opposite result for the betweenness centralisation, indicating that the removal of CERN leaves a network where more centres play the role of relay between peers in a more horizontal network.

If we now move to data transfer, we observe that CERN's role as a "hub" (in-degree centralisation) is less prominent, while the network of data exchanges is strongly organised around CERN (out-degree centralisation drops by 30% with CERN's removal). This is consistent with the fact that CERN is the ultimate "source" for the experimental data. This is mitigated by the fact that Monte Carlo data are generated at all the nodes of ALICE's Grid. A large change in the number of cliques is also consistent with CERN's role of producer of data and receiver of elaborated data for custodial storage which makes of it a preferred vertex of interconnected subsets.

The above suggests that at the "physical" layer the ALICE Grid is horizontally rather than vertically organised, with little hierachization. The data transfer is characterised by CERN's special role as source of the data, but otherwise shows little verticality. Regional densities for the "physical layer" shown in table 2 differ largely for the different regions.

If we now turn to the "relational layer" we observe a largely different situation. The density of help, emails and contacts show a large change when CERN is removed. The nature of the changes can be understood looking at the other parameters. The betweenness centralisation is sharply reduced by removing CERN and the effect is even more dramatic for the Help category, implying a structure where the largest number of links point to CERN, as can be easily seen in figs. 1–5. The drop of in-degree centralisation is also very large, which clearly indicates the "hub" role of CERN in providing help, email contacts and as a target for wished collaboration. Similar consideration for the number of cliques in wished collaborations. CERN is seen as the "missing vertex" for multilateral collaborations.

The drop in out-degree centralisation, which measures the hierachization of the system, is however less pronounced. This suggests that the hierarchical structure remains even if the "top" is removed. Looking at the graphs, this hierachization clearly depends on geographical regions. The density of wished collaborations within regions seems to be inversely proportional to the perceived existing level of help and email exchange, which is an expected result.

These findings should be referred to the hierarchical organisation foreseen by the MONARC model [8] which has largely served as a blueprint for the WLCG Grid. The structure for the MONARC "physical layer" was influenced by the foreseen limitation on the network bandwidth and the projected need to optimise resource utilisation within this constrain. The actual evolution of the network capacity has gone beyond all expectations. At the end of last century, when the MONARC proposal was put together, researchers were hoping to have 622 Mb/s links, while now links with a capacity one order of magnitude more are commonplace. As a consequence of this, at the "physical layer" this model has been replaced by a model evolving toward a more "democratic" cloud paradigm. In the case of ALICE, this move has been accompanied by precise architectural choices in the design of the middle ware and in the operation of the Grid, with the aim of optimising the usage of resources. The fact that the data transfer follows this pattern reveals that the architectural design is reflected in the current data-flow.

At the "relational" layer a completely different picture emerges. Geographical regions play, in this case, an important role in modelling the infrastructure, which looks highly hierarchized in a way similar to the one suggested by the MONARC model. While the machine-to-machine exchange draws a "cloud", the human relationships, both actual (email exchange), perceived (help) and wished (wished collaborations) follow a hierarchical scheme strongly influenced by geographical and national/cultural elements. This layer looks like an example of dynamic "self-organisation" capable to adjust itself in order to "optimise" the Grid usage from the "user's perspective" within the constrains coming from the "physical layer". Of course the "physical layer" itself also evolves as a result of the way in which the system is really used, aiming at improving efficiency and resource utilisation. Analysing the mutual influence between these two layers could be an interesting subject for future studies. Moreover, these two layers, although different, can work together in a complementary fashion. How much this relates to the evolution of Internet as a whole outside HEP is also a very relevant question, and it would be important to compare our results to similar studies in different fields. At the same time it would, of course, be very interesting to repeat this study in the future to see how the situation evolves.

The authors wish to thanks Dr. Iosif Legrand for his insightful reading of the draft.

References

1. S. Wasserman, K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge University Press, Cambridge, 1994).
2. Worldwide Large Hadron Collider Computing Grid, `http://lcg.web.cern.ch/LCG/public`, last updated on 05-10-2010.
3. E.D. Widmer, F. Carminati, C. Grigoras, G. Galli Carminati, *Proceedings of the 13th International Workshop on Advanced Computing and Analysis Techniques in Physics Research, ACAT2010, Jaipur* (PoS, 2010).
4. R. Burt, *The social capital of structural holes* (Russel Sage Foundation, New York, 2001).
5. J. Scott, *Social Network Analysis* (Sage, London, 2000).
6. S. Borgatti, M.G. Everett, L.C. Freeman, *Ucinet 6 for Windows. Software for social network analysis* (Technical report, 2002).
7. L.J. Hubert J. Schultz, Br. J. Math. Stat. Psychol. **29**, 190 (1976).
8. The MONARC Project, `http://monarc.web.cern.ch/MONARC`, last updated on 13-01-2000.