USING AN AGENT-BASED SIMULATION APPROACH TO STUDY INFORMATIONAL ASPECTS OF ANSWERING BEHAVIORS IN MAILING-LISTS

Emilie Marquois-Ogez
France Telecom R&D
38-40, rue du Général Leclerc 92794
Issy-les-Moulineaux Cedex - France
Emilie.Marquois@francetelecom.com

Frédéric Amblard
IRIT/Université Toulouse 1
21, allées de Brienne
31042 Toulouse - France
frederic.amblard@univ-tlse1.fr

Cécile Bothorel
France Telecom R&D
2, avenue Pierre Marzin
22307 Lannion Cedex - France
Cecile.Bothorel@francetelecom.com

KEYWORDS
Multi-agents simulation, social simulation, virtual communities, mailing-lists, participation behaviour.

ABSTRACT

As a complement to a classical sociological analysis concerning participation and reputation effects within mailing-lists, we built a collection of agent-based simulation models in order to test hypothesis concerning the generative mechanisms of the behaviors isolated from data analysis. In particular, a first probabilistic model dealing with the way participants answer to messages causes us to refine our hypotheses in order to take into account the discrepancy between the level of knowledge included in the messages and the knowledge of the agent willing to participate.

INTRODUCTION

The use of agent-based social simulation (ABSS) widens as a complementary tool to a classical sociological analysis. It enables especially to describe observed phenomena but also to propose formalized hypothesis concerning the generating mechanisms (Manzo 2005) of theses phenomena (by the simulation) at an individual level (following the multi-agents approach) or of regularities observed at a macro level (Gilbert and Troitzsch 1995; Amblard and Phan 2006). In this paper, we apply ABSS to virtual communities in order to identify the key elements of participation in mailing-lists.

Many interesting questions have been addressed concerning the participation in virtual communities (for instance, why do people initiate a discussion? Why do people contribute to certain discussions but to others? …). The studies identified in the literature aim for instance at understanding why an individual participates or not in the exchanges (for instance, surveys on the members who never post or very rarely) and to identify the factors (essentially socio-psychological ones) that lead individuals to take part in the life of a virtual community (for instance the building of a climate of trust (Tung et al. 2001) and of a feeling of belonging to a virtual community (Blanchard and Markus 2002)).

In this work, we take the stance that the knowledge a participant has on a subject and the interest he accords to it, can explain the participation in mailing-lists. We also consider that reputation effects are elements that may play a significant role and may influence the different rates of types of messages (sending information, request or answering to a request for a basic typology) as well as the level of participation in virtual communities. In this way, we study virtual communities from two different points of view: an informational dimension (the participation of the agents depends on their interest in the topic and the knowledge they have about it) and a social one (the participation of the agents depends on their reputation and on the one of the other agents). This paper focuses specifically on the informational aspects of answering behaviours.

The main definitions of the notion of virtual community we adopt are sociological (Rheingold 1993; Wellman 1997): a virtual community is composed by individuals who interact using tools from the information technologies such as mailings-lists, chat, blog… Moreover, being a community implies the consciousness of its existence, either from the inside (I belong to a community and I am able to define it) or from the outside (I have a representation of a community I can identify even if I don't belong to it) and/or to share cultural and/or communicational elements in the group.

The literature exhibits a multitude of typologies for virtual communities, each typology corresponding to particular principles and perspectives of research (Schubert et Ginsburg, 2000). Within the framework of our work, we chose to use the typology proposed by (Stanoevska-Slabeva and Schmid 2001). Contrary to the other typologies, it is based on the goal of the virtual community. It constitutes an interesting starting point; knowing the goal of a virtual community allows us to better understand the nature of the stakes (social and/or informational) that characterize them. Moreover, (Stanoevska-Slabeva and Schmid 2001) examine for each type identified, some of the tools (generally text-based or graphical and synchronous or asynchronous) they recommend. We focus on the first of the four types of virtual community identified: the e-mail-based discussion lists of a community of practice.

The paper is organized as follows. After presenting the case background, a first model and the research methodology are introduced. Next, the different agent-based simulation models are described and the results discussed. The conclusion presents the main directions for future research.

CASE STUDY: THE MAILING-LIST ERGO-IHM

We obtained from its moderator, two-years (2002 and 2003) archives of the French-speaking mailing-list ergo-ihm (http://listes.cru.fr/wws/info/ergoihm). It is a community of practice created in 1999 for the exchange of ideas, information, and so on, in the area of human computer interface and human factors. It holds actually more than 1000 members and the exchanged information is free and not strategic.

The first step of our work on this case study consisted in analyzing the archives of the mailing-list (with 613 members, 698 e-mail address, 1880 discussion threads identified and 4101 exchanged messages) and the results of two surveys of individual interviews. The first one was conducted from July 11, 2004 to September 30, 2004 and answered by 88 members of the mailing-list. It deals with the activity of the participants to the mailing-list. A deeper questionnaire was conducted from February 16, 2005 to March 31, 2005 and answered by 23 identified members. It aimed to know if the interactions were motivated either by social or informational factors.

The analysis of the surveys results and of the archives exhibits different types of participation and messages (post of information, of requests or answering to requests). Focusing on members, differences can be found in levels and types of participation. More than 80% of the members of the mailing-list sent between 1 and 10 messages for the two years we analyzed.

Thus, we find the different categories of contributors identified by Millen and Dray (2000): regular contributors, sporadic contributors, very infrequent contributors and lurkers. According to the participants in the survey, some members are “inevitable” and a very active minority of 20 members is clearly identified by the members. Moreover, respondents to the questionnaire evaluate the percentage of active members from 1 to 20 %.

Concerning the types of participation, there are also some differences: some members only answer messages (156 members), some of them only initiate messages (96 members), and others both answer and initiate discussions (206 members). For the three cases, the proportions vary from an individual to the other. The analysis of the survey results showed that some individuals who participate a lot in the exchanges, are considered as experts by the other participants. One can notice too that the mailing-list constitutes a resource, a source of information for the members, and most of them subscribed to the mailing-list to stay informed concerning the area of ergonomics, (76 %) and consider the other members as persons that could help them (66 %). Finally, for them, answering messages is a way to help the community of ergonomics (27 %). Participants answer a message if they have sufficient knowledge to share or if they can improve or precise the previous answers and they make requests if they basically need of information. It is also important to say that members don’t read all the messages. They make a selection according to the topic of the message and/or according to the author of the message.

Joining a mailing-list, people want to get a general understanding of the topic and to satisfy needs for information.

According to the analysis of the results of the second survey, the identity and the reputation of the author of the messages influence the participation of the members. For instance, some of them said that they answered more readily to known members and others that they were enable to answer messages because of the perceived difference of knowledge between them and the author of the message.

DEFINITION OF A FIRST MODEL

On the basis of this empirical work, we developed a first model of the mailing-list dynamics. This model allows us to describe in a homogeneous way (considering the structure of the agents) the different aspects of information exchange (or knowledge sharing) and acquisition in discussion lists. Our model aims first to organize the knowledge we acquire on the system from the first analysis. Our objective is then not to reproduce real behaviours or to make predictions, but to test hypothesis we have concerning the behaviour of the individuals in this community. In a few words, we try to explain the way the modelled system functions.

Figure 2 depicts a simple UML class diagram for the conceptual model. The next section is about the objects and the relations between objects.

Description of an agent

We consider a population of *N* agents. Each agent of the mailing-list is described by three kind of attributes: the level of knowledge it has on each of the *n* subjects: < K_{subject} >, the level of salience for each of the *n* subjects: < S_{subject} >, and a matrix ($n*N$), for the evaluation of the reputation of each one of the *N* agents in the population on each one of the subjects (< $E_{\text{agent,subject}}$ >) including self-reputation.

The example of an agent is given below (Figure 1).

Reading of the messages

In our model, each message posted by an author, is relative to a subject and contains a level of knowledge on this subject. This level of knowledge is lower or equal to the level of knowledge of the author of the message has on this subject. The agents of the mailing-list receive all of the messages but read the message according to the probability corresponding to their interest on the subject $S_{subject}$.

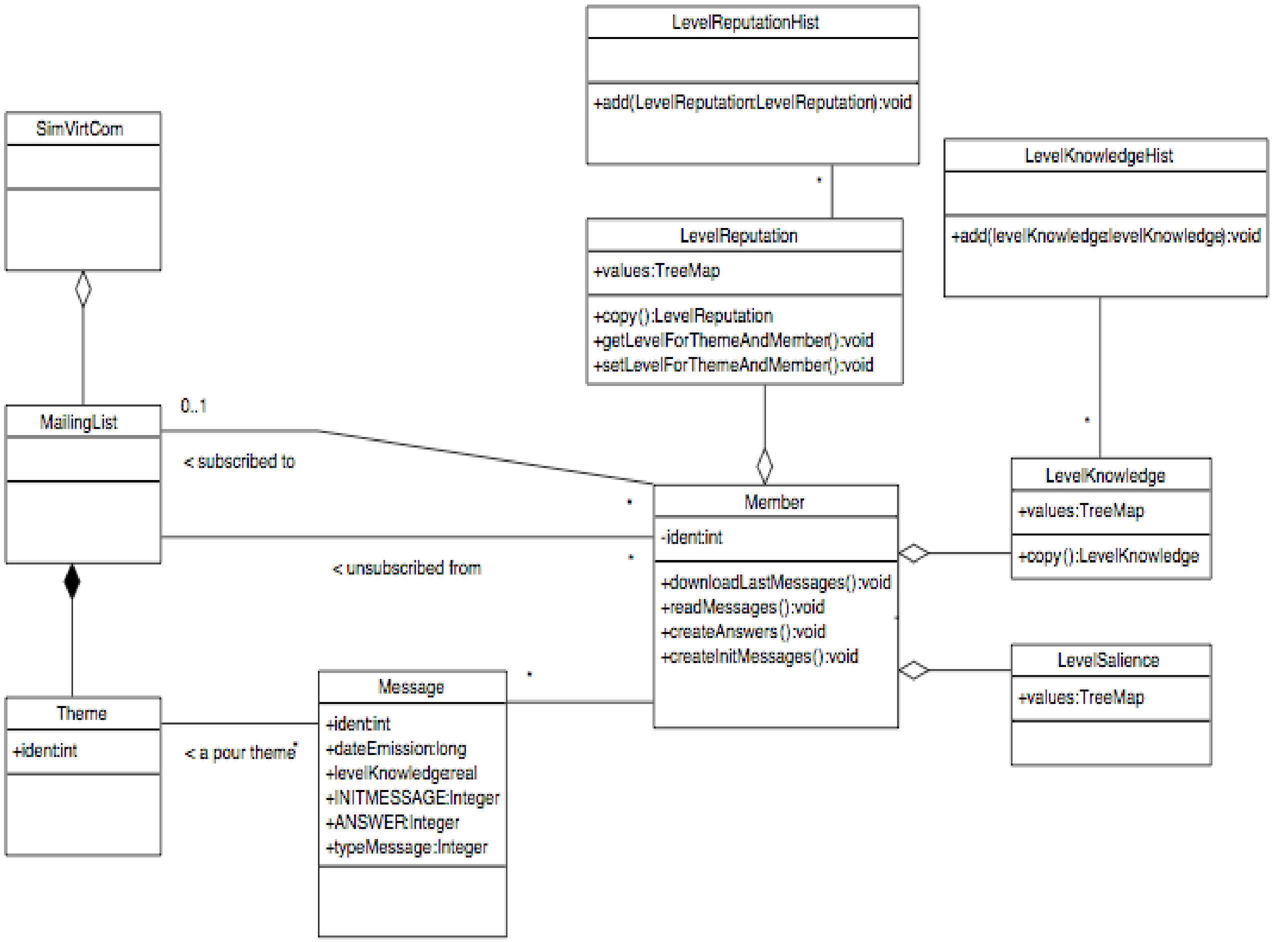

Figure 1: UML class diagram of the model

INFORMATIONAL DIMENSION

levelK

subject1	subject2	subject3	subject4
0,2	0,3	0,1	0,5

levelS

subject1	subject2	subject3	subject4
0,5	0,6	0,2	0,8

SOCIAL DIMENSION

levelR	subject1	subject2	subject3	subject4
agent1	0,2	0,3	0,8	0,5
agent2	0,4	0,2	0,5	0,1
agent3	0,1	0,9	0,5	0,2
agent4	0,7	0,3	0,4	0,3

Figure 2: Attributes of the agent 1.
The level of knowledge of the agent 1 on the subject 1 is 0,2; the interest the agent 1 gives to subject 4 is 0,8; the level of expertise of the agent 1 on the subject 2 is 0,3; according to the agent 1, the level of expertise of the agent 3 on the subject 2 is 0,9

Updating of the levels of knowledge and reputation

After reading a message on a particular subject, and if his level of knowledge (*levelK*) is lower than the knowledge contained in the message (*levelKmess*), the agent increases his knowledge with a part of the difference *(levelKMess-levelK)* .

He updates also his level of reputation (*levelR*) and the level of reputation of the author of the message (*levelRauthor*), on this subject according to the result of the comparison between the knowledge contained in the message and his own knowledge.

The level of salience doesn't vary in this version of the model.

Send messages on the mailing-list

We distinguished two kinds of messages: the messages that initiate a discussion such as requests or calls for papers and the answers to these initiating messages. The answering to a message varies in the article and is therefore detailed below.

SIMULATION OF THE ACTIVITY OF A MAILING-LIST

Protocols and system parameters

Following a decreasing abstraction methodology proposed by (Lindenberg, 1992) and operationalized by (Amblard et al., 2001), a simulation framework has been implemented in Java.

For each model, the system parameters are the following: the number of agents in the population N that is kept static in the presented simulations, the number of iterations *nbIter* and the number of themes. The first two parameters vary from a simulation to the other but in a first attempt we consider only one theme.

For each model, we observed the following indicators: total number of messages, number of answers and number of messages initiating a discussion. The number of messages for each discussion and the level of knowledge of each message are given in some cases. Ten replications are run for each simulation.

At the algorithm at each iteration is the following: 1) an agent is randomly selected; 2) it downloads the messages sent to the mailing-list since his last connection; 3) he reads them or not depending on his level of salience S; 4) if he reads the message, he can answer it depending on his particular behaviour detailed below; 5) he decides to initiate a discussion depending on his level of salience S.

Elements concerning the level of salience don't change from a model to the other: The level of salience lies between 0 and 1 and is the same for all the agents in the population.

The changing algorithms from one model to the other concern the calculation of the probability to answer, the calculation of the probability to initiate a discussion and the calculation of the level of knowledge contained in the messages. The three alternative models we developed are described in Tables 1.1 and 1.2.

The model A corresponds to a null hypothesis and allows us to understand the basic properties of the core model concerning the answers whereas the following models explore some plausible hypotheses concerning the posting behaviour of the individuals.

Table 1.1: Description of the rules applied for the three models –concerning the calculation of probaInit, probaAns and levelKnewMess

probaAns : probability to answer a message; probaInit : probability to initiate a discussion; levelKnewMess : level of knowledge contained in a new message

	Probabilities		
	ProbaAns	ProbaInit	LevelKnewMess
Model A	static [0.01,0.1] – 0.01	= levelS probaInit > nb	-
Model B	= (levelKmess – levelKMemb)	= levelS probaInit > nb	[0,Kmemb]
Model C	= (levelKmemb – levelKmess)	= levelS probaInit > nb	[0,Kmemb] [Kmess,Kmemb]

Table 1.2: Description of the initialisation of the three models

	Attributes that characterize an agent		
	levelS	levelK	readMess
Model A	[0.1,1] – 0.1	-	levelS
Model B	[0.1,1] – 0.1	[0.1,1] – 0.1	levelS
Model C	[0.1,1] – 0.1	[0.1,1] – 0.1	levelS

Results

Model A: basic properties

This model makes us identify a burst in the number of messages when applying only a probability to answer incoming messages: basically because the more responses there are, the more messages to answer to there are. In Figure 3, curve shows exponential tendency with respect to the probability to answer($r^2 = 0.994$).

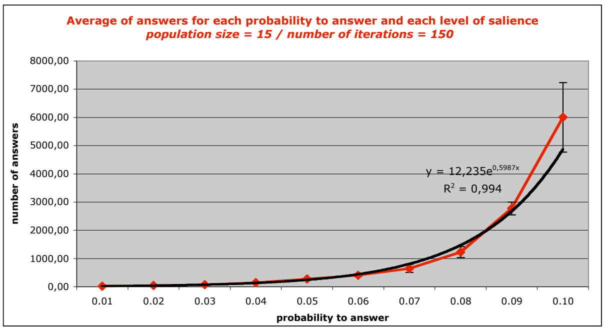

Figure 3: Average of answers for each probability to answer and levelS = 1, population size = 15, number of iterations = 150

In each of the curves presented in this paper are represented the standard deviation curve and the average curve.

Of course, the empirical system does not follow to such a tendency for many reasons: individuals lack of time, are not interested in all the subjects, delete sometimes some messages without reading them because there are too many messages in their mailbox… If we could observe this type of behaviour in real systems, it would be impossible for each agent to follow all the discussions. Thus, this probabilistic model doesn't account for the answering behaviour observed in a mailing-list: the generating mechanism of the burst following the simple rule: the more there are responses in the mailing-list, higher is the opportunity to answer. Therefore, it conducted us to propose more realistic mechanisms that would act as corrective process to inhibit this burst of messages.

Model B: Participating idiots or the decreasing wealth of the mailing-list

In this model, we used an equation instead of a simple variable to calculate the probability of sending an answer. Even difficult to argument, the basic behaviour is: if the knowledge contained in a message is lower than mine, the agent decides to not answer, else the agent answers with the probability *(Kmess-K)* that is the less knowledge agent has, the more chances there is for him to participate. Such a behaviour is not observed for every participants in a mailing-list but everyone experienced such a situation where discussion on the list became so stupid that it inhibits ourself to participate.

In Figure 4, we observe that the data fit to a logarithmic curve; the number of answers increases slowly for each level of salience at a controlled rate.

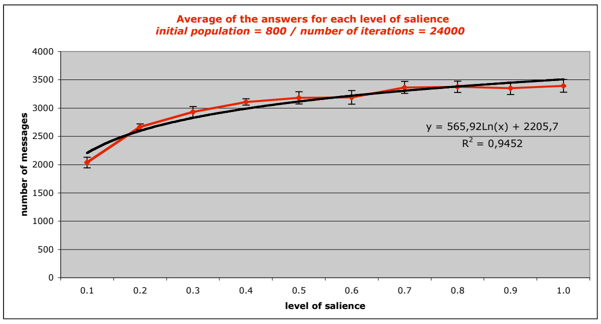

Figure 4: Average of answers for each levelS
population size = 800, number of iterations = 24000

Knowing that the level of knowledge of the new message (answer to a message) is still comprised between 0 and the level of knowledge of the author of the anwser. Our basic behaviour controls the burst observed in model *A*. As only agents with a lower level of knowledge answer to messages leading to a decrease in the level of knowledge of the posted messages. Such a behaviour excludes from the discussion the agents that have a higher level of knowledge. One can notice also that the increase of knowledge at the population level is naturally very low. The consequence of such a situation is the decrease of the knowledge content in the messages (Figure 5), which could lead in a real situation individuals to unsubscribe because exchanges are less interesting and because they do not acquire knowledge any more.

levelKmess of the initiating post	levelKmess of the answers (chronological order)
0.9245081	0.26937804 0.21386528 0.06880118 0.03542916 0.057888873 0.0398916

Figure 5: Example of a discussion
population size = 50, number of iterations = 100

Such a situation even noticed sometimes, is not as frequent as the inverse one corresponding to the case that to answer to a message, an agent must have a level of knowledge superior to the one contained in the message. The rule is "every contribution must be useful". That's the reason why we decided to to test this new hypothesis (model C).

Model C: Participating elites or increasing the wealth of a mailing-list

For this model, an agent can only answer to a message if he has something to bring to the discussion, that is if his level of knowledge is superior to the level of knowledge of the message he answers to. Moreover, he has to bring something new, so by comparison with previous version where the knowledge contained in the answer was drawn at random between 0 and the knowledge of the author, in this version, the level of knowledge of the answer is drawn at random between the level contained in the message he answers to (basically he can't do worth) and his own knowledge.

We then obtain a reasonable growth in number of messages as we can see in Figure 6.

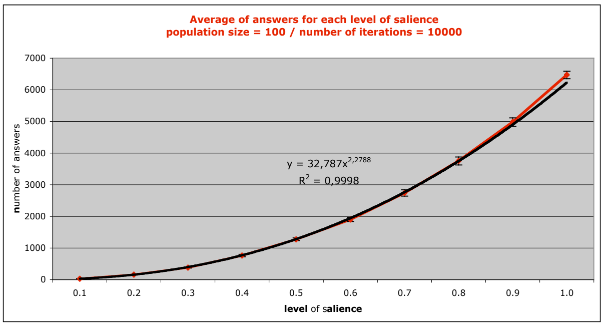

Figure 6: [Kmess,Kmemb] Average of answers for each level of salience
initial population = 100, number of iterations = 10000

In this model, only the agents who have a higher level of knowledge can answer. Thus, the level of knowledge contained in the messages becomes higher and higher and the agents with a low level can't answer them. But they acquire knowledge and thus are more susceptible to participate in the future. The consequence of a situation like this is the increase of the wealth of the messages sent to the mailing-list (Figure 7).

levelKmess of the initiating post	levelKmess of the answers (chronological order)
0.23061639	0.5001727 0.81811464

Figure 7: Example of a discussion
population size = 100, number of iterations = 1000

If we modify the existing model considering that by answering to a message, an individual can do worth and bring less information in the answer than the information contained in the previous message; i.e. if the knowledge contained in the answer is drawn at random between 0 (instead of *Kmess*) and *levelK;* we obtain the same tendency (Figure 8).

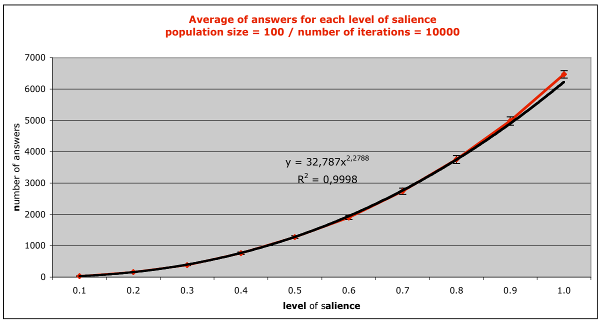

Figure 8: [0,Kmemb] Average of answers for each level of salience
population size= 100, number of iterations = 10000

CONCLUSION

The model A allows us to see the limit behaviours of our collection of models concerning the number of answers. We observed a burst in messages due to the following rule: the more responses there are, the more messages to answer to. The analysis of the results lets us suppose that the change of the static variable by different equations taking into account the level of knowledge contained in the message and the level of knowledge of the agents allows controlling this burst. This hypothesis is verified (with models B and C).

But, we have to remark that even the hypotheses we took for model B and C, can be realistic in some cases cannot be applied systematically, and surely the case where population is heterogeneously composed of agents having one or the other behaviour has to be studied. Another step consists in considering the reputation effects in the behaviour of the individuals on mailing-lists in order to concentrate on the social dimension of the phenomenon.

Of course we also plan to consider several subjects (the levels and the types of participation can vary from one subject to another one for a unique agent).

REFERENCES

Amblard, F., Ferrand, N. and Hill, D.R.C., "How a conceptual framework can help to design models following decreasing abstraction", Proceedings of SCS 13th European Simulation Symposium, Marseille, octobre 2001, p. 843-847, 2001.

Amblard, F. and Phan, D. 2006. *Modélisation et simulation multi-agents pour les Sciences de l'Homme et de la Société : une introduction*, Hermès, à paraître, 2006.

Blanchard, A. and Markus, L. 2002. "Sense of Virtual Community-Maintaining the Experience of Belonging". *Proceedings of the 35th HICSS Conference*, Hawaï.

Gilbert, N. and Troitzsch, K. 1995. *Simulation for the social scientists*, Open University Press.

Lindenberg, S. 1992. *The Method of Decreasing Abstraction*. Coleman & Fararo, p. 3-20.

Manzo G. 2005. "Variables, mécanismes et simulations. Une combinaison des trois méthodes est-elle possible ? Une analyse critique de la literature". *Revue Française de Sociologie*, vol.46, n°1.

Millen, D. R. and Dray, S. M. 2000. "Information Sharing in an Online Community of Journalists". *Aslib Proceedings*, vol. 52, n°5, p. 166-173.

Rheingold, H. 1993. *The Virtual Community: Homesteading on the Electronic Frontier*, Massachusetts, Addison-Weslley.

Schubert P. and Ginsburg M. 2000. "Virtual Communities of Transaction: The Role of Personalization in Electronic Commerce". *Electronic Markets Journal*, vol. 10, n°1, p. 45-55.

Stanoevska-Slabeva, K. and Schmid, B. 2001. "A Typology of Online Communities and Community Supporting Platforms". *Proceedings of the 34th Hawaï International Conference on System Sciences*, Hawaii, USA.

Tung, L., Tan, P., Chia, P., Koh, Y. and Yeo, H. 2001. «"An Empirical Investigation of Virtual Communities and Trust". *Proceedings of the 22nd International Conference on Information Systems*, p. 307-320.

Wellman, B. 1997. "An Electronic Group Is Virtually a Social Network". *Culture of the Internet*, (Kiesler, S. ed.) Lawrence Erlbaum Associates, Mahwah, NJ, USA, p. 179-205.