

Social Network Analysis Used for Modelling Collaboration in Distance Learning Groups

Christophe Reffay¹, Thierry Chanier¹

¹ Laboratoire d'Informatique de l'université de Franche-Comté
16 route de Gray, 25030 Besançon cedex, France
{Christophe.Reffay, Thierry.Chanier}@univ-fcomte.fr

Abstract. We describe a situation of distance learning based on collaborative production occurring within groups over a significant time span. For such a situation, we suggest giving priority to monitoring and not to guiding systems. We also argue that we need models which are easily computable in order to deal with the heterogeneous and the large scale amount of data related to interactions, i.e. models relying on theoretical assumptions which characterise the structures of groups and of interactions. Social Network Analysis is a good candidate we applied to our experiment in order to compute communication graphs and cohesion factors in groups. This application represents an essential part of a system which would enable tutors to detect a problem or a slowdown of group interaction.

1 Introduction

In computer science, as soon as several prototypes belonging to a sub-domain have been developed, we often try to establish a categorization among them, categorization often based on system functionalities. We then feel more secure, after having reduced the size of open-ended problems, introduced some ways of comparison and judgment among systems, and even indicated a direction for future research. This is what happened in CSCL –Computer Supported Collaborative Learning– with the often referred to Jerman & Al's paper [8], entitled "From mirroring to guiding: a review of the state of the art technology for supporting collaborative learning". Its authors classify CSCL environments according to their type of intervention defining three main categories: Mirroring Systems that reflect actions; Monitoring Systems that monitor the state of interaction; Guiding Systems that offer advice.

But this simplified presentation of CSCL work conceals several levels of problems which may be important issues in ITS. Among them: the nature and size of the data from which systems can make computations; the variety of learning situations; the relationship between tools/systems, models and theoretical perspectives.

Let us look at the data perspective first. At one end, Mirroring Systems seem to be reduced to basic computations of raw data –one if an email message has been sent, zero otherwise and, from there on, computation of the number of sent messages–, and, at the opposite end, Guiding Systems process various sorts of highly structured data. Experience from research in AI or from Student Modelling shows that, if we aim at

developing systems that can make decisions and give advice to a learner on a specific learning task, it needs to heavily rely on domain-knowledge and on detailed task-descriptions. But domain-dependent approaches can hardly provide generic solutions. Systems developed to support collaboration during the learning process need to rely on the basic data, which often are textual data, coming from communication tools. CSCL literature presents interesting Monitoring Systems built on textual data, partially structured with a subpart made of sentence openers. Free input linguistic data are even accepted when the system includes NLP treatments. But what happens if the system has to deal with thousands of emails and conferences messages, thousands of speech-turns in synchronous dialogs –see, for example, the Simuligne figures? Then we are faced with a size and scale problem. In this case, it is worth reconsidering basic data: knowing whether one has opened/read a message –and not only whether a message has been sent– is a piece of information from which interesting inferences can be made, as we will see. Moreover, gathering and structuring communication data in large-scale environments is not straightforward.

The question of scale leads us to the second level of the problem, i.e. the non distinction often made among the variety of learning situations. As Fjuk & Al. [5] says:

"the problem area within most CSCL research in general, and in distributed collaborative learning in particular, is that their ecological validity could be considered low, since most studies are experiments or small-scale field trials. [Some] studies [...] are limited to experimental settings, or field trials where the time span of the learning activities is of short duration. [...]"

Besides time span, another characteristic of learning situations in CSCL is whether learning happens, on the one hand, in face-to-face or a mixture of face-to-face and distance situations or, on the other hand, in real distance learning situations. Confusing both and asserting that learning and teaching issues are the same is hardly convincing. Of course, we are not claiming that small-scale collaborative experiments, involving learners who have part of their syllabus in face-to-face courses are of no interest. We simply claim that the aim of supporting collaboration and interactions in real distance learning environments raises specific research priorities [7]. For example, because the role of the human tutor is critical in distance learning and her/his workload is more important than in face-to-face situations, it is worth having Distance Learning Management Systems –DLMS– that can automatically compute and show the structure of learning groups, as well as their cohesion and send warnings before the situation becomes irreversible.

Mentioning collaboration in distance learning groups brings us to the third concealed level of problems: the relationship between tools/systems, models and theoretical perspectives. Someone looking from outside how we, computer-scientists, sometimes deal with issues in CSCL may be surprised. On the one hand, from time to time, we may pretend developing generic tools unrelated to any theoretical concerns, and on the other hand, desperate for semantics, we may quote a citation of Vytgostky, coming from the thirties, as if it could exactly fit into our current concern! Even if the latter sentence is a caricature –for the sake of understanding–, it helps us introduce the fact that there can be a middle way. When involved in the improvement of existing DLMSs we are faced with new issues where interactions are not restricted to learner-system nor learner-learner pairs interactions but should be considered at a group level.

There already exists models in sociology that see interaction at a structural level – "interactionisme structurel" in French or "Social Network Analysis" –SNA– in English. These models are computable and lead to tools which may be reused in our field. If we do think these tools are useful for the improvement of DLMSs –as Nurmela & Al [10] and Wortham [14] did when implementing specific SNA tools– we should also take into account the corresponding models and theoretical assumptions. Consequently, at a research level, it is worth considering spending time developing Monitoring Systems strongly linked to theoretical assumptions, before attempting to build Guiding Systems without knowing what exactly is at stake.

In the first section of this paper, we will give an overview of the Simuligne experiment. The learning situation will then be fixed and from there, the kind of DLMS we are concerned with. The second part will introduce the SNA approach and relate it to CSCL concerns. Researchers in SNA always had problems when collecting data. Within our electronic environments, data are accessible, provided that we decide which one to consider as relevant. This issue will be discussed in the third section. The following section will show how SNA-based-graphs algorithms can be applied to build our first learning group structures and measures of interactions on a subset of data collected in the Simuligne experiment.

2. The Simuligne Experimentation

Special thanks to which supports the ICOGAD project.

Simuligne was born in a trans-disciplinary research project named ICOGAD (Great-Britain), the Computer Science Laboratory of the Université de Franche-Comté (Besançon, France) and the Psychology Laboratory, Université de Nancy2. ICOGAD, sponsored by the French Minister of Research (MRT) and its cognitive science programme (Programme Cognitique 2000) whose partners are the Department of Language Learning at Open University, is the whole research project. It includes the conception, production and delivery of the online learning stage named Simuligne.

In Simuligne, we had 40 learners –English adults in professional training, registered at the Open University–, 10 natives –French teacher trainees from Université de Franche-Comté–, 4 tutors –teachers of French from the Open University– and one pedagogical coordinator. They can all be classified in one of the three classes of actors of this distance learning experiment: learners, experts –natives– and teachers –tutors and the coordinator. All actors were dispatched into four learning groups, namely Aquitania, Lugdunensis, Narbonensis and Gallia. In another group, the trainers' group, the coordinator, tutors and natives could share questions and answers while the simulation went on. French as a foreign language was the learning subject. Looking for a collaborative production-oriented project, we decided to adapt the method called "Simulation globale" for the first time to a distance situation. The global simulation method is based on role playing and is often used in intensive face-to-face language learning. Distance was the rule: everybody worked at a distance; no one had ever met before Simuligne, except the natives from Besançon. The only people working face-to-face were the technical team, some of the designers and the pedagogical coordinator in Besançon, where the DLMS server stands. Learners and

tutors did not know the technical features of the DLMS beforehand. Consequently we had to train all the tutors, at a distance, before Simuligne really started. They were trained on the technical aspects of the platform as well as on the global simulation, a pedagogical method that most of them had never practised.

Other important factors of that experiment are sequence and duration. Simuligne spanned over 10 weeks, broken down into 4 parts:

- Stage 0: 2 weeks: self-introduction to the group and acquisition of technical skills,
- Stage 1: 3 weeks: designing the place of the simulation, –city, campus map, ...
- Stage 2: 3 weeks: defining the various characters and putting them in various situations to solve some problems –explosion on campus, buses on strike...
- Stage 3: 2 weeks: discussing and voting for the favoured project.

Three groups out of four achieved the simulation, which is a high ratio in distance learning. When the Lugdunensis group broke up, its most active learners were transferred to another group. Posters produced by the three remaining groups presented rich and high quality language productions [3]. But what can be assessed afterwards by teachers on interaction and collaboration in the group need to be more carefully understood if we want to improve learning environments. The following section gives some representation of communications from which it is possible to have an overview of the interactions happening inside a distance learning group. These are the first steps if we want a system to be helpful to evaluate interaction.

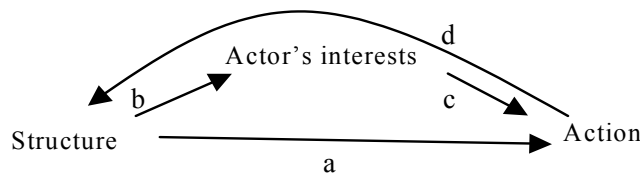


Fig. 1. Basic assumptions in structural theory.

3. Introduction to Structural Interactionism

Social Network Analysis [4,12] is a large research field in sociology and ethnology. Its major objective is to characterise the group's structure and, in particular, the influence of each of the members on that group, reasoning on the relationships that can be observed in that group. SNA has developed a theoretical approach thoroughly different from the traditional social analysis. In a traditional social quantitative analysis, the population is chosen for its variety, representativeness and is classified according to its individual characteristics such as age, sex, social class, ... Then, the study compares some extra attributes and uses statistic tools to give laws of dependency between some of these attributes. The main problem with this approach is that categories are defined before the analysis by describing the various attributes of individuals. On the contrary, SNA focuses on the relationships between individuals instead of the individuals themselves. In other words, in SNA, the basic

item is the group where, as Block [2] says in holism theory: "*an individual acts according to the group he belongs to*".

In distance learning based on collaborative production, we start with individuals that have to socialise in order to form a group which shares goals and values. The existence of this group is a key issue if we want the group to produce collaboratively. Once this group exists, each visible action of each individual will alter the structure of the group. What holism says in the Structural Theory (Burt 1982 in [4]) –see figure 1– is: the weight of this structure will influence a) the actions of the members and b) the members' interests so that if the member is a rational person, c) he will act according to his interests. The resulting new action will again d) modify the group's structure. The evolution of the structure, always modified by actions is well illustrated by Leydesdorff 1991 [4,p.15] who introduces time.

The structure of Burt (in figure 1) represents a basic collaborative learning group also called community. Woodruff [13] explains how cohesion is the key issue for collaborative learning and defines four cohesive factors he called "glue factors":

"1) function, 2) identity, 3) discursive participation, and 4) shared values. Briefly, function is the goal or purpose of the community; identity is the validation of 'self' through membership; discursive participation is the means by which the members' discourse helps to advance the function or goal of the community; and, shared values are the global beliefs held by members which unite them and help to promote an emerging discourse."

The third glue factor namely "discursive participation" is what we call interactions and is the most visible to the members, in so far as it gives a measure of group activity. For the researcher, this cohesive factor may be quantitatively valuable. More likely in a classical context, Homans in [4, p.95] pointed out the cohesion concept related to appreciation and interaction. He says that the cohesion of a given group enhances the appreciation of each of its members. The more people appreciate each other, the more they will interact. Interaction being a glue factor, it will reinforce cohesion. "Appreciation", in Homan's model, is probably very close to Woodruff's fourth cohesive factor, named "shared values". In fact, Woodruff says that the glue factors he identified are closely linked to one another: changes in one factor will inevitably have an effect on all the others.

These characterisations can be automatically computed by using matricial tools of the graph theory [1]. The difficulty in SNA in general, is to collect the large amount of data which define the relationships between individuals: collecting data is often a hand-made process. In our distance learning context, these data are stored in the DLMS. We have access to a large amount of data where it is sometimes very hard to find relevant information, as we will see in the next section.

4. Data for Communications and Interactions

The basic tools for communication in DLMS are e-mail, discussion forum –also named "conferencing system"– and chat. Communication is based on textual data and happens either asynchronously with the first two tools or synchronously for the latter one. Here are the figures coming from the Simuligne experiment:

- Discussion forum: 879 015 characters in 2686 messages, which represent: 45.11% of the communication flow measured in number of characters;
- E-mail: 834 753 characters in 4062 messages: 42.84%;
- Synchronous chat: 234 694 characters in 5680 speech turns: 12%.

Synchronous communications represent 12% of the whole set of data. Activities based on chat are interesting in language learning: it increases motivation, it develops abilities in conversations on the fly, provided that there is a limited number of participants who know which role to play and who have prepared it in advance. But activities using chat are difficult to integrate in the agenda and impedes the flexibility that characterises and make success of Distance Learning context.

Moreover e-mail and e-conference are generally considered as the core tools in distance learning, because they –not only– convey communications around the knowledge domain but also about the whole learning process [7,11].

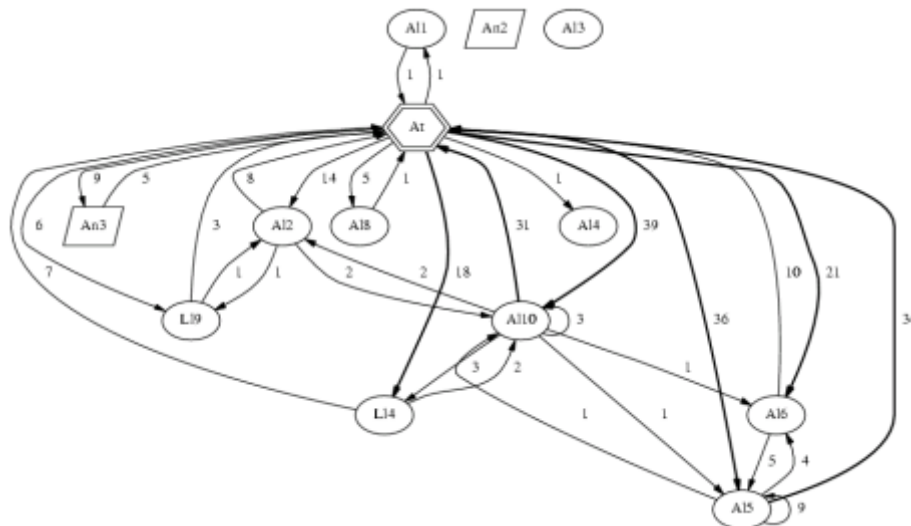


Fig. 2. E-mail graph of group A (Aquitania) on the whole period

Consequently, looking for structures that could automatically be built in order to reflect communications inside groups, we decided to apply SNA models to these asynchronous data, starting with e-mail. This supposes to retrieve detailed information on each of the messages. This information varies from one tool to another and is not straightforwardly accessible. Technically, the problem is to get the precise information among a very large and amount of basic data.

5. Building Graphs to Reflect Group Communications

In our model, we will try to ignore/mask all the messages that have not been opened by their recipient. Because such messages have been sent, we agree they have

some influence on a *participation* index, but they may be ignored when dealing with an *interaction* index.

Using the resulting database, it becomes possible to compute, for each group, a graph of communications where we only select *messages that have been opened*.

From there, we can define a graph of group communications:

Let $G_o = (U, I)$ an oriented and valuated graph of order n where:

- U is the set of n vertices: the n group users: X_1, X_2, \dots, X_n ;
- I is a family of oriented and valuated relations representing interactions between the users of U . Each edge belongs to $U \times U \times \mathbb{R}$ represented by (U_i, U_j, v) where source user U_i and destination user U_j are users of the set U and v defines the volume –number of messages– emitted by U_i and opened by U_j .

Having selected and restructured the kinds of data we need and defined our graph we are now able to design various representations.

The first step is to represent the volume and destination of all the communications sent by each user or the volume and source of all the communications received and opened by each user. Let us now illustrate this with e-mail messages.

The following example focuses on one of the four learning groups (Aquitania=A) during the whole period of the Simuligne experimentation (10 weeks). The communications represented are restricted to e-mail messages received and opened by their addressee(s). Extracting data from the databases mentioned previously, an e-mail matrix is built. Then, using the open source GraphViz package [6], we can automatically generate the visual representation of the e-mail graph (figure 2)

The edges of the graph are valued by the number of messages sent –by the user at the origin of the directed edge– and consulted –by the user at the end of the same edge. This picture immediately gives an idea of the central role of the Tutor –At– from/to whom the majority of the messages, sent and consulted, converge.

Let us now see what happens when we remove the tutor –At– and all the messages he is concerned with (see figure 3, top left corner for Aquitania). The notion of cohesive subgroup begins to be clear on that graph. It is not only the list of users interrelated by e-mail, but also who is in relation with the maximum number of other people and who is absolutely not connected –by e-mail– with others. In particular, we can see in this group that there is no e-mail communication with An2 and An3: the two natives of the group who were supposed to bring support as language experts if needed. A12, A15, L14, L19, A16 and A110 are the learners who successfully finished the Simuligne training by an actual production in the final group project result. Note that L14 and L19 are transferred students from the “dead” group Lugdunensis (L) to Aquitania (A).

The comparison given in figure 3 of the four groups on their consulted e-mail graph without tutor seems to be relevant for the way a group collaborated and produced.

Firstly, the data exposed in figure 3 are part of those needed by the global coordinator in charge of the inter-group regulation and of the tutor support. But these data are only partial data even for the communication part of the training.

Secondly, the main fact we are interested in, is to get a real value of a cohesion index for each group at each instant in order to give a comparative representation of the cohesion evolution of the four basic groups for each stage.

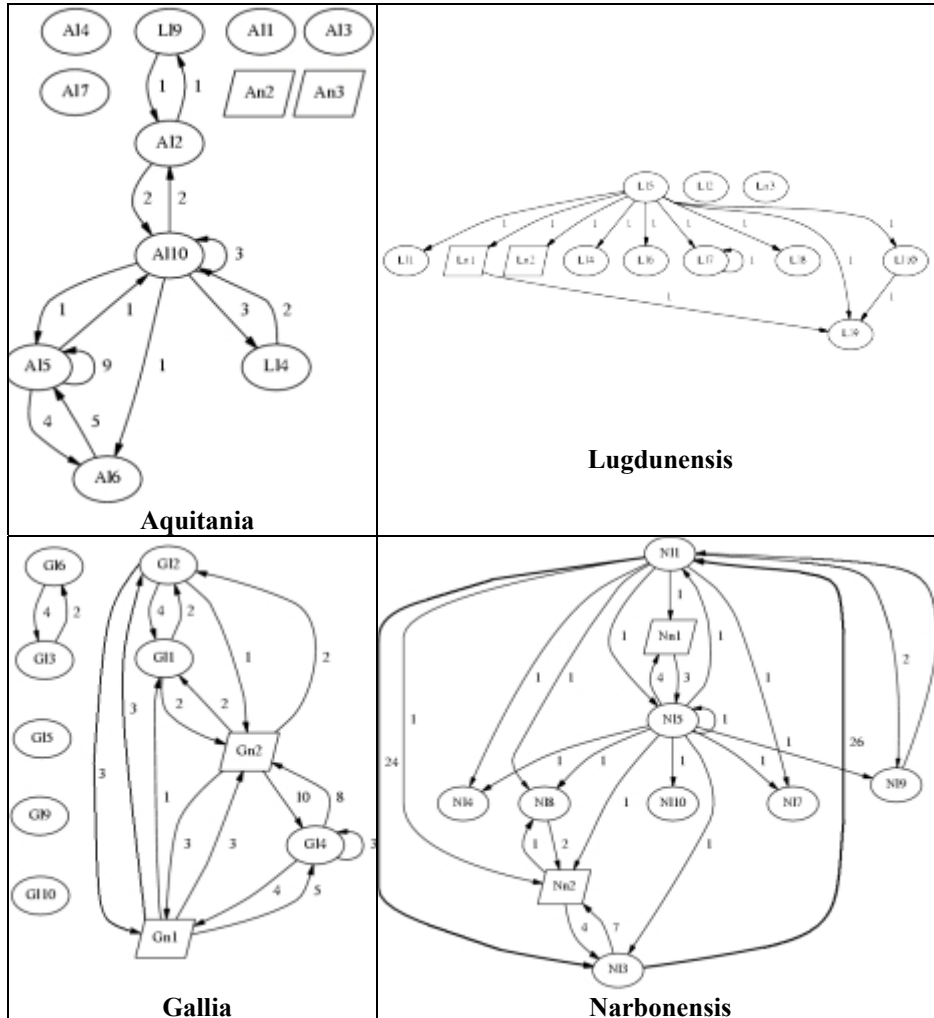


Fig. 3. Comparison of the consulted e-mail graph (without tutor's messages) of the four basic groups during the whole period of the Simuligne Experimentation.

6. Cohesion in Communications

Starting from the list of all forum messages and for each of them, the list of the users who read it, let us describe the computing process of the cohesion index of the group for a given period –adapted from [4, p. 100].

Firstly, for each couple (L_i, L_j) of learners, we compute the number x_{ij} of forum messages posted by L_i in the target period and read by L_j . We build the matrix A

where $a_{ij}=1$ if $x_{ij}>0$. Then, the symmetric matrix S is given by $s_{ij}=\max(a_{ij},a_{ji})$ for weak cohesion –a strong cohesion factor would be obtained replacing *max* by *min* function. S is the adjacency matrix representing the relations existing in the group. s_{ij} is 1 if L_i or L_j read at least one message posted by the other in the given period. Let n_i be the number of relations of L_i including himself. We have $n_i=\sum_k s_{ik}$, and if n_{ij} counts the relations shared by L_i and L_j , $n_{ij}=\sum_k \min(s_{ik},s_{jk})$. It is then possible to compute d_{ij} :

the recovering degree of relations circles of L_i and L_j given by: $d_{ij}=\frac{n_{ij}}{n_i+n_j-n_{ij}}$. If

$d_{ij}=1$, it means that L_i and L_j share all their relations and if $d_{ij}=0$, they don't have any common relation. The cohesion index of the group is then given by the means of d_{ij} values for all possible pairs L_i and L_j of learners in the considered group for the target period.

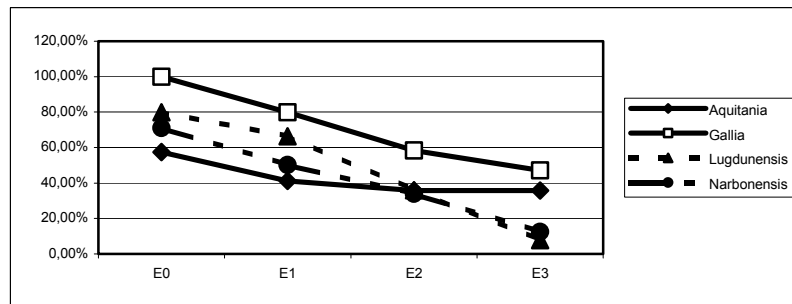


Fig. 4. Evolution of the cohesion factor in the four basic groups

By separating Simuligne in its four stages E0, E1, E2 and E3, we computed this cohesion factor for each group and for each stage. The resulted graph is shown in figure 4 where erosion –due to the abandon of some learners in normal proportion for distance learning– forces the evolution of the cohesion factor for each group to decrease. We can also notice that, during E0, the Gallia group obtained a maximum score for cohesion revealing that each member of Gallia had at least one interaction with all the others in the group.

It is also perceptible that in moving two active members from Lugdunensis to Aquitania, the Lugdunensis's cohesion factor fell while Aquitania's one remained the same during E2 and E3. Taking into account only the existence but not the number of interactions, this index much reduces the intensification of interactions between the remaining members of each group. The final means of d_{ij} is heavily thinned down by the entire lines and columns of zeros concerning the abandon, despite the fact that the values concerning the remaining learners are nearly all equal to one. The first method presented here to compute the cohesion index does not illustrate properly the growing cohesion of reduced and more and more active subgroups –empirically identified during the Simuligne experience. To avoid such a problem, we suggest we take into account the number of messages between each couple of members and we limit the computation of the cohesion index to the “interesting” subgroups.

7. Conclusion

We have described a situation of distance and online learning based on collaborative production. For such a complex situation we suggest we give priority to monitoring and not to guiding systems, even if other works in progress in our research group reconsider a whole DLMS as a multi-agent architecture in order to sustain interactions [9]. We have to understand exactly what happened in the groups during such an experiment. We showed that Social Network Analysis gives an interesting theoretical background to compute various global indices such as communication graphs and the cohesion factor of a group. A complete access to communication data is needed in order to reorganise them in fruitful databases. We have started to show the interest of some sociometric measures using a large volume of data. This work is the first step in developing a software that would enable tutors or pedagogic coordinators to detect a problem or a slowing down of group interactions. In the near future, we need to refine the computation of the cohesion index and assess other SNA indices with respect to our Simuligne experience feedback.

References

1. Berge, C.: Graphs and Hypergraphs. Dunod, Paris (1973)
2. Block, N.: "Holism, Mental and Semantic". In The Routledge Encyclopedia of Philosophy (1998), <http://www.nyu.edu/gsas/dept/phil/faculty/block/>
3. Chanier, T.: Créer des communautés d'apprentissage à distance. Les dossiers de l'Ingénierie Educative, no 36 sur "Les communautés en ligne", CNDP, Montrouge (2001) 56-59, <http://life.univ-fcomte.fr/RECHERCHE/P7/pub/cndpIE/cndpIE.htm>
4. Degenne, A., Forsé, M. : Les réseaux sociaux. Armand Colin, Paris (1994)
5. Fjuk, A., Ludvigsen, S.: The Complexity of Distributed Collaborative Learning: Unit of Analysis. Proceedings of EURO-CSCL'2001 Conference, Maastricht (2001) <http://www.mmi.unimaas.nl/euro-cscl/Papers/51.doc>
6. GraphViz Home page (2000) <http://www.graphviz.org>
7. Henri, F., Lundgren-Cayrol, K.: Apprentissage collaboratif à distance. Presses de l'Université du Québec, Québec (2001)
8. Jermann, P., Soller, A., Muehlenbrock, M.: From mirroring to guiding: a review of state of art technology for supporting collaborative learning". Proceedings of EURO-CSCL'2001 Conference, Maastricht (2001) <http://www.mmi.unimaas.nl/euro-cscl/Papers/197.pdf>
9. Mbala, A., Reffay, C., Chanier, T.: Integration of automatic tools for displaying interaction data in computer environments for distance learning. This issue, Biarritz, France (2002)
10. Nurmela K.A., Lehtinen E., Palonen T.: Evaluating CSCL log files by Social Network Analysis. Proceedings CSCL'1999 Conference, Palo Alto, CA: Stanford University (1999) 434-444. <http://kn.cilt.org/csc199/A54/A54.HTM>
11. Salmon, G.: E-moderating: The key to teaching and learning Online. Kogan, London (2000)
12. Scott, J.: Social Network Analysis. A handbook. SAGE Publication, London (1991)
13. Woodruff E.: Concerning the Cohesive Nature of CSCL Communities. Proceedings of CSCL'1999 Conference, Palo Alto, CA: Stanford University (1999) 677-680 <http://kn.cilt.org/csc199/A81/A81.HTM>
14. Wortham D.W.: Nodal and matrix analyses of communication patterns in small groups. Proceedings CSCL'1999 Conference, Palo Alto, CA: Stanford University, (1999) 681-686 <http://kn.cilt.org/csc199/A82/A82.HTM>