Alma Mater Studiorum – Università di Bologna

PhD Program in Business Administration

Ciclo XXI

Sector: SECS-P/10

Title

Evolutionary Dynamics of Coordination-Communication Networks In Open Source Development

PhD Candidate:

Vitaliano Barberio

Board:

Prof. Alessandro Lomi

Prof. Raffaele Corrado

Dott. Guido Fioretti

Phd Coordinator:

Federico Munari

Esame finale anno 2009

INDEX

THESIS OVERVIEW

Research Idea

This research deals with the issues of **representing, detecting and understanding the**

evolutionary dynamics of communication networks, as a means of coordination, which enable both traditional and distributed production processes over the time. In terms of general research questions I ask: (i) How should we theoretically represent coordination-communication patterns in modern complex production? (ii) How a network representation of coordination helps us to detect organizational structures in fuzzy bounded and fast changing environments? (iii) What is the impact of institutionalized structures on communication-coordination networks? (iv) How communication practices, institutionalized in *genres,* dynamically enable and constrain coordination patterns?

Why this issues became generally interesting today even more than in the past? This is neither the place nor the time for a socio-philosophical reflection on how innovation in ICT technology changed the human life in the last few years. However let us suppose that tomorrow we will wake-up and the Internet, for some reason, will be totally disappeared. It is not so difficult to predict a generalized chaos. In other words, it is undeniable that today we live in a 'connected' world and the Internet enables a significant part of this connectivity. At a minimum level, we use emails at work, Voip (Skype) and chat devices for Internet calls, but at a major level of involvement, a lot of people also search for, exchange and share an impressive amount of information every day.

We should acknowledge that since the diffusion of the Internet in the early nineties[1], the scope of business possibilities dramatically changed. On the one hand, let us think to very traditional businesses indeed, like chemicals or automotive, whose innovation system needs are today complemented recurring to on-line marketplaces for solutions and intellectual assets like yet2.com[2]. On the other hand, Internet made possible the emergence of worldwide

1 We can date the birth of the World Wide Web (WWW) back to 19991. http://en.wikipedia.org/wiki/Internet

2 Founded in 1999, yet2.com is focused on bringing buyers and sellers of technologies together so that all

distributed user communities for many products and technologies. Producers today get product user feedbacks very often and very fast if compared to just few years ago. Moreover users exchange information and beliefs among themselves gaining momentum as communities when a collective opinion emerges. Producers have progressively made users more involved in product development[3] and customization until the distinction between producers and users is now more fuzzy than in the past.

Along the change for traditional businesses new ones emerged contrasting many traditionally accepted economical principles. Who actually could imagine ten years ago the appearance of a market for 'Internet-search engines' and that the leader of such a market – Google – would have five billion dollars revenues in a quarter of year? Moreover who could imagine that these revenues are earned by Google without asking its users to pay for the core developed technology (search engine)? Finally, we have new innovation-production models like the one of Open Source Software, where geographically diffused communities of users-developers create and maintain, on a voluntary base, high quality technologies never or very rarely meeting each other face-to-face. These collective development efforts represent a combination of all the innovative characteristics mentioned above: (i) the use of Internet as a means of knowledge sharing; (ii) the fuzzy boundaries between producers and users; (iii) the free distribution of core products.

Why it Is Important for Economics and Management Science?

Recent innovation in ICT made socio-economic systems more connected and dynamic rising

parties maximize the return on their investments on intellectual assets.

[3] The book Wikinomics contains a lot of stories of user communities involvement. See www.wikinomics.com

the complexity of coordination activities for organization that operate within them. The way in which people exchange knowledge is changing productive structures and innovation systems. It seems that technological advances are vindicating the forward-looking vision of porous, boundary-less and networked organizational form that light-handedly coordinates the activities of heterogeneous, geographically diverse team members (Weisband, 2008).

Using 19.9 million papers over 5 decades and 2.1 million patents Wuchty, Jones and Uzzi (2007) empirically demonstrated that teams increasingly dominate solo authors in the production of knowledge. If the way in which knowledge is produced and reproduced it is changing then, it is expectable that the way in which work is organized it is changing as well. Ancona et al (2002) both surveyed and observed 169 teams through 6 industries and over a time period of 2 years looking for some correlation between team characteristics and team performance. They found that successful teams, which they call *x-teams*, have distinctive traits, like fluid membership, expandable tiers and extensive (in-in and in-out) connectivity, which contrast with the traditional rules handed-out by best-selling books. The current environment – with its flatter organizational structures, interdependence of tasks and teams, constantly revised information and increasing complexity – requires a networked approach. X-teams have emerged to meet that need (Ancona et al. 2002).

Despite the importance of communication for new models of distributed work and innovation, very few attention has been payed to its forms and patterns in modern complex organizations. Since Barnard (1938) recognized the importance for communication channels within big corporations those channels have been mainly considered as an informal parallel structure and researchers mainly focused on the formal side of coordination. In a recent paper some authors (Kleinbaum et al. 2009) used millions of email messages, calendar meetings and teleconferences for may thousands of employees of a single multi-divisional big company over a three months time period, in order to explain the emergence of communication networks. They essentially confirmed that formal hierarchies are the most important driver for communication channels in a big corporation. It has also been claimed that a communication theory of the firm should consider the evolution of communication networks in organizational communities (Monge et al., 2008).

However the multi-divisional big corporation today is not the only one available productive form and with the progressive increase of external connectivity, it is going to overlap more with different (non corporate) entities affecting its technological core. A general communication theory of the firm should also consider new productive forms, like virtual communities, where formal structures are barely implemented but collective decision are made efficiently. Addressing the evolution of these organizational fields (Powell et al. 2005) in which complex knowledge exchange links together diverse and distributed teams is the general contribution of this thesis to a potential communication theory of the firm.

From my point of view, which is on coordination of product-development tasks, a central issue for economics and management today is: **how the use of IT mediated communication both enable and constrains the boundaries of production possibilities.** According to my perspective communication is an enabled structure of coordination among interdependent tasks which production is arranged in. The more productive tasks and resources are distributed across different agents, the more communication will be a relevant means of coordination. Given this assumption the choice of the Open Source Software seemed to me just a perfect empirical setting where to test my arguments. In fact as I mentioned, the most of open projects are mainly managed in a distributed way, that is by means of massive use of email communication.

I would like to make clear since now, that the Open Source development for this research is mainly an empirical setting which provided free-public communication data to be 'easy' collected. As I have been saying at very beginning of this overview, I will have three main objectives leading this research on coordination-communication dynamics: (i) representation; (ii) detection; and (iii) understanding (both predicting and interpreting). Hence, the general aim of this thesis is not to show how Open production-innovation models are, to some extent, superior than more traditional ones. **This research is rather intended as an effort to build and test a general research strategy, for both the intelligence (detection) and the prediction-interpretation (understanding) of communication-coordination processes, coherently with a network theoretical framework on productive organizations (representation).**

The exposition of contents will be articulated in four sections reflecting my thesis' objectives. These four sections, are intended as self-containing, then: the fist one will provide a general theoretical framework for the construction of research questions and for the interpretation of further results; the second one will deal with the problem of finding organizational structures using empirical communication data; both the third and the fourth ones should be considered as two complementing steps of understanding structures. They focus respectively on exploring the impact of institutionalized structures on communication (third) and on a more dynamic point of view on communication practices (fourth). For this reason each part will have its distinct but related research strategy. The contents for all the three sections are briefly introduced below.

Representation (section one)

Economic theory has been representing production since the end of 19^{th} century. My work will begin with an issue for economic and management theories today: *how good traditional representations of production are in order to model communication dynamics as enacted coordination structures in organizations over the time?*

Mainstream economics used, as a main representation tool, the 'production function' (Cobb-Douglas, 1928; Arrow et al., 1961) which transforms given input-factors in outputs. In order to satisfy tight simplifying assumptions, like decreasing marginal productivity and convex technologies, Neoclassical production functions lay very far from offering a representation of how actually production is arranged within and among organizations (Simon, 1991; Landesmann and Scazzieri, 1996; Lomi, 1977). Then, I will review some research in organization theory which paid more attention to interdependencies within and among productive units, focusing on its potential contribution to the issues of coordination, communication and distributed development. Finally in this section I will present a short literature review on open source development focusing on those contributions dealing with distributed development and communication networks. In this paragraph I will both look backward to organization theory and look forward building research question to be explored and eventually tested in the second section of this thesis.

Detection (section 2)

The second section of this thesis deals with the issue of community structures detection and inspection. Assuming that communication networks mirrors underlined interdependencies to be coordinated, the research questions leading this section are: *can we find organizational principles leading the emergence of community structures using email communication empirical data? How can we partition a wide and fuzzy-bounded community in smaller groups approximating traditional organizational units? how organization principles work at this tighter level of analysis?*

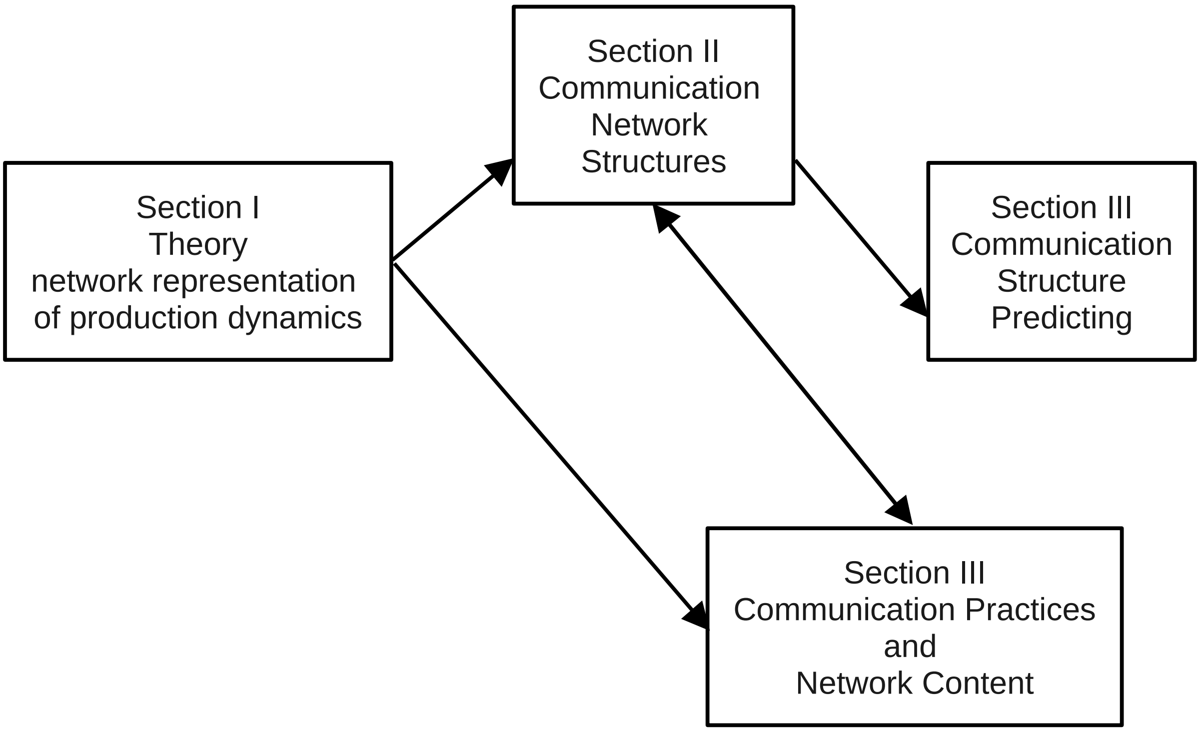
These questions in summary state a kind of null hypothesis which has to be rejected before every further research effort. **This null hypothesis is: communication networks are random graphs or, in other words, it is impossible to find any organizing principle.** As a first step I will present how I collected and organized data. Then in order to explore these general questions, I will use some metrics of structure for describing the topology of a very wide community of production belonging to the Apache web server project. This project as a whole is a collective development effort involving (with variable intensity) thousands of programmers who wrote almost a million of emails over a ten years time period (1995-2004) in order to coordinate their work. In spite of a widely accepted vision of self-organization dynamics leading the emergence of structures in virtual communities, I found that formal institutional arrangements could have been playing an organizational design role of simplification for coordination complexity. Assuming that organization structure mirrors product structure (Shilling and Mahoney, 1996) I expected a near decomposable (Simon, 1962 social structure. then using network clustering techniques (Newman and Givran, 2003), I will partition the whole community in sub-components and I will show how it is possible to extract a core component. This core unit interestingly corresponded to what is called a 'Platform' (to which modules are added) in software modularity literature (Rusnack et al., 2006; Baldwin and Clark, 2006; Haefliger, et al, 2008).

Understanding (section 3 and section 4)

Section 3 (predicting structures). Using as a new sample the network cluster extracted in the previous section I performed more refined estimations for the emergence of communication networks. Namely I used MQUAP regression models for dyadic data (Krackhardt, 1986, 2007). More precisely I will take a communication network as a kind of dependent variable an test for some structural attribute of nodes potentially explaining the emergence of links within it. Because of the experimental stage of this field of research no consolidated strategies for data aggregation and analysis are consolidated. Hence, this section concludes with an interpretation of results and a proposal for the extension of the regression models based on both a different level of data aggregation and the extraction of further attributes.

Section 4 (interpreting communication contents). In the former section I studied the *structure* of communication networks as repeated interaction tracing institutional dynamics. In this section I will take a more dynamic perspective on institutions. The overall question for this section is: how communication works in practice linking both material and symbolic constituting elements of production processes? How sets of communication practices, institutionalized in *genres*, lead decisional processes toward given objectives?

I will use semantic network analytical for a short case study exploring the semantic content of organizational communication networks. This study aim to show how communication practices, condensed in genres, work as a powerful and flexible structure (culture) which co-evolve along with a more material structure of product modification (action). More precisely, I will build networks whose nodes are concepts referring to agents, resources, tasks, organizations and so going forth. This will be for two distinct networks corresponding to distinct discussions. The first one will be about working on the software code.



SECTION I

REPRESENTING PRODUCTION: AN ISSUE FOR ECONOMIC THEORY

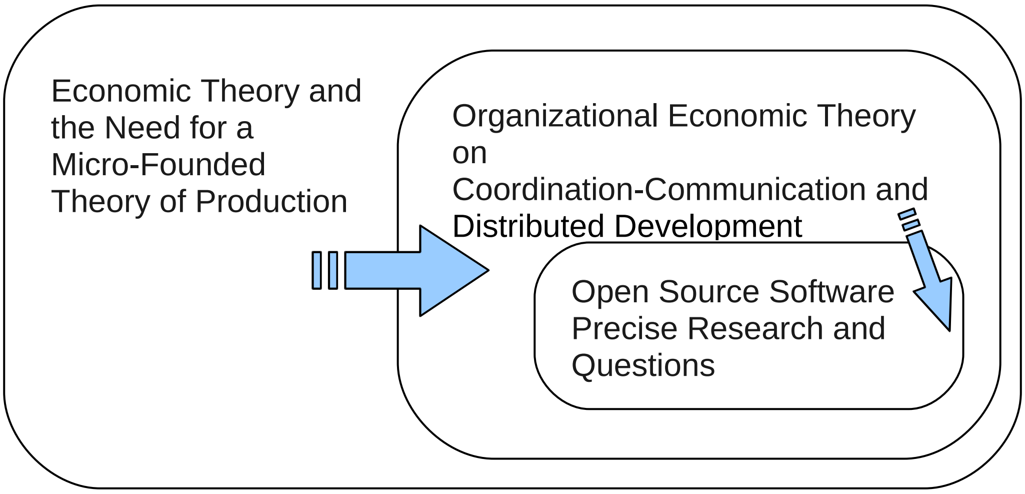
Introduction

As argued in the overview of this thesis, this section construct a theoretical issue of production dynamics representation for economics and management today. Innovation in information and communication technologies is increasing the connectivity of agents in social and economic systems. As a consequence, more complex problems of coordination arise for which traditional organizational solutions seem increasingly inadequate. Despite this trend toward increased connectivity and interdependence, engineering and economic models of production are still mainly rooted in simple input-output analytical frameworks. The overall research question for this section is: *which actually could be the contribution of recent studies on networked virtual communities to the 'eternal' debate on the need for an organizational representation of production?*

Neoclassical economic theory, which is still the mainstream in economics today, have been representing production processes with the same tool – the Production Function – . Since its fist empirical estimation in 1928 the so called Cobb-Douglas production function remained unchanged or barely changed (Arrow et al, 1961). A number of controversies have been posed to the validity of Neoclassical production theory, which could be summarized in the absence of micro-foundations (Georgescu-Reogen, 1953; Sraffa, 1961; Simon, 1957). In particular it seems that very few attention has been payed, from Neoclassical economics, to how the elements of productive dynamics (materials, tasks and agents) are arranged in interdependent components or units (Landesman and Scazzieri, Padgett, 2003). Mainstream economics never clearly articulated the issue of productive organization's ontology, that in other words is, it never developed an organic theory of the firm.

Hence, in order to give a coherent frame to interpret my own contribution, I will shortly introduce some alternative point of view on the ontology of productive organizations. In this not exhaustive review I will try to highlight how different theories of the firm motivated coordination structures, then also communication, as a response to uncertainty. In particular Contingency Theories, Knowledge Based theories, Population Ecology Theory and Theory of Practice. will be considered for its contribution to my ideas on coordination and communication in traditional and distributed production.

Because of its relevance for the understanding of new organizational forms relaying on Internet as a means of coordination, the issue of distributed production will be deeper constructed in the last paragraph of this section. In particular, I will conduct a further literature review on Open Source Software development paying particular attention to those contributions that focused on communication and other governance structures. In this last paragraph of the section one, I will highlight some research question to be empirically either further explored or tested in section two.



Economic Production Theory and Functions

With the marginal revolution, Jevons, Menger and Walras developed pure exchange models in the 1870s that shifted the explanation of prices away from the classical difficulty-of-production focus to the neoclassical focus on utility and relative scarcity. Adam Smith's diamond-water paradox was no longer a paradox, since price was explained as proportional to marginal utility, which depended on scarcity (Cohen and Harcourt, 2003). Neoclassical capital theory was the arena for extending the general principle of relative scarcity to explain all prices, including factor prices in models with production and time (Hennings, 1985). This paragraph presents an incomplete summary of production function's main characteristics as used by neoclassical theorist (Cobb-Douglas, Arrow, Solow, Samuelson, Hicks, among the others).

The Production Function has been used as an important tool of economic analysis in the neoclassical tradition whose general question has been about the rational decision on the optimal level of production for a firm, an industrial sector or an economical system, according to both available productive factors and technologies. It is generally believed that Philip Wicksteed (1894) was the first economist to algebraically formulate the relationship between output and inputs as $Y = f (x_1 , x_2 ,..., x_n)$. Then, whatever Production Function, of this shape, aims to formally represent the laws according to which some input factors (like capital and labor) are transformed into an output (goods and services). The available technologies of transformation are thinkable as recipes for inputs' combinations.

Theories of production are important components of general economic theory because the representation issue they accomplish should provide a micro-foundation for general models of growth (like Samuelson, 1962) and technological change. The production technology for the one-output/two-inputs case is depicted in Figure 1. Output (Y) is measured on the vertical axis. The two inputs, which we call L and K which can be thought as labor and capital, are depicted on the horizontal axis. In the Neoclassical theory of production, *all capital is assumed to be endowed*. The hill-shaped structure depicted in Figure 1 is the *production set*. Notice that it includes all the area *on* the surface and *in* the interior of the hill. The production set is essentially the set of technically feasible combinations of output Y and inputs, K and L.

A *production decision* is a feasible choice (a recipe) of inputs and an output is a particular point on or in this "hill". It will be "on" the hill if it is technically efficient and "in" the hill if it is technically inefficient. Properly speaking, the *production function* $Y = f\ (K, L)$ is *only* the surface (and not the interior) of the hill, and thus denotes the set of *technologically efficient* points of the production set (the maximum producible output). As normally assumed by Neoclassical economists: Let there be m factors of production and let vector $x = (x_1, x_2, ..,$

x_m) denote a bundle of factor inputs. We shall define an *input space* as the acceptable set of inputs for our economy.

Commonly, a bundle of factor inputs x is deemed "acceptable" if every entry in that vector, i.e. the quantity of every factor, is a non-negative, finite real number. Thus, any input bundle x lies in $\mathbb{R}^m{}_+$, the non-negative orthant of m-dimensional Euclidian space.

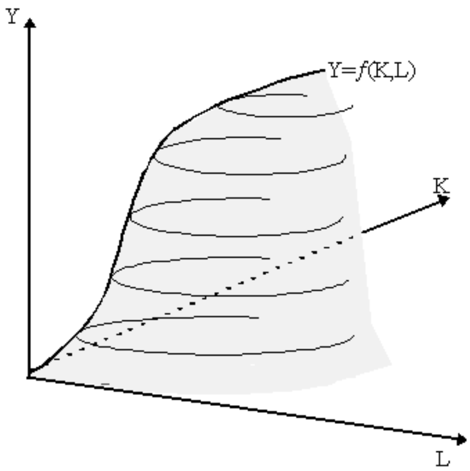


Figure 1. Production function for one-output/two-inputs.

Thus, $\mathbb{R}^m{}_+$ is our input space. Let y be output, which is assumed to be a single, finite number, i.e. $y \in \mathbb{R}$. Thus, a production function f maps acceptable input bundles to output values: f: $\mathbb{R}^m{}_+ \rightarrow \mathbb{R}$. More specifically, $f(x)$ is the *maximum* output achievable for a given set of acceptable inputs, $x \in R^m{}_+$. A deeper examination of Neoclassical Production function's properties and conditions is offered in appendix I.

Cobb-Douglas production function

It is generally believed that Philip Wicksteed (1894) was the first economist to algebraically formulate the relationship between output and inputs as $Y = f(x_1, x_2, ..., x_n)$. Then, whatever Production Function, of this shape, aims to formally represent the laws according to which some inputs are transformed into an output. The Cobb-Douglas (1928) formulation of the production function is still today the most ubiquitous representation in theoretical and empirical analyses of growth and productivity. A notable change in his formulation of production function came only in 1961 – after a gap of 33 years – with the work of Arrow, Chenery, Minhas and Solow (1961), which, however, is only an extension, not an alternative paradigm.

The classical formulation of the Cobb-Douglas (1928) is: $Y = A K^{\alpha} L^{\beta}$, where Y is the total production (value of production); A is the total factor productivity (a constant); K and L are respectively capital (profits) input and labor (wages) input; and α and β are the output elasticities of labor and capital, respectively (parameters to be estimated). These values are constants determined by available technology and their total effect is equal to one ($\alpha + \beta = 1$). The estimation of the parameters of aggregate production functions is central to much of today's work on growth, technological change, productivity, and labor. Empirical estimates of aggregate production functions are a tool of analysis essential in macroeconomics, and important theoretical constructs, such as potential output, technical change, or the demand for labor, are based on them (Felipe and Adams, 2005).

Around 1920s, Douglas was studying the elasticities of supply of labor and capital, and how their variations affected the distribution of income. To make sense of and interpret the numbers obtained, Douglas needed a theory of production. Assuming neoclassical restrictions like exogenous given resources and technology, constant returns of scale, diminishing marginal productivity and competitive equilibrium, a 'simple' way to empirically estimate the Cobb-Douglas production function was the statistical fit of a least square regression for the α and β coefficients in the logarithm form of the equation with Y, K and L measured with empirical data. The OLS model estimated by Cobb-Douglas (1928) with data from the American manufacturing sector (1899-1922) has been taken as an implicit validation for the existence of the aggregate production function, as well as for the validity of the marginal productivity theory of distribution (Felipe and Adams, 2005).

The practice of estimation for production functions was started and a new generalization of the production function, the so called Constant Elasticity of Substitution (CES) was proposed by Arrow et al. (1961) for which the Cobb-Douglas (perfect substitution of input factor) and Leontief (1941) (zero elasticity of substitution) were two extreme cases. This attempt was also an further step in the consolidation of aggregation practice. Aggregation means that a global (sector- economy) production function could be modeled by an additive function where Y, K and L are the simple sum of production capital and labor for a sector or a whole economy.

Main Critiques to Neoclassical Production Functions

Aggregation critique. Even if Cobb-Douglas-like *aggregate production functions* are the pillars of neoclassical growth models (Solow, 1956, 1957), the significance of estimated productive function is a very turbulent area in the economic debate. The early results of estimated production function (Cobb-Douglas, 1928) were accepted with a certain hostility (Cobb, 1967). Commenting those results Samuelson reveled multicollinearity problems, the presence of outliers, the absence of technical progress, and the aggregation of physical capital. Then, all the aggregate Cobb-Douglas function regression captures is the path of the value added accounting identity according to which value added equals the sum of the wage bill plus total profits.

This problem was actually already highlighted by Simon and Levy (1963) concluding that the existence of a fitted Cobb-Douglas function with a value of with a value of α (obtained by fitting the general equation $Y = A\ K^{\alpha}\ L^{\beta}$ to empirical data on Y, K and L) in agreement with the actual α (labor's fraction of total product) does not imply that the underlying function is truly Cobb-Douglas. The general critique of aggregation in production function resulted in a consistent corpus of works named aggregation literature whose main exponent is Franklin Fisher (see Fisher 1993). In a recent work from this literature Felipe and Adams (2005) repeated the estimation of Cobb-Douglas with the same finding that the results are particularly poor when adding an exponential factor accounting for the technological change over time. Moreover those authors (Felipe and Adams, 2005) confirmed that:

> *"An algebraic transformation of the identity, under the appropriate assumptions about the data, yields a form that resembles a production function. This implies that if the correct form of the identity, written as a production function, were fitted, one should always conclude that the aggregate production function exhibits constant returns to scale, and that factor markets are competitive... The important aspect of this argument is that it can parsimoniously explain why, despite the fact that aggregate production functions do not have a sound theoretical basis, they appear to yield meaningful results at times"*.

The conclusion is that neither the existence of the aggregate production function, nor the standard neoclassical hypotheses of constant returns to scale or competitive markets, can be tested empirically since they cannot be refuted.

Capital Theory Critique. Another critique to the Neoclassical production theory is the one of the so called *Cambridge controversies*. This name reflects that two competing points of view on economic production have been very animatedly debated between neoclassical theorists (Solow, Samuelson, Bliss), Mainly rooted in Cambridge USA, and Neo Ricardian or Sraffian (Robinson, Sraffa, Pasinetti) economists, from Cambridge UK especially during the period 1950s-1970s. Joan Robinson (1953) viewed aggregate production function with a remark:

> *"... the production function has been a powerful instrument of miseducation. The student of economic theory is taught to write Q = f (L, K) where L is a quantity of labor, K a quantity of capital and Q a rate of output of commodities. He is instructed to assume all workers alike, and to measure L in man-hours of labor; he is told something about the index-number problem in choosing a unit of output; and then he is hurried on to the next question, in the hope that he will forget to ask in what units K is measured. Before he ever does ask, he has become a professor, and so sloppy habits of thought are handed on from one generation to the next."*

A problem here seems to arise with the measurement of capital. As argued by Wicksell (1911) long ago, heterogeneous capital goods cannot be measured and aggregated in physical units; instead valuation must be used. The evaluation of goods could be linked to the cost of production or to the present value of the future production. Both of these two criteria involve time then presume a rate of interest. A problem of circularity arises when in the general models, like the one of Samuelson (1962), the rate of interest is determined in a one way manner by the quantity of capital.

Resources Critique. A famous justification given by a recognized voice from neoclassical theory in defense of aggregated production functions is that: “until thermodynamic lows are not violated I will continue to add production functions” (Solow, 1978). However in my opinion, a fundamental point is that thermodynamic laws indeed are violated by Neoclassical assumptions for production functions. Since the production function is often explained as a technical recipe, following (Daly, 1997):

> *“we might say that Solow's recipe calls for making a cake with only the cook and his kitchen. We do not need flour, eggs, sugar, etc., nor electricity or natural gas, nor even firewood. If we want a bigger cake , the cook simply stirs faster in a bigger bowl and cooks the empty bowl in a bigger oven that somehow heats itself” ... “Furthermore we can make not only a cake, but any kind of dish – a gumbo, fried chicken, a paella, bananas foster, cherries jubilee – all without worrying about the qualitatively different ingredients, or even about the quantity of any ingredient at all!”*

Georgescu-Roegen (1951, 1975,) who was an economist but also an engineer and a physicist, claimed for some attention to be payed in economic theory to the issue of resources undergoing production. Solow-Stiglitz replied, to this critique with a new production function accounting for 'natural resources' alongside with Capital and Labor (Solow et al., 1978). However Georgescu-Roegen again (1979) further clarified that he wanted to rise an issue for economic theory not just about the consumption of a new productive factor (Materials) but rather on the recognition that in real-world productive factors enable production entering in complex nexus of interdependence. As far as I know, this controversy has still not been addressed by Neoclassical Economics until today.

Optimizing Rationality vs Bounded Rationality. When talking about critiques to the

Neoclassical thought a prominent position have been assumed by Herbert Simon. Simon's thought go to the deepest micro argument of neoclassical theory – the rationality of economic agents – for which traditional neoclassical economics relies on very thigh assumptions of maximizing behavior and perfect information. In his seminal contributions, Simon (1946, 1957) stated that in decisional processes: (i) not all the possible actions are known ex-ante; (ii) not all the consequences of eventual actions are clearly definable; and (iii) not all the preferences are unequivocally comparable. Then a moderation of the Neoclassical rationality what he ironically defined 'Olympic' at his the Nobel Lecture (Simon, 1978), comes with the acceptance of satisfying (rather than maximizing) behavior of economic agents. As we can see in the next paragraph this assumption is of fundamental importance for motivating the ontology of whatever organization.

An Alternative Point of view on Economic Production

Neoclassical economists never convincingly addressed all the above mentioned critiques (Silos Labini, 1985) resulting in economic production to be one of the most problematic areas of economic theory ever. Because of the lack, in mainstream economics, of a micro-founded 'realistic' theory of production there is a possible contribution for my research, which goes beyond the emerging literature on virtual communities, embracing the general theme of organizational ontology.

My approach to the above mentioned internal (to economic theory) critiques is rather constructive. In fact I think that a valid integration should be provided to standard production economics that could address some open issue and maybe pose further issues for economics today. As mentioned before the main issue for economic theory of production is one of organization of interdependent productive factors. Hence an initial insight has been provided by the emergence in contemporary economic theory of a possible link with organizational theory. This link is created by a stream of researches continuing both the Sraffa's and Georgscu-Roegen's critiques to neoclassical theory of production. In particular way the work by Landesman and Scazzieri (1996) proposed a representation of production dynamics as co-evolution of multiplex networks. Those networks are essentially: (i) a network of materials undergoing production dynamics; (ii) a network of tasks to be performed; (iii) a network of agents who perform tasks. Co-evolution means that during production, that is product modification operated by agents, the productive structure change itself. Inf fact Products, flowing through productive agents, build an organizational structure which in turn learns and changes by means of knowledge embedded within them (Powell et al, 1996; Padget, 2003).

A study which uses network representations of production over time could also contribute the methodological issue of shifting from a static representation of networks to the so called network dynamics (Smith-Doerr and Powell, 2003). Network Dynamics, when compared with traditional 'social network' is more concerned with factors influencing the emergence and the evolution of networks with: (1) no limitation to just one kind of nodes; (2) no limitation to just one kind of relationship among nodes; (3) no limitation to just small (under fifty nodes) networks; and (4) finally recovering the symbolic-semantic dimension of social interaction (Pattison and Carley, 2003). From the above considerations a proposition on the eventual contribution of this thesis follows:

Proposition 1: *with the aim of providing micro-foundations to general production theory,* *economic production dynamics can usefully be represented as a co-evolutionary process with multi-mode (different kinds of nodes) multiplex networks (different kinds of relations).*

In my opinion, an important contribution to the extension of this micro-economic level representation could be offered from organizational theories. In particular organizational theory hold a consolidated tradition of research on factors shaping structures and vice-versa. In the following paragraph I will review some classical and recent papers in organization theory focusing on its contribution to the argument of representation for today's networked and communicative organizations.

Foundations for an Economic Organizational Theory of Production

Organizational Rationality under Uncertainty and Interdependence

Chester Barnard (1938) proposed that the ontology of organizations (and firms) resides in individual's limits to accomplish complex objectives. He stated that although the study of economic organization deals principally with markets, a grate deal of economic activity takes place within firms. He observed that managerial functional hierarchies are powerful means of coordination when organizational environment is characterized by complexity and change. The argument of a 'visible hand' (managerial coordination) doing organizational adaptation working, opposed to the 'invisible hand' (of market) as a means of natural selection, has been central in the further functional analysis of north American mass-production corporate enterprise (Chandler, 1962, 1977).

Barnard (1938), who was a telecommunication executive, intended managerial hierarchies as formal structures where both power and incentives leads coordination toward an equilibrium in a Paretian fashion. But what is more interesting for my research theme, is his recognition for the existence of communication structures within firms as a parallel informal mechanism of coordination whose importance grows going up along formal hierarchies. Considering the above a second proposition follows:

Proposition 2: *hierarchical coordination and the communication it implies is very central for the existence of organization.*

Herbert Simon (1951, 1991) whose work surely has been influenced by Barnard's thought, argued that Economics rather than with markets should represent organizational dynamics. He also (Simon and March, 1958) argued that the ontology of organization lays into a superior information processing ability of hierarchies, when compared to markets (Hayek, 1945). Simon's thought depart from the Barnard's one when reasoning about hierarchies. According to Simon, those are Information processing structures where some task must be performed before in order to enable other tasks completion. Coordination is the way in which organizations cope with uncertainty required to integrate interdependent tasks in presence of both uncertainty and bounded rational agents (Simon, 1951).

I think it is important to acknowledge that Simon's representation of organizations goes beyond the separation between the formal side and the informal side of structures, stating that communication is thinkable as an information exchange process either enabling or constraining the possibilities of action. James Thompson (1967) argued that organizational structures should be designed in order to protect organizations' technological core from environmental uncertainty. Uncertainty is a bi-dimensional factor, that is both technological and institutional, determining different degrees of complexity for the task environment to be coordinated from agents. As an example we can think policies[4] (institutions) and digitalization (technologies) are introduced in universities (organizations) in order to avoid the distraction of professors (agents) from the core activity (research and teaching).

According to Thompson's (1967) influential work, technological interdependencies may assume three basic forms. Arranged in an increasing degree of complexity, the forms of interdependence are: (i) pooled (for mediating technology); (ii) sequential (for long-link technology), and (iii) reciprocal (for intensive technology). For each form of interdependence Thompson proposed a dedicated way of coordination which in turn shapes the emergence of structures. Arranged by an increasing cost of implementation the forms of coordination are: (i) categorization and standardization, leading to flexible structures changing by near independent units' reproduction or elimination; (ii) planning, leading to vertical integration of interdependent units; and (iii) mutual adjustment leading to high integrated horizontal structures of communication and coordination (Thompson, 1967).

Because of the prototypical nature of Thompson's reasoning on interdependence and coordination, it is not dependent on historically contingent communication technologies or institutions, then his ideas still stand as an important starting point for whatever construction on organizational dynamics. Directly building on Thompson's ideas two proposition for my theoretical construction follow.

Proposition 2: *in presence of rising levels of complexity for the task environment to be coordinated, communication networks (as a means of coordination) will change toward more complex forms.*

Proposition 3: *the adoption of organizational arrangements, in order to protect the technological core, intentionally simplifies the task environment leading toward lower of*

[4] Mohr's (2004) study on racial-diversity policies at the University of California is a possible example.

complexity for communication networks.

Production Networks as a System of Transactions and Contracts

Thompson's (1967) theory of organizational design has been very central to further developments of organizational economics. Building on the transactional-contractual view of the firm (Coase's, 1953; Alchian and Demsetz, 1972), Oliver Williamson (1975, 1985, 1991) proposed a Transaction Costs Economics (TCE) theory of design. TCE especially confronts with the issue of institutions intended as rules for managing transactions which are formalized in contracts. He especially confronted with the issue of vertical integration (Williamson, 1975, 1985), but in more recent works also proposed an organizational discrete analysis (Simon, 1978) based on institutions. Williamson offered a representation of productive structures boundaries based on neoclassical contract theory and property rights. Technological separable units are represented as transacting agents that try to minimize the cost of its transactions (the cost of using the market) by means of efficient coordination. He argued that both market and hierarchy are good systems of coordination (Williamson, 1985).

Distinguishing Attributes of Market, Hybrid, and Hierarchy Governance Structures*

	Governance structure		
Attributes	Market	Hybrid	Hierarchy
Instruments			
Incentive intensity	+ +	+	0
Administrative controls	0	+	+ +
Performance attributes			
Adaptation (A)	+ +	+	0
Adaptation (C)	0	+	+ +
Contract law	+ +	+	0

* + + = strong; + = semi-strong; 0 = weak.

Figure 2. Governance strructures and organizational features in TCE.

Markets are more efficient when prices embed all the information needed to take opportunism under control and for transactions to be completed. When the cost of a transaction using the market is higher than the cost of using hierarchy, that transaction should be incorporated within formal organization's boundaries (Williamson, 1985). Between these two polar forms of governance a number of hybrid combinations are also thinkable (Williamson, 1991). For representing each institution of governance (market, hybrid and hierarchy) a different institutional environment (rules for transactions like in North, 1986) is needed. Each governance form has attributes to be considered in organizational design, those are reported in the following table.

On the one hand, markets uses more price incentives for coordination, hierarchies are supposed to use more administrative control. On the other hand, while markets perform a kind of autonomous adaptation, hierarchies are supposed to perform more mutual adaptation and information exchange. The structure of transactions, that are the fundamental unity of analysis (Commons, 1934), is ultimately determined by: the uncertainty to which transactions are subject, the frequency whit which the transaction occurs and both the type and degree of asset specificity involved in supplying the good or service in question (Williamson, 1979).

Even if TCE formalized the argument of a continuum spectrum of feasible governance solutions for coordination it has been subject to several critiques. A first one possible point is the loss for attention to the nexus ox complex interdependencies coming with different kinds of technology (Thompson, 1967). A second point rises about the static nature of technology which is very in the spirit of Neoclassical Economics. Finally a third point rises whit the issue of institutions which are considered as mere systems of rules (North, 1978) losing the social evaluation interpretation given to this concept by precedent studies (Thompson, 1967; Berger and Luckman, 1966). All these issues has been central central research questions for different streams of research which are briefly presented in the following paragraphs of this section.

The Issue of Contingencies in Organizational Design

TCE offers only a 'residual' explanation for the ontology of organizational hierarchies, they are useful coordination tools when the market fails allocating productive factors. A more sophisticated explanation for the emergence of organization has been offered which is based on hierarchy's superior ability to manage complex information flows (March and Simon, 1958). In organizational design research, the most widely accepted interpretation, for the so called information processing view of organization (March and Simon, 1958) has been developed by Contingency Theorists (Lawrence and Lorsch, 1967; Galbraith, 1973; Daft and Lengel, 1986). A common argument of all these works is that matching increasing task uncertainty and complexity with less formal modes of coordination leads to better performance.

Thompson (1967), already mentioned the issue of contingencies in his theory of rational organizational design: these are knowable, but not designable (controllable), elements for a focal firm. From a more positivist point of view, Lawrence and Lorsch (1967) proposed that in order improve performance, organizations should pro-actively adapt (design structures) to given both internal (among units) and environmental (between a focal organization and the external environment) contingencies.

When focusing on communication, Contingencies Theory could offer an important contribution to my reflection which comes form information system design (Daft and Lengel, 1986). In one of the most quoted paper in organizational studies ever, Daft and Lengel (1986) proposed a design framework based on two contingencies – 'Uncertainty' and 'Equivocality' – affecting information processing in organizations. That is, a rational information system design reduces uncertainty by allowing organizations to access relevant information and by reducing the ambiguity of such information.

These authors also proposed a classification of media solutions matching the complexity of tasks to be coordinated with coordination tools to be adopted. According to this reasoning on the one hand, less rich and more informal media (like rules, formal information systems, special reports and planning) are better solutions for uncertainty reduction (obtain additional data, seek answer for specific questions). On the other hand, more rich and personal media (like direct contact, integrators and group meetings) are better solutions for equivocality reduction (clarify, reach agreement and decide which question to ask). A new proposition comes from the above:

Proposition 4: *the more uncertain is an interdependence among two or more tasks to be coordinated, the bigger is the amount of information that should flow among those tasks and the more formal should be the media.*

A strength of contingency theory is, like in Thompson's design theory, the recognition of complex interdependencies in organizational work. However, when dealing with information-processing it assumes that the environment is predictable enough to characterize existing interdependencies and that predefined mechanisms can be designed for various contingencies (Organizational Design).

Decisions, Knowledge and Organizational Learning

The contemporary quest for flexibility and adaptability of organizations to fast changing scenarios is vindicating the long ignored vision of loosely coupled and knowledge based organization (Cohen, March and Olsen, 1972; Weik, 1976, Nelson and Winter, 1982; Kogut and Zander). The loosely coupling refers to decisions that are just barely leaded by formal structures, but rather relate to the 'socialization' of individual (subjective) cognition (Weik, 1976). Collective decision-making in organizations, means that individual community members may develop diverse and competing ideas about the best way to organize (Weick, 1995). Cohen, March and Olsen (1972) proposed a '*garbage can*' model for organizational decisions (maybe still the only one) where problems encounter solutions just on the base of their co-occurrence in time and space, then regardless for formal structures.

I such a boundary-less representation of organization a different metric (from formal governance arrangements) should be proposed in order to explain performance differentials among economic agents (firms, individuals and institutions). In particular 'Knowledge', alternatively labeled as routines, competences or capabilities, has been proposed to shape organizations' performance possibilities (Nelson and Winter, 1982; Kogut and Zander, 1992). Because knowledge is a dynamic entity, it always implies some kind of learning by means of which organizations and firms have shifting knowledge boundaries over the time. A particular kind of organizational learning is reflected in modular organization architectures that have been claimed to provide an 'evolutionary' strategy for decoupling interdependencies (Simon, 1962-1996) among tasks.

Communication can provide the necessary means of coordination (von Hippel, 1990) when task interdependencies: (i) are no further reducible by re-arrangement (Warglien and

Levinthal, 1999); (ii) are not manageable in a less complex (more formal) way (Thompson, 1967). Communication and other information exchange patterns could then be implied to trace knowledge boundaries of organizations. Organizational learning in turn is thinkable as reflected in the evolution, generation, selection and retention of nodes and links, of communication networks over the time.

Both TCE and Contingency Theories gave to organization a Strategic interpretation. That is they respectively asked what 'make-or-buy' decision or what contingency impact reduction could improve the economic performance of an organization (firm) according to a rather positivist costs-benefit trade-off calculation. Cognitive and Knowledge based theories usefully re-introduced a kind of 'subjective dimension' of organizing, but essentially offered a view of boundary-less organization as systems of decisions and actions.

Isomorphism and Diversity, the Informal Side of Coordination

A complementary approach (to all the aforementioned ones) focused on a different feature of organizational systems, that is the 'informal' side of coordination to which somebody also refers as 'Organizational Culture' (DiMaggio and Powell, 1983). The key difference between this approach to the study of organization and more 'strategic' ones lays on the definition of institutions which departs from the most accepted (rules and contracts) in economic theory (North, 1973; Williamson, 1991).

In fact, by means of repeated interaction, social actors are supposed to institutionalize a collectively constructed reality (Berger and Luckmann, 1966) made of informal norms and values. Over the time, this social process can result in what economic sociologists called an

'O*rganizational Field'* (Powell and DiMaggio, 1983) that is an ensemble (population) of organizations whose action tend to conform to the aforementioned shared view because of the individual search for legitimation. This pressure for legitimation is supposed to lead organizational forms to resemble each other that is also called “isomorphism” (Powell and DiMaggio, 1986). This kind of social structure enhances organizational coordination by constraining the possibility of action and the access to resources an by reducing the uncertainty in decisional processes. In this sense, Neo-Institutional Sociology (Powell and DiMaggio, 1991) goes beyond the separation between formal and informal side of coordination which is latent in the 'strategic view' (Willimson, 1991).

Another approach confronting with the study of organizational forms over organizational populations has been Population Ecology (Hannan and Freeman, 1977; 1987). Similarly to Neo-Institutional sociology this research program focused on the study of large sample (populations) of organizations, but asking the opposite question, that is: why there are so many kind of organizations? (Hannan and Freeman, 1977). Building on the seminal contribution of the sociologist Stinchcombe (1965) this stream of research, in extreme summary, inquired the relation between organizational forms and organizational survival determining over long time periods an evolutionary dynamic at the population level.

Both Neo-Institutional organizational sociology and Populatio Ecology have been influenced by early works on Social Structures and Social Networks arguing that economic action is embedded in 'social evaluation' systems (Granovetter, 1973; 1985) or that organizational relations implies the duality of persons and groups (Breiger, 1974). However when dealing with communication and other informal channels for coordination, the underlined social structure is often empirically operationalized as a stable (given) entity constraining action possibilities and shaping organizational economic performance (Uzzi, 1996, 1997; Podolny, 2001).

Practices in Socio-Technical Systems

A more dynamic view on informal coordination structures (cultural institutions) has been provided by Practice theorists (Giddens, 1984; Bourdieu, 1990). A key argument put forward by “practice theorists”, is that neither the material world (the world of action) nor the cultural world (the world of symbols) can exist or be coherently structured independently (Mohr and Duquenne, 1997). The duality of culture and practice imply that practices become institutionalized over time by means of use. The ongoing interaction between individuals and institutions could be viewed as a 'structuration' process (Giddens, 1984). Structuration concerns the production, reproduction and transformation of social institutions, which are enacted by the use of social rules. These rules shape the action taken by individuals in organizations; at the same time, by regularly drawing on the rules, individuals reaffirm or modify the social institutions in an ongoing, recursive interaction.

When work situations are characterized by novelty, unpredictability, and ever-changing combinations of actors, tasks and resources it could be very complicated, even not useful, to ex-ante specify systems of routines and formalized plans of action. In such context the concept of 'trajectories' (Strauss, 1993) as sequences of actions toward a goal, could better emphasize the interplay between contingencies and interactions among actors. Trajectories are also a useful concept because it deals with deviations of the course of action from the desired objective. In those scenarios decisional processes are more dealing with the situation rather than with formal organizational arrangements (Miche and White, 1998).

Faraj and Xiao (2006) suggested that in complex knowledge and fast changing environments the 'lens of practice' are more suitable to understand coordination than traditional contingent approaches. Practices as suggested by Bourdieu (1990) have at their principle not a set of conscious, constant rules but practical schemes, opaque to their possessors varying according to the logic of situation. We argue that distributed-governance in virtual communities is a changing entity over time and place. We believe that rather than 'contingent', it is 'coherent' to a project domain and situation (Mische and White, 1998).

Contemporary Developments

I will consider now three contemporary streams of research which propose synthetic elaborations of the organizational economic debate over the last fifty years. These research streams are: (1) Modularity Theory; (2) Learning Networks; (3) Organizational Genres and Categorization.

Modularity Theory

In a seminal contribution Herbert Simon (1962) proposed the Idea of modular architectures as an emergent (successful) evolutionary strategy in complex systems. The basic idea behind modular production systems is that processes could be decomposed in near-independent components. Modular architectures of problem solving systems (like production) allow for more flexibility of coordination in response to uncertain environments (Dosi and Marengo, 2005). however this flexibility comes at the price of a just 'local' optimization (Ulrich, 1995).

Modularity theory has recently been proposed (Baldwin, 2008) as a way to realize a synthesis among Transaction Cost Economics (Williamson, 1991), because it punctually points to the locus of transactions (modules); and Knowledge Based View (Nelson and Winter, 1982; Kogut and Zander, 1992), because it accounts for the knowledge embedded in product design over the time. A central assumption for the more 'strategic' literature on modularity is that interdependences among product components mirror coordination needs among organizational units that realize it (Shilling and Mahoney, 1996). Recent empirical literature on complex product development (Eppinger et. al 2006) interpreted this concept in a rather contingent way, where the new contingency is product design (when components are realized by not controllable entities): organizations that would have a modular organizational structure should design modular products. Focusing more on organization's knowledge bases (as patents) it has also be shown that the use of near-decomposable firm's knowledge structure leads to increased usefulness of inventions (as patent citations) and also to the knowledge malleability or capacity for change (Yayavaram and Ahuja, 2008).

Learning in Inter-organizational Networks

Neo-institutional sociology of organization (Powell, 1991) also showed a considerable interest into knowledge transmission over the time (Powell et al. 2005). a key assumption is that "*when the knowledge base of an industry is both complex and expanding and the sources of expertise are widely dispersed, the locus of innovation will be found in networks of learning, rather than in individual firms*" (Smith-Doerr and Powell, 1996). This approach departs from the 'strategic' one (Pisano, 1989, Williamson, 1991) in which the form of collaboration is purported to vary according to the specific types of skills and resources to be exchanged. According to Brown and Daguid, learning is rather a social construction process.

Similarly to Cohen and Levinthal's (1990) 'absorptive capacities', a Network serves as a locus of innovation because it provides timely access to resources that are otherwise unavailable, while also testing internal expertise and learning capabilities. Then, inter-organizational collaborations are not simply a means to compensate for the lack of internal skills; nor should they be viewed as a series of discrete transactions (Smith-Doerr and Powell, 1996).

Categorization, Forms and Genres

Both Population Ecology Research (Hannan et al., 2005; Hannan et al., 2007) and Practice based research (Yates and Orlikowski, 2002; Im et al., 2005) have recently been devoting a growing attention to the problem of categorization in organizational affairs. In markets, as in all social domains, actors rely on systems of categories to interpret experiences. Category systems appear as social facts—they set rules about market boundaries and tell what appropriately lies within those boundaries. These shared understandings stabilize a market by channeling perceptions and actions in predictable ways.

Audiences pressure agents to conform to categorical expectations with implicit, or even explicit, threats of social and economic sanctions (Meyer and Rowan 1977; DiMaggio and Powell 1983; Podolny 1993; Scott 2001). Audience members rely on category boundaries to identify and make sense of producers (those agents who put goods and services on offer in the market), and producers that span diverse categories are likely to be ignored (Zuckerman 1999; 2000) or explicitly devalued (Pòlos, Hannan and Carroll 2002; Hsu 2006).

In support of this notion, Hsu (2006) finds that audiences express greater dissensus about the category memberships of films that target multiple categories (genres) as compared to those that target a single genre. Another empirical test for genres in films has been provided by examining how the diversity of genres that audiences associate with a film affects its appeal to critics and filmgoers as well as its success at the box office (Hannan, Hsu and Kocak, 2007).

When considering communication, as a fundamental means of coordination for distributed teams of developers, a central issue rises about its role as a set of practices linking both the material domain of action and the symbolic domain of culture. Drawing on Giddens' (1984) structuralist perspective, Orlikowski and colleagues (Yates and Orlikowski, 1992; Orlikowski and Yates, 2002; Im et al. 2005), proposed an approach based on C*ommunication Genres* (emails, meetings, expense forms, reports, etc.) as a social structure constituted through individuals' ongoing communicative practices. As they suggested in a recent work (Im et al., 2005): *"These genres are socially recognized types of communicative actions that are habitually enacted by organizational members over time to realize particular social purposes in recurrent situations (Yates and Orlikowski, 1992). Through such enactment, genres become institutionalized templates that shape members' communicative actions. Such ongoing genre use, in turn, reinforces those genres as distinctive and useful organizing structures for the community ..."*.

Whether used explicitly or implicitly, as organizing structures, *genres* shape beliefs and actions and, in doing so, they enable and constrain how organizational members engage in communication (Im et al., 2005).

OSS Development, Coordination and Communication

Open Source Software projects are development efforts based on the contribution of users-developers communities of productive agents who are geographically distributed around the world. Relaying mainly on ICT mediated communication as a means of coordination for development tasks, these communities create and maintain high quality and innovative technologies engaging in collaboration mainly on a voluntary base. Even if the theme of “OSS development” just recently appeared in economic and managerial literature it became a central topic for these disciplines over the last five years. In particular, due to the public regime for the availability of information about the development, the empirical research about OSS projects grew very fast.

So often empirical evidences coming from the same settings, while differing in level of analysis or data organization, have been used to validate contrasting economical and managerial theories and authors' personal beliefs on organizational human behavior. This resulted in potentially misleading interpretations and a lack of clarity on the important issue of new ITC-based productive models (Rossi, 2004). A very central problem for the OSS literature is the fuzzy overlapping of two polar position on the ontology of coordination, that is self-organization (von Hippel, 2007) vs. design (O'Mahoney and Ferraro, 2007). According to what I have just mentioned about the fast growth of OSS literature, it is not surprising that this 'fuzzy overlapping' perfectly mirrors the eternal economic debate on the spectrum of feasible solution between Market and Hierarchy (Simon, 1978; Williamson, 1991).

In the following paragraphs of this section I will rise some issues about the boundary between 'self-organization' (market-like coordination) and 'design' (hierarchy-like coordination) coming from some papers on OSS development projects coming from

Economics, Management and Software Engineering. This review has two main aims: (i) looking backward to economic organizational theory, that is to highlight the potential contribution of OSS development literature to the general debate on representing coordination of productive activities by means of communication; and (ii) looking forward, that is to formulate precise research question to be simply further explored or sometime, more precisely, empirically tested in the next two sections.

Motivations to contribute an OSS Project Shaping Communication Structures

Because developers are not directly reworded for the activities they perform in OSS projects, a significant attention has been payed in economic an managerial literature to the 'non economic' motivations leading individuals contributing OSS projects. Bagozzi and Dholakia (2006) surveyed hundreds of developers from Linux user groups finding that the main reason leading contributions was the sense of belonging to the group itself.

Lakhani and Wolf (2003) in a web-based survey administered to 684 software developers in 287 F/OSS projects find user personal needs, both work and non-work related, to be the overwhelming reason for contribution and participation. Gosh et al. (2002) in the so-called FLOSS survey also find a significant percentage of respondents indicating motivations that can be associated to the category of user needs.

All these findings are not in contrast with the more 'economic' explanation for programmers contributing in Open Projects. Economic motivations have been proposed to come from signaling incentives for individual programmers on the labor market (Lerner and Tirole, 2002) when being 'visible' in successful projects. Based on this kind of delayed reward a production function which is very much in the spirit of neoclassical ones, but without the capital, has been proposed (Lee, Moisa and Weiss, 2005).

Considering both the direct utility of the produced software for programmers who create it and the signaling incentives it has been proposed that organizational structures in OSS projects assume a flat self-organizing structure (Weiss et al. 2006; von hippel 2007). That is in network analysis' words, the coordination-communication network among agents assumes a topology with a high degree of randomness. More precisely in case of self organization communication networks assumes a scale free topology over the time. From this line of reasoning a proposition follows to be empirically explored in the next section (II):

Proposition 5: *self-organization is a key principle leading coordination in OSS development communities that results in scale-free topologies for communication networks over the time.*

Communication as an Information Processing Network

For virtual communities, where neither formal authority nor central planners are responsible for the so called 'organizational design', organizational structures could be seen as emanating from product architecture (Sanchez and Mahoney, 1996). Following this reasoning, since software has a more modular architecture than other products, virtual communities producing software will tend to have more distributed structures of governance than other productive organizations (von Hippel and von Krogh, 2003; Baldwin and Clark, 2006). Decisional processes in OSS projects seem to scale well for rising complexity[5] of the activity to be

[5] Our use of the word complexity is very much in the spirit of H. Simon's thought (Simon, 1963; Simon, 1996). according to him complex systems are those made of a lot of components among which interdependencies are not negligible.

managed because decisions are taken at the module-level rather than at level of the whole product. The use of Application Programming Interfaces (API) makes parallel development and components re-integration easier.

However, API(s) are static pictures (trees) of functional architectures while development is a dynamic process of learning by peer reviewing (Lee and Cole, 2003). Hence, communication is essential to interactively coordinate tasks to be performed by developers. Communication provides the selective principle according to which product and task structures co-evolve.

According to this view, the content of communication could be thought in this case as the mutual assistance that programmers provide each other. From a slightly different point of view Kuk (2006) proposed that programmers use communication as means of epistemic search for the knowledge that they need in order to solve their technical problems. In doing so they try to interact with those other programmers who control more 'valuable' knowledge, but also accept a general rule of reciprocity.

Communication as a Socio-technical Evaluation Network

In 1999 Eric Raymond presented his metaphor of the 'Cathedral' and the 'Bazaar' (as a way) to describe the dramatic difference between OSS development and commercial software development. The Bazaar's metaphor concerns a distributed-production system, involving a large number of developers and characterized by: (a) the absence of a centralized decision-making unit defining ex-ante the direction of development of the software code; (b) parallel design and debugging; (c) the integration of users into the production of software code; (d) self-selection of programmers for the tasks that best match their abilities.

The idea of flat self-organizing networks (Weiss et al., 2006; von Hippel, 2007) have been contrasted by a rather consolidated literature arguing that very different commitment levels characterize individual contributions and based on that, different roles are played by contributors in open projects (Mokus et al., 2002, Lee and Cole, 2003; Howison and Crowston, 2005; Kuk, 2006).

In fact, Virtual communities of production have less fluid boundaries than other virtual social networks like for example Facebook and MySpace and repeated interaction among programmers highlight a positional specialization where different roles[6] have different control on development activities. Mockus, Fielding and Herbsleb's (2002) study on both the 'Apache web server' and the 'Mozilla web browser', provided evidence for the existence of teams of 10 and 15 people who controlled the development of the majority of the source code. Following these findings Lee and Cole's (2003) work departed from the Idea of totally flat structures of production in OSS project, proposing that core-periphery structures are very likely to take place. They empirically tested this argument with an empirical case study on the Linux Kernel.

This kind of structural dynamics have been summarized in the so called “onion model” (Howison and Crowston, 2005), where nested levels of governance roles for programmers expand from the core of development. If we know that small communities of core members either produce more or access key development tasks (Mokus et al., 2002; Lee and Cole, 2003) very few is known about communication patterns linking them and it is not clear what

[6] Both the words 'role' and 'position' here are used in the spirit of structural sociology (White et al., 1976). A 'role' is the pattern of relations held by an actor, while the 'position' is the connectivity pattern of a role in a system of roles.

analytical techniques are more useful to inspect such 'development cores'. From the above considerations a further proposition to be developed, in terms of research strategy in the next section, follows:

Proposition 6: *cores of developers are responsible for the most part of the code development, however more attention to analytical techniques for extracting such components from communication networks are needed.*

Chen and O'Mahoney (2007) proposed, supporting their claim with an ethnographic study of four virtual communities, that Distributed Forms of governance come from the need of members for balancing two competing logics: (i) a so called 'expression logic' which pushes for the absence of formal organizations as a way to respect differences in members motivations, abilities, timeliness, and accountability and encouraged broad participation; and (ii) a so called 'production logic' which endorses rationalized, bureaucratic practices, such as a division of labor and rules.

Communication among community members operates a synthesis among these two logics over time. In a Study on the Linux Debian community, O'Mahoney and Ferraro (2007) used electoral debate data to show how leaders' vision of governance shifted over time from a more 'technical' (writing code) concept of merit to a more 'organizational' one (community building). Then, inverting the logic they also proposed that individual performance, as the likelihood for candidates of being appointed on community-management formal positions, was affected by the congruence over time between the individual behavior (writing code or community building) and the socially rewarded kind of merit. Communication as repeated interaction among members provides a means of social evaluation of others and reproduces trust resulting in status positions (O'Mahoney and Ferraro, 2007) and social structures.

Some authors (Grewal et al., 2006) showed that embeddedness (Granovetter, 1985, Uzzi, 1996), measured as centrality of both projects and individuals, in communication networks, could increase (project-programmer) legitimation and then could positively affect the access to resources and performance. Both 'trustt and 'familiarity' are similar social (to the aforementioned ones) dynamic which has been showed to affect individual and project performance in such virtual environments where formal arrangements are very rarely used (Ferraro and O'mahoney, 2004; Stewart, 2005). Because I believe that all these dynamics implying a social evaluation process could affect the structure of communication networks, but they received a minor attention from literature, a work proposition follows that will provide the underling reasoning for the section three:

Proposition 7: The impact of social evaluation dynamics on the emergence and change of communication networks over the time, in OSS projects, should be further tested.

Avoiding the Separation Among Technological and Institutional Domains

I did not find any managerial or software engineering literature addressing the issue of communication practices in OSS projects, hence I will elaborate here my own theory about this. Due to the recency of the argument of communication practices in OSS environments I had to borrow some concepts for my theoretical construction from previous consolidated research on coordination theory (Thompson, 1967) and structuration theory (Giddens, 1984). I also borrowed from practice-based research in the commercial software development field (Im et al., 2005).

Instead of considering communication either as the structure of epistemic information search/processing (Kuk, 2006) or alternatively as the structure of social evaluation (Grewal et al. 2006), I propose here that communication is an evolving set of coordination practices (Im et al., 2005). Indeed Thompson's (1967) seminal contribution already avoided the dichotomy between an objective domain of action and a subjective domain of collective sense-making considering the organizational environment as a changing ensemble of tasks to be coordinated over the time. Both technological and institutional uncertainty contribute to explain environmental complexity for tasks to be coordinated.

Governance structures in virtual communities of production, like for example OSS projects, emerge and change as bounded rational agents (Simon, 1957) attempting to control two dimensions of uncertainty shaping environmental complexity: (i) technological uncertainty; and (ii) institutional uncertainty (Thompson, 1967).

Distributed governance in fast growing virtual community is, in my opinion, a general concept that underlines a vision of decentralized/informal decisional processes. In fact on the one hand, decision are commonly (not always) taken at module-level, but to some extent hierarchies among tasks (as an emergent community development agenda) always exist. On the other hand, projects rely on cultural coordination (OSS is also a social movement) until the institutional uncertainty is under control, then they recur to more formal governance systems.

My argument is that each project in a different measure borrows, from both distributed software development experience and the OSS social movement, reproducing work practices. Over the time, by means of use of those practices, each project specifies its own governance system for tasks to be completed (decision making). The decision on how to admit new members to the community of a growing project, can be used as an example. Projects generally borrow the general concept of 'merit': “who writes the code is allowed to take the decisions” from the OSS social movement. Over the time a project, which has been using that admittance rules, discovers its own practices (way) to make its decision. When talking about communication practices there is some empirical evidence that project members institutionalize practices in so called communication genres (Im et al., 2005). Communication genres are not fixed institutions but rather accepted practices that could be used in a flexible way and that could change according to its social use over the time. Following Im et al. (2005), they could be operationalized as email tags like the [bug], [proposal], [vote] etc in email subject lines.

Considering the above, communication is the means of coordination for flexible trajectories of action. Hence, a first general research question to be empirically explored in section four is: *how communication genres are flexible to the situation but at the same time provide a coherent domain for action in distributed systems of governance leading discussions to its objectives?*

A second research question about communication practices, to be empirically explored in section four is: *how much communication practices, institutionalized in genres, could simplify the complexity of decisional processes coming from diverse communities of decision-makers?*

SECTION II

DETECTING AND TESTING STRUCTURES FROM EMPIRICAL COMMUNICATION DATA

Introduction

In section one, I proposed a theoretical representation of productive dynamics which is based on networks change over the time. In particular I argued that especially (but not only) in distributed development efforts, communication flows could mirror an underling technological structure of interdependent tasks to be coordinated. In this section I will illustrate alternative analytical strategies for representing such coordination networks in innovative virtual communities whose boundaries are defined by socio-technical interdependences among participants. I will extract and organize communication data from thousands of emails retrieved by the Apache Open Source email archive. then I will build communication networks to give empirical contents to my theoretical arguments and to substantiate my claims that: (i) Self-organizing networks provide the basic principles of coordination in such communities; (ii) Once in place, deliberate governance arrangements affect coordination patterns within virtual communities; (iii) Structural properties of communication networks change significantly over time depending on their internal organizational logics, and (iv) Affiliation (a.k.a. two mode) networks provide a useful representation for detecting community structures.

Data I collected cover a ten year time period – from 1995 to 2004 – of distributed software development. Beginning in 1995, the Apache community created (and still now maintains) the most widely implemented web server software in the world. The second part of this section consists of an analytical development of my perspective on the endogenous organizational dynamics of communication and coordination. Each of my general arguments is introduced by an abridged survey on the state of the art in OSS literature. Then each issue is developed by means of network analytical tools and results are discussed. The section three concludes with research questions that could further extend and strengthen the preliminary results presented.

Innovative Networked Communities

Recent years have seen the emergence of new conceptual models of innovation which rely on ICT mediated communication to coordinate production and exchange activities. Such models tend to assign a rather limited role to formal governance mechanisms that are viewed as restricted in scope (to regulated tasks) and time (adoption in advanced stages of growth). These two simple assumptions are of great relevance for the study of organizational dynamics of innovation.

When the knowledge needed to generate innovation is both complex and distributed across different organizations or units (Powell et al., 1996), network partners and institutions affecting patterns of exchange become of central importance for our understanding of innovation processes. For example, Ancona et al., (2002) argued that successful teams (X-teams) within organizations, today are characterized by porous boundaries and fluid membership allowing organizations to reach the knowledge they need to sustain high innovation rates over the time.

In order to achieve some collective objective organization members are supposed to rely on some form of shared knowledge. When such knowledge is distributed around a community of interacting actors, interdependent tasks typically require some information exchange in order to be coordinated (von Hippel, 1990). In the project management literature, for example, information exchange has been argued to map a kind of 'state of the world awareness' to sequences of 'possible actions' (Pich et al. 2002). Then, both the amount and structure of known information determine the complexity[7] of decisional processes for project teams in order to perform tasks. Literature on problem solving (Nicherson and Zenger, 2003; Levinthal, 2006) proposed that modular (decomposable in near independent parts) and barely formal organizational architectures should display an evolutionary advantage, when compared with more traditional ones, in complex and fast changing environments.

Despite this recent recognition of the advantages for both distributed and networked models of innovation, comparatively little attention has been paid to communication patterns. We think of communication as an important means of coordination, an enacted structure that links interdependent tasks[8] to be performed (Kleinbaum et al., 2009, Monge et al., 2008) in order to 'feed' innovation processes. I also think that Open Source Software (hereafter OSS) projects could be a perfect empirical setting to both develop and test a reflection on distributed organizational dynamics. In this section we will focus on email

7 A complex system of whatever nature is intended here as one which is made of a large number of simple but interdependent component parts (Simon, 1962).

8 The paper by Cataldo et al. (2006) is a notable exception.

communication intended as the main means of coordination for distributed development in a successful OSS project.

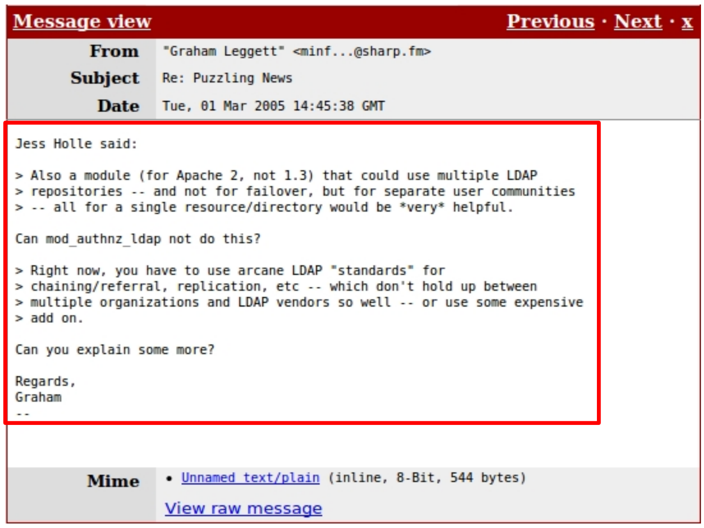
I will discuss general issues about the dynamics of organizational structure, and the adequacy of available analytical strategies for detecting it and represent its change over the time. More precisely we want to explore four issues: (i) the evolution of information exchange structures defined in terms of communication networks; (ii) the impact of adopting formal governance arrangements on communication structures; (iii) the usefulness of direct communication networks as a basis for networks decomposition, and (iv) the detection of community structures in communication networks, based on the dual association between programmers and mailing list (Breiger, 1974, Simon, 2002).

Gathering Data

In order to gather email traffic data I wrote an ad-hoc script (in Phython[9]). These kind of (scripts) programs are also known as spiders. The typical email in the Apache archive has a structure like in figure n below. From these standard email formatted data I gathered the text bodies in order to perform semantic analysis (see section four). Then like showed in figure nn. from the email header (row text) of each email I retrieved six kinds of information. These are respectively: (1) *message id*, that is a unique id for each sent email (2); *In-Reply-to*, that is a message id indicating whether and to which message the current mail is a reply-to; (3) Date, that is the exact time point in which a message has been sent to a mailing list (4); *Subject*, that is the email subject I used in section four in order to address the issue of communication *genres;* (5) *From,* that is the author of an email message (6) *To*, that is the mailing list to

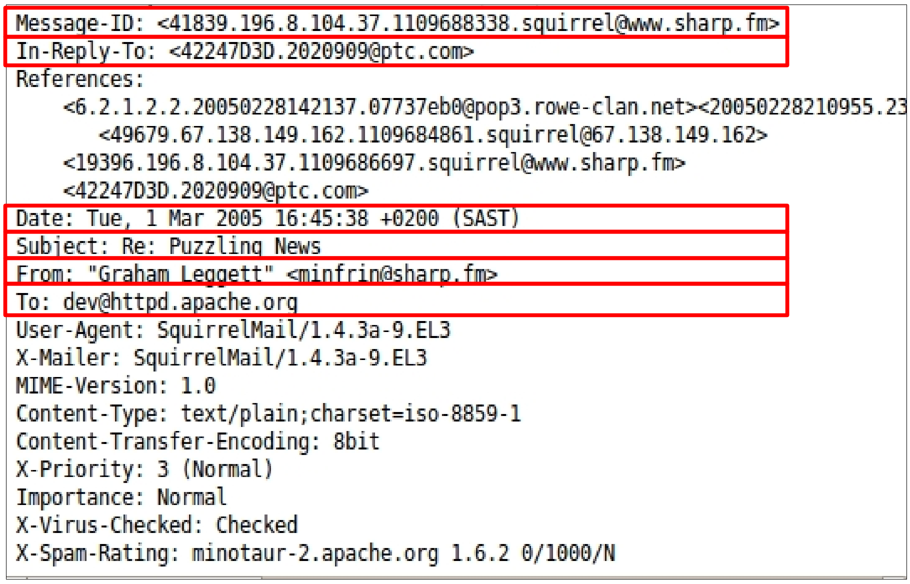
[9] Python is an Open Source Programming Language. Home http://www.python.org/

which a particular message has been sent.



```
Message view                                   Previous · Next · x
From     "Graham Leggett" <minf...@sharp.fm>
Subject  Re: Puzzling News
Date     Tue, 01 Mar 2005 14:45:38 GMT

Jess Holle said:

> Also a module (for Apache 2, not 1.3) that could use multiple LDAP
> repositories -- and not for failover, but for separate user communities
> -- all for a single resource/directory would be *very* helpful.

Can mod_authnz_ldap not do this?

> Right now, you have to use arcane LDAP "standards" for
> chaining/referral, replication, etc -- which don't hold up between
> multiple organizations and LDAP vendors so well -- or use some expensive
> add on.

Can you explain some more?

Regards,
Graham
--

Mime     • Unnamed text/plain (inline, 8-Bit, 544 bytes)
         View raw message
```

Figure 3. the email classic format for messages in the Apache mail Archive



```
Message-ID: <41839.196.8.104.37.1109688338.squirrel@www.sharp.fm>
In-Reply-To: <42247D3D.2020909@ptc.com>
References:
    <6.2.1.2.2.20050228142137.07737eb0@pop3.rowe-clan.net><20050228210955.23
       <49679.67.138.149.162.1109684861.squirrel@67.138.149.162>
    <19396.196.8.104.37.1109686697.squirrel@www.sharp.fm>
    <42247D3D.2020909@ptc.com>
Date: Tue, 1 Mar 2005 16:45:38 +0200 (SAST)
Subject: Re: Puzzling News
From: "Graham Leggett" <minfrin@sharp.fm>
To: dev@httpd.apache.org
User-Agent: SquirrelMail/1.4.3a-9.EL3
X-Mailer: SquirrelMail/1.4.3a-9.EL3
MIME-Version: 1.0
Content-Type: text/plain;charset=iso-8859-1
Content-Transfer-Encoding: 8bit
X-Priority: 3 (Normal)
Importance: Normal
X-Virus-Checked: Checked
X-Spam-Rating: minotaur-2.apache.org 1.6.2 0/1000/N
```

Figure 4. The row text header an Email from the Apache archive.

In order to complement the social side of the development process, I also gathered data about modification operated by programmers to software code. These data cover a more 'technical' component of the development. From the very beginning of the Apache project, community member used a CVS (Control Version System) and lather a SVN (Sub Version System) for managing issues and modifications to the software code. From these code repositories I essentially retrieved three information kinds: (1) *commits* modifications to the software code operated by individual developers (who are also called committers); (2) *reviews*, that are ensembles of code modifications (*commits*) affecting potentially interdependent files. (3) *time*, that means in what time point commits and reviews have been operated. These data is intended to partial integrating the main analytical level of this thesis which rather lays on communication networks.

Discovering and representing structures with data

For my exploratory analysis I arranged relational data that have been presented above (figures n and nn) in two kinds of communication networks for the Apache community over a ten year time period (1995 trough 2004). Those network kinds are: (i) direct communication networks; and (ii) affiliation networks. The first one is a so called “one-mode” (only one type of nodes) social network meaning that it represents a one-to-one email exchange amoong programmers. In the second one, individual programmers are connected through their dual association with mailing lists to which they contribute.

Direct communication networks are intended here as networks whose nodes are community members and whose links exist between two nodes when an agent (developer) sent a message in-reply-to another message by another agent (developer). Links were weighted using the number of exchanged messages among dyads of agents. Affiliation networks are built with two kind of nodes – programmers and mailing lists – and nodes of one type only connect with nodes of the other type.

For the examination of network topologies I generated random graphs (whit normal distribution of nodal degrees) in order to compare them with actually observed networks. For analysis aiming to partition the networks I just operated a network reduction cutting lines and nodes under a given threshold of connectivity. For final analysis I fist, folded two-mode networks by multiplying original matrices (algebraic representation of networks with nodes of type one on rows and nodes of type two on columns) for transposed ones (which have the same kind of nodes on rows and columns). In this way we obtained one mode networks, weighted for the number of shared nodes (of the other type). On these networks I used Newman clustering algorithm for finding community structures (Newman and Givran, 2003). This algorithm hierarchically decompose networks in sub component progressively removing nodes with highest betweenness centrality (Freeman, 1977).

Evolutionary Dynamics of Communication: Scale-Free networks and Self-Organization

The diagrams reported in Figure 5 show the evolutionary trajectory of such 'direct communication' networks for the Apache project over ten years time period 1995-2004. The number of nodes increased from 28, in 1995 to 6353 in 2004. The number of edges (network ties) increased from 38 to 16100 during the same time period. The components count was 1 in 1995 and grew to 113 in 2004. In Figure 6 we can see how both network degree centralization (average centrality for the overall network) and network density decreased by one and two orders of magnitude respectively over the observation period.

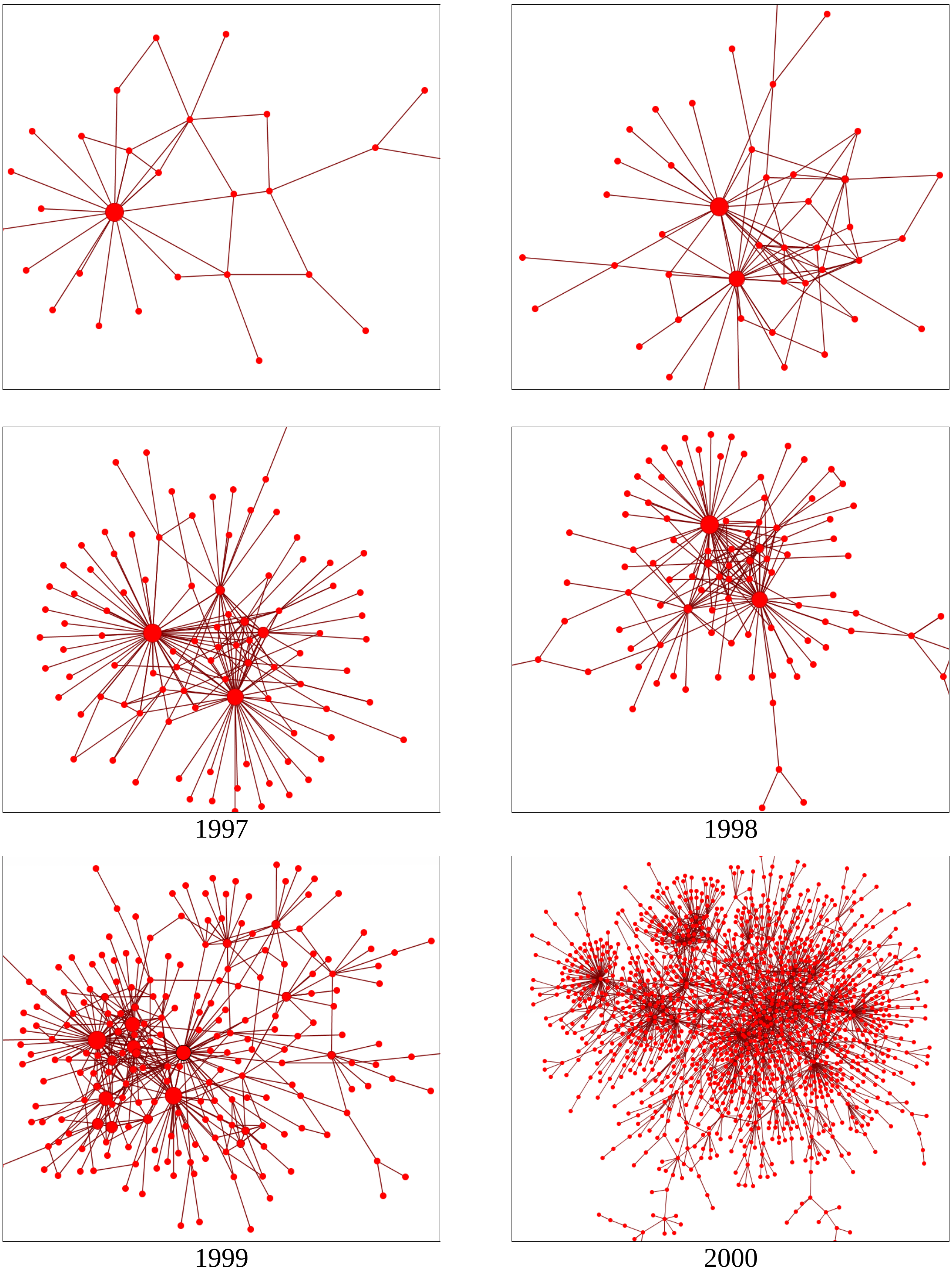
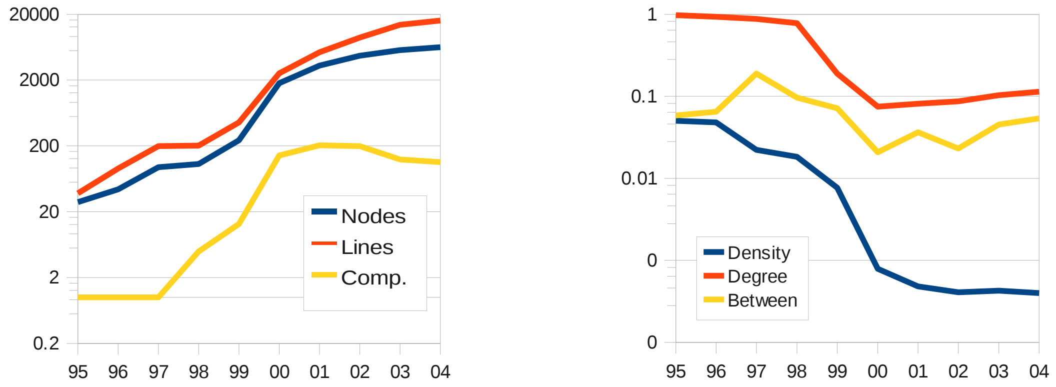


Figure 5. The evolution of direct communication networks over a ten years time period.



(a) (b)

Figure 6. (a) Node count, edge count and component count (Comp.). Values on Y axes are in logarithmic scale. (b) Density, Degree Centralization and Betweenness Centralization. Y axes is on logarithmic scale.

Weiss and colleagues (2006) studied the degree (number of lines incident to a vertex) distribution for the Apache email archive finding that only few developers held a high value (degree distribution follows a power law). They also controlled for the existence of the so called preferential a attachment phenomenon according to which over the time more connected nodes are more likely to become even more connected than others (rich gets richer). My analysis confirms this results (see figure 7 below) and we also controlled for both the clustering coefficient values and average distance values over time (see figure 8 below) (Watts and Srogatz, 1998). Both average distance and clustering coefficient values, were higher then the correspondent values in random networks with the same density and number of nodes.

This first examination is particularly important because it eventually will provide the rejection of what is a kind of 'null hypothesis' for the thesis as a whole, that is: the structure of communication networks is a random one. In fact a random network (Erdos and Renì, 1952) of communication would be interpreted as the total absence of coordination patterns emerging over the time, and then as the absence of an organizational structure itself.

This result could be interpreted as the overall network holding a scale-free topology (Barabasi and Albert, 1999). It has already been shown how scale-free networks could be generated from an initial network according to a variety of self-organization mechanisms. A well studied mechanism is preferential attachment. If over the time new nodes attach themselves to others according to the simple 'preference' for already highly central others a scale-free network will be obtained.

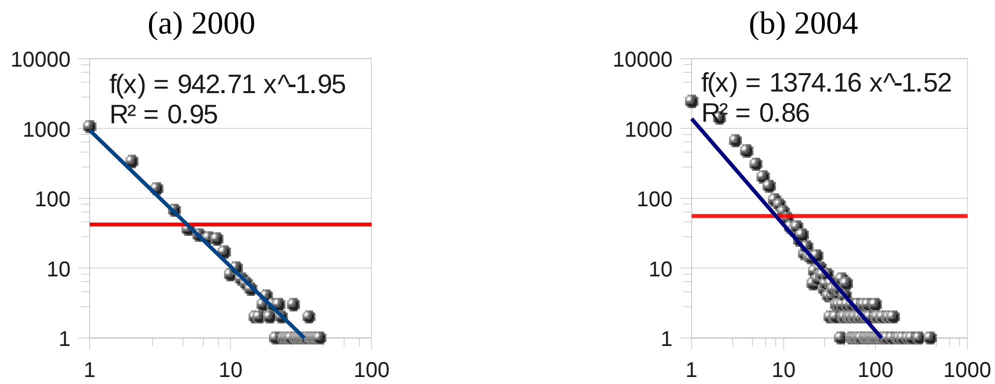


Figure 7. Degree distribution for 2000 (a) and 2004 (b). Degree values are reported on x axes, the number of nodes holding that degree are reported on Y axes. Both axes are in logarithmic scale.

This concept, that literally means the emergence of organizational structures in absence of central planners, seems to be of a certain interest for the study of virtual communities. In economics it has been argued that this behavior could be explained by the signaling incentives for individual programmers on the labor market [Lerner and Tirole, 2002; Paul and David, 2006] it.

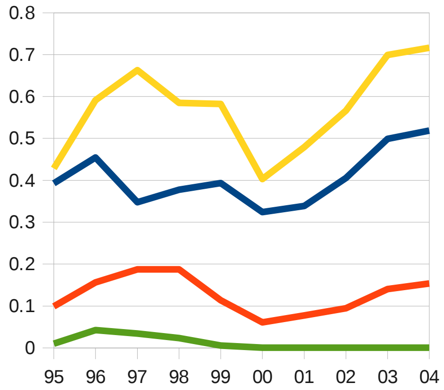


Figure 8. Original networks. Both the average distance and the clustering coefficient for the real networks are over the values for the random generated networks.

Degree Based Assortativity

As an further exploration of attachment dynamics among nodes, I also computed an assortativity coefficient 'r' (Newman, 2003) for the overall communication network over ten year time period (1995-2004). This measure consists of a Pearson correlation coefficient computed on the networks nodes decomposed in dyads.

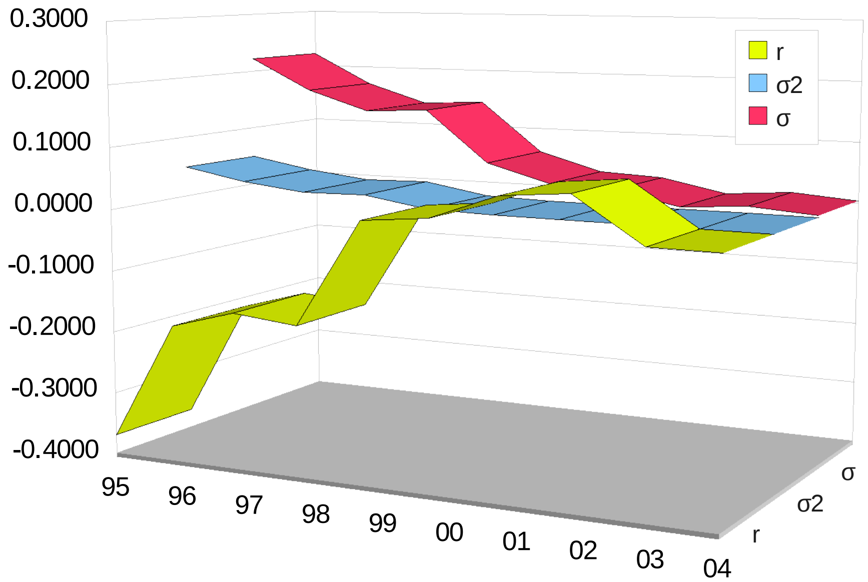


Figure 9. Degree based Assortativity over the time for in_reply_to networks.

It practically offers a measure of how much nodes with a given centrality either tend to attach to similarly central nodes (assortativity) or tend to attach to differntly central nodes (disassortativity).

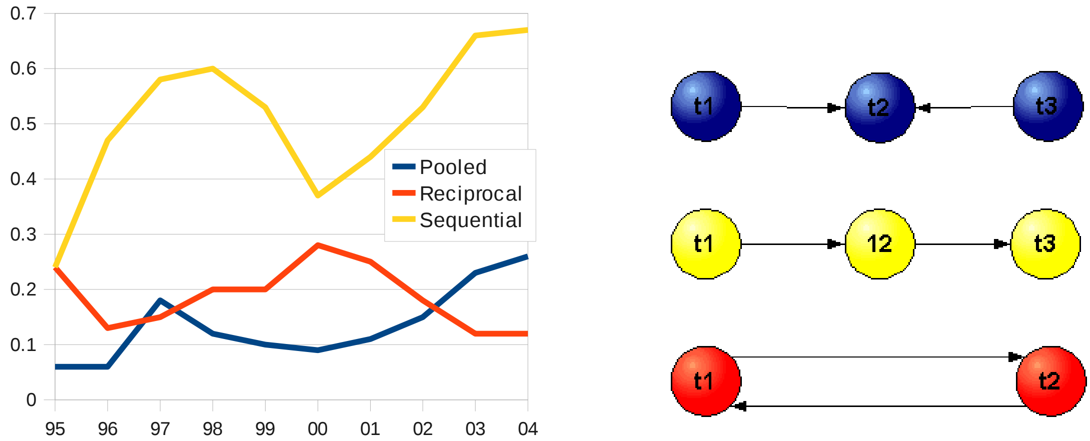
The Impact of Formal Institutions on Communication Networks

Self-organization in virtual communities is an important coordination mechanism. However, recent research has shown that successful fast growing projects need to be balanced self-organization with formal governance arrangements which are designed to lead the development process in desired directions [O'Mahoney and Ferraro, 2004, 2007]. Then I expect that not only institutional mechanisms are in place in virtual communities but also that such arrangements affect the technological dimension and the communication needs coming from technological interdependences. This is a further reformulation of 'null hypothesis' for further examinations, that is: not only self organization leads to non random organizational networks, but the intentional action of community members shapes the development process and structure over the time.

According to James Thompson’s influential statement, technological interdependencies may assume three basic forms. Arranged in an increasing degree of complexity the forms of interdependence are: (i) pooled; (ii) sequential, and (iii) reciprocal [Thompson, 1967]. According to Thompson, organizational structures should be designed in order to cope with the different degrees of complexity coming with task interdependencies to be coordinated.

Basing on our initial assumption that in virtual communities coordination should be mirrored by communication patterns, I expected that those patterns would change after the design and implementation of formal governance arrangements. More precisely, the formalization of organizational structures, which in my case study could be intended as the creation of the Apache Software Foundation (ASF) in 1999, would lead coordination toward simpler forms (patterns).

In order to explore this argument I simply counted, in direct communication networks over time, how many patterns were corresponding to Thompson's interdependencies as percentage of the total number. As showed in figure 10, my expectations are confirmed because: on the on hand, both 'pooled' and 'sequential' interdependencies tend (on average) to increase before 2000, while they tend to diminish after that time; on the other hand, reciprocal interdependencies – the more complex type – increased before 2000 and diminished after that time.



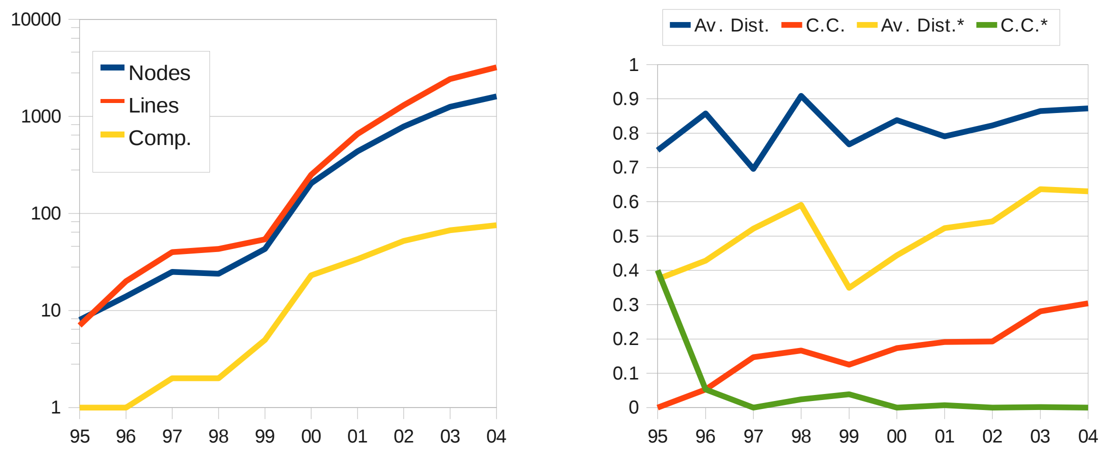
(a) (b)

Figure 10. (a). Thompson's interdependencies count (percent of total) in direct communication networks measured as percentage of the total number of links among nodes; **(b).** Interdependency shapes: (blue spheres) pooled; (yellow spheres) sequential; (red sphere) reciprocal.

The Adequacy of 'in-reply-to' built Networks to Network Decomposition

Another point that should moderate the extent of findings on self-organizing dynamics in OSS projects is about coordination and division of labor in large communities. Early literature on OSS development highlighted how small groups of developers actually accounted for writing the most of software code in Apache and Mozilla [mokus], and Gnome [Koch and Schnider, 2002]. This studies also found that the number of contributors who fixed bugs was one order of magnitude higher than the number of those who wrote the code [Mokus et al, 2002]. It seems that, when looking at productivity, large communities display core-periphery structures [Lee and Cole, 2003] and nested layers of roles [Howison and Crowston, 2005]. So we ask here: what happens to communication networks when we just consider the core of interaction processes?

In order to explore this issue, we assumed that the more a developer writes code the more hi will use email communication in order to coordinate his actions with other community members. Then, we applied a simple cut (lines and nodes) reduction on our direct-communication networks.



(a) (b)

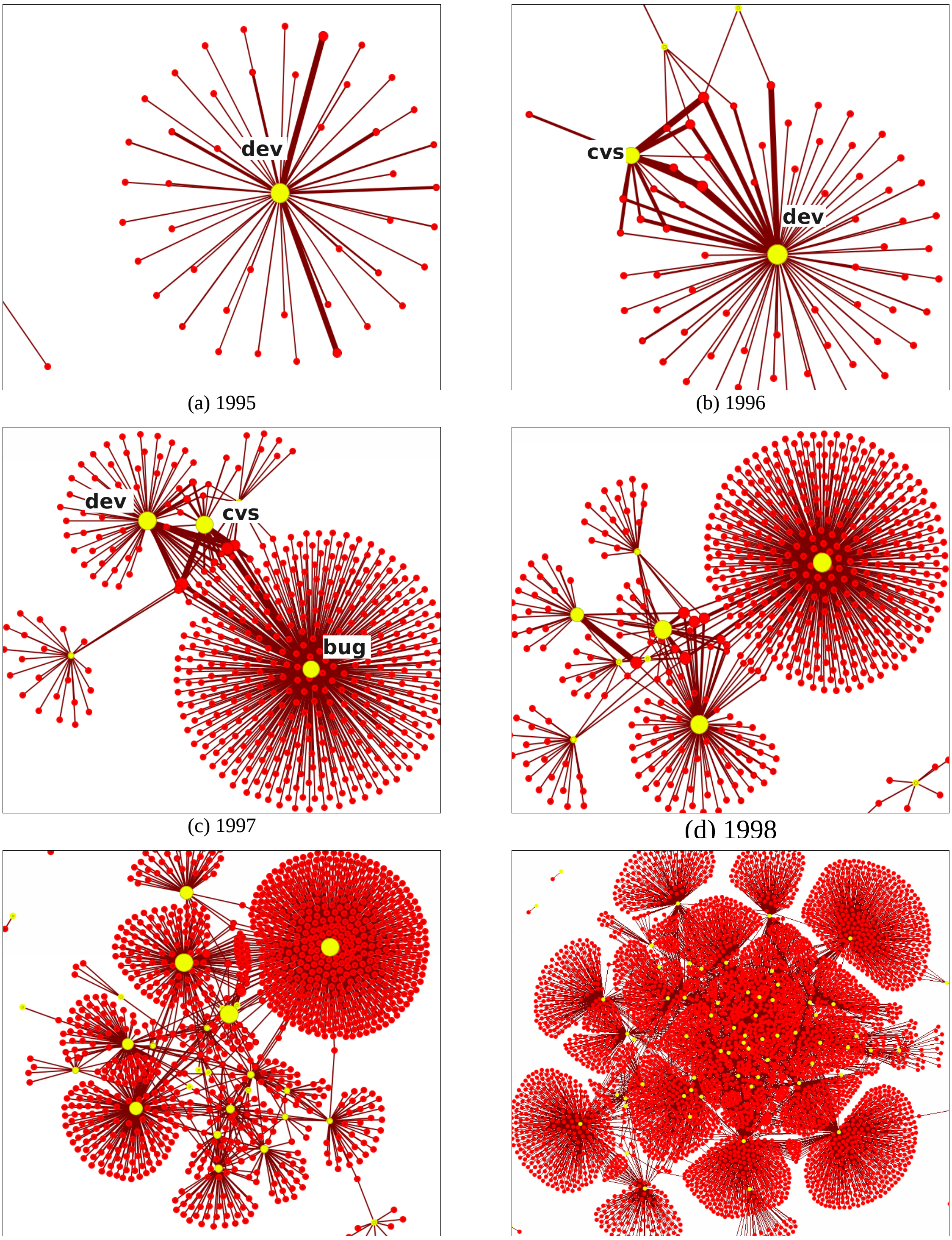
Figure 11. (a) Nodes, lines and components after network reduction with cut-threshold = 3; **(b)** Clustering coefficient and average distance after network reduction with cut-threshold = 3. Values marked with * refers to random generated networks.

This means that we removed from networks that lines with a value lower than a given threshold (say 3 exchanged emails) and then we removed which those nodes that resulted to have a total degree (in + out) less than 1 (say isolate nodes). The results of this procedure are showed in chart 1.b. The so reduced network 'captures' on average (over the time) the 21% of nodes, the 17% of lines and the 63% of components. It is also to notice that in reduced networks the density is on average the 27% higher than in the original networks. We interpret tis result as a higher connectivity among more active (core) members of the community. It is also to notice that networks, whose links were created using the in-reply-to filed on email headers, are very sensible to cut-like method of reduction.

When we look at values from the reduced networks at least two things are to notice: first, the clustering coefficient is monotonically growing (chart 2.b.) instead of floating (chart 2.a.); second, the values of average distance for real networks is higher than the correspondent values for random generated networks. Combining these two findings we could say that core members tend, over the time, to form clusters which are characterized by high inbound connectivity and low outbound connectivity.

Affiliation Networks, Modular Architectures and Newman clustering

The Evolution of Affiliation Networks. In this paragraph we graphically represent a story which has already been told a lot of times (Fielding, 1999; Mokus et al., 2002), the one of the Apache project. The process of development begun for Apache with a single mailing list, the dev(elopers) one (the second yellow node in figure 1.a. only appears at the end of the firs year). Since the second year of activity the Apache Group implemented a Current Version System for code management whose activity is mirrored in the 'cvs' list (figure 1.b). As we can see in figure 1.b the most active users writing on the 'dev' list are at the same time that users who have been allowed to access to the 'cvs'.



(a) 1995 (b) 1996 (c) 1997 (d) 1998 (e) 1999 (f) 2000

Figure 12. The evolution of programmers to projects affiliation. Yellow nodes are mailing lists, while red nodes are programmers. The size of nodes is weighted with the nodal degree centrality, while edges weight expresses the number of sent emails during per year.

The findings in paragraph 2.3 made us thinking about another strand of organizational literature on modular structures[10] on OSS projects whose major claim is that coordination patterns should mirror technical interdependencies (Baldwin and Clark, 2006; Rusnack et al., 2006). Because software has a more modular architecture than more traditional products have, the organization that produces it should have a modular structure as well.

On the one hand, it is very reasonable to assume, coherently with modularity theory [Simon, 1962], that a programmer working in a peripheral module probably just knows very little about what the development concerns in another 'distant' periphery of the community. On the other hand, community members who frequently work on the same modules should be supposed to reciprocally communicate. Then we think that coordination network structures could be decomposed in modules (mailing lists) according to affiliation patterns of agents (developers).

In order to explore this issue, as mentioned at the beginning of this analytical section, we built a two-mode network where nodes of type one are programmers and nodes of type two are mailing lists (see respectively red nodes and yellow nodes in figure 12). The weight of this affiliation is computed as the number of email that a programmer sent to a mailing list per year.

From the two-mode network we 'derived', a new one-mode network (folded) whose nodes are only mailing lists. The underlying assumption when we build this new network is that the higher the number of programmers who use the same mailing lists the more those mailing list refer to interdependent activities. By construction, two mailing lists were linked when at least a developer wrote an email on both. The weight of these relations have been imposed equal to

[10] A modular structure, or architecture, is intended as one in which components (building blocks) are barely interdependent among them. A practical consequence for product-project management is that near independent components can be developed in parallel.

the sum of developers shared by mailing lists dyads (and adjusted for the weight of affiliation). These new (folded) network loses the property of representing 'exact' communication patterns but it is less sensitive to cut-reduction. This means that we can consider only the developers who, wrote at lest a given number of emails (for example 10) over a year time period without dramatically altering the network structure.

In order to find sub-communities of coordination modules (represented for example as clusters of mailing lists), we used a folded one mode (mailing list to mailing list) networks (year 2000) reduced applying a cut-with threshold = 10. Then, we used the Newman clustering algorithm for detecting community structures. The modularity level was measured by a clustering coefficient Q ranging from 0 (non modular structure) to 1 (totally modular structure). We found a Q = 0.2013 in the mailinglist-mailinglist network (see figure 3.b where the same color is assigned to nodes that belong to the same cluster).

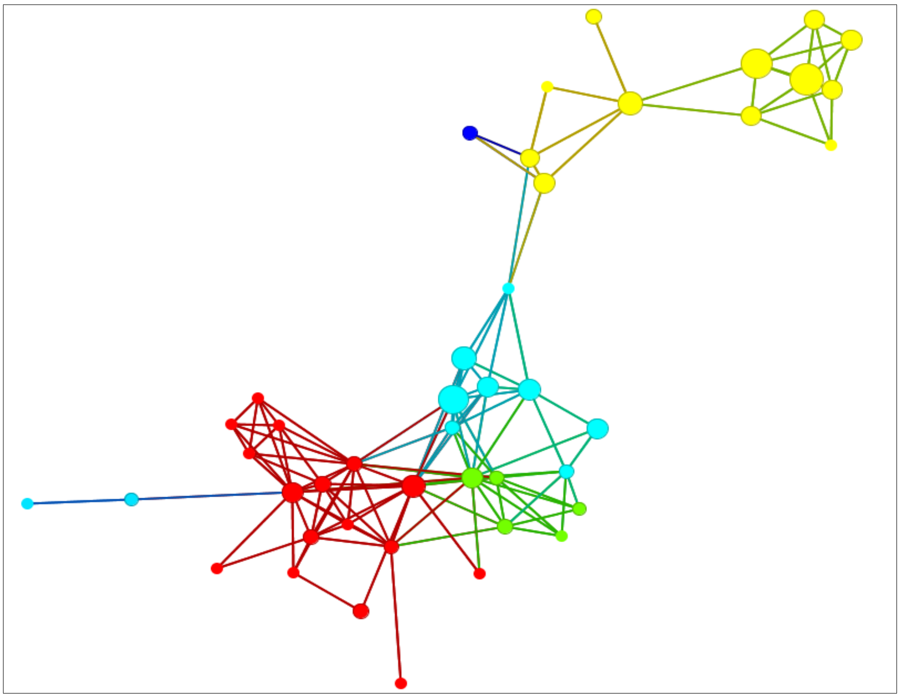


Figure 13. Module-module (lists) network, cut threshold = 10, Q = 0.2013.

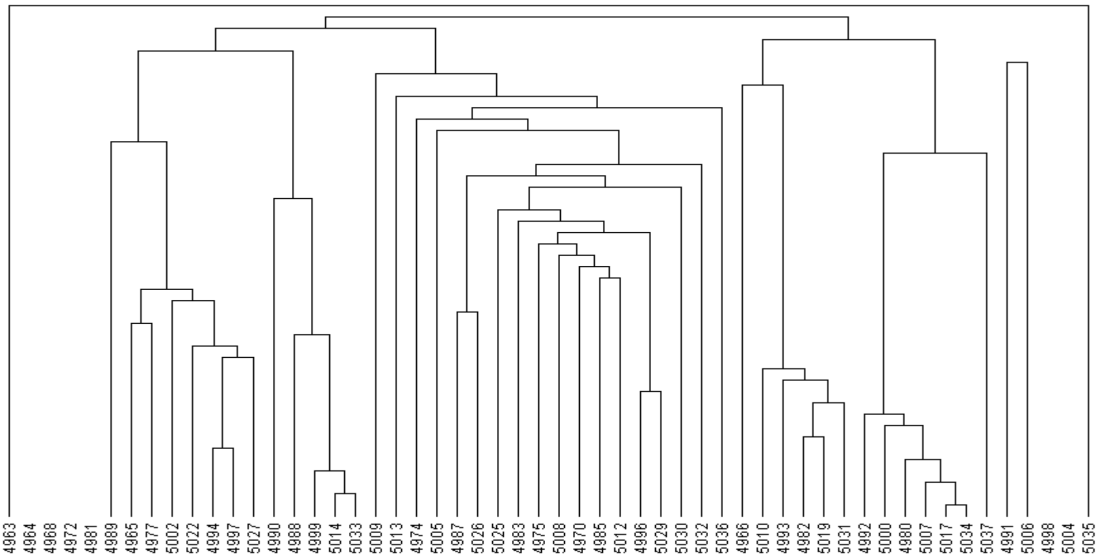


Figure 14. Hierarchical representation of Newman clustering algorithm at work, Dendogram of the clustered network in figure 13.

This research strategy was intended as a test for the 'resistance' of folded networks to cut-like reductions, we tried it for increasing cut-thresholds, ranging from 0 to 10, before running the Newman clustering algorithm and we found that the Q (modularity coefficient) only changed by a 0.1% for that range. The resulting modularity coefficient (Q) values could be interpreted as detecting a low modular organizational structure (of coordination). We also repeated the clustering process using a folded network with only developer-nodes and obtained (Q) values which were very close to 0.8, highlighting a very modular social structure, for both original and reduced (having cut threshold from 0 to 10) networks.

I further used this partitioning technique for detecting a sub-community with more manageable dimensions in order to perform more refined analysis. I have found that for a cut threshold = 30, that means removing from the affiliation network programmers who just sent less than 30 emails over one year, the Newman algorithm isolated the core community of the apache project. That core community is composed by the founders of the Apache project itself, the so called Apache group (see section three). This result is showed in the following figure (x) where the isolated green colored cluster on the left side perfectly capture the mailing lists belonging to the software server platform. The main component captures projects which complement the platform's functionality according to a modular architecture (Rusnack et al. 2006).

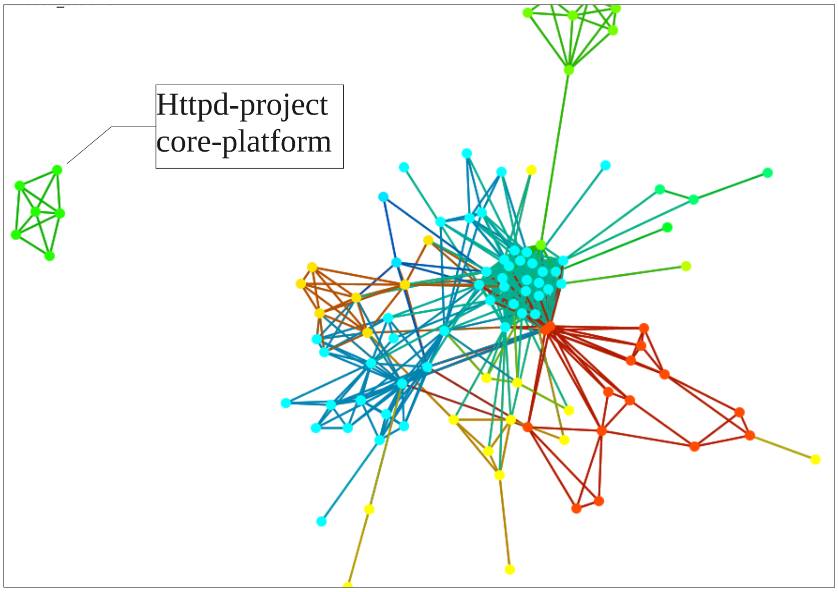


Figure 15. Newman Clustering. Modularity coefficient is Q = 16031356. The Platform or the core project results isolated see the green component in the left side of the picture.

Conclusions and further research

The results that we reported in this paper confirm that, when we build direct communication networks using the in-reply-to field of email headers for generating links, the overall network topology tends to develop scale-free qualities. This could be interpreted as the presence of self-organization in virtual communities, that is coordination structures could be thought to emerge in absence of central planners.

Despite this finding, we showed how the same networks could reveal that organizational design, which may be viewed as an almost opposite exogenous organizing principle, could have been affecting coordination-communication patterns over the time. We think that a further exploration of connectivity patterns could advantage the knowledge in the field of emergence of governance in virtual communities. In particular it could be interesting to control for the existence of eventual correlations among developers attributes (productivity, tenure etc.) and Thompson's typical interdependencies.

Driven by contrasting (or balancing) dynamics that we have documented, we further explored the issue of finding core interaction components in the overall networks. At a macro-level we observed a more clustered structure after reduction. However, direct-communication networks resulted very sensible to a low cut-reduction threshold, that is the shape of networks changed a lot when we just assumed that core community members exchanged at least three emails over a one year time period. This means that further micro-level analysis, like the one conducted in paragraph 2.2 could not be significant anymore.

A possible way to cope with this issue is presented in paragraph 2.4, where we proposed a different way to represent communication networks based on the idea of affiliation of developers to mailing list as 'modules' of the overall coordination structure. We have shown that 'folded' networks, either agent-agent or list-list from affiliation ones (with a cut threshold of ten) are respectively highly modular and low modular ones. A further contribution in this direction could be the construction of networks where developers affiliate to a more micro-level of coordination-communication, that is emails threads. This could offer a representation which is closer to direct communication without suffering from obvious problems of sensibility to cut reduction.

SECTION III

TESTING THE INTERPLAY AMONG INSTITUTIONS,COORDINATION AND CONTRIBUTION INEQUALITY

Introduction

Despite the progresses in representing networks, we still know few about causes affecting their emergence (Smith-Doerr and Powell, 2003). Moreover, Despite a consistent body of product-development research on technological interdependencies (Perrow, 1978) as shaping organizational structures (Rusnak et al, 2006; Cataldo et al. 2006, Sosa et al. 2005), very few attention has been payed to the role of institutions (O'Mahoney and Ferraro, 2004, 2007; Stewart, 2005) as the social construction of reality (Berger and Luckman, 1966; DiMaggio and Powell, 1983) complementing the technological domain of uncertainty (Thompson, 1967).

Assuming that communities institutionalize social roles and positions for productive agents over the time (Howison and Crowston, 2005; Stewart, 2005; O'Mahoney and Ferraro, 2007) two general research questions arise for economic production theory: *(i) do institutional factors affect the emergence of coordination-communication networks and if yes how much? (ii) do institutionalized coordination structures predict agents productivity inequality and if yes how much?*

This section provides an empirical exploration of institutional factors potentially shaping the emergence of coordination structures and then, inverting such a logic, attempts a test for the effect of such institutionalized structures on contribution differentials for productive agents. In the following paragraphs of this section first, I build empirically testable hypothesis presenting literature contributions which I already discussed broader in section one. Second, I present data organization and the research strategy which will be used for hypothesis testing. Third, the results of analytical results are presented. Finally, I discuss such results and propose further extensions and improvements for the used research strategy.

Hypothesis

In economical and managerial literatures the most accepted hypothesis for the emergence of coordination-communication networks is a contingent one. In fact, communication and other coordination forms are expected[11] to mirror (Shilling, 2001) technological interdependencies among product components (Rusnak et al, 2006; Baldwin and Clark, 2006). This line of reasoning has been tested in commercial complex product development (Sosa et al, 2007, Cataldo et al. 2006) but never in OSS projects.

Because of the absence of traditional formal hierarchies (Barnard, 1938) in OSS projects, I expect a different impact of technological interdependencies on the emergence of communication networks. In particular, also because different communication media (Bug databases and Control Version Systems) are used in order to manage the development, interdependencies among code modules should estimate a very few component of the e-mail public communication network (Mokus et al., 2002). Finally as showed in the previous

[11] There is no empirical evidence in the OSS development research. However an interesting case, reinforcing this line of reasoning, is the paper by Cataldo et al. (2006) which is developed in a commercial (but geographically distributed) software development project.

section direct communication networks do not properly reflect that all the programmers participating at discussion at a thread level are to some extent affiliated to a same task. From the above a first testable hypothesis follows:

Hypothesis 1: *technological interdependencies among software code modules have a little impact on the overall direct-communication networks.*

A second line of reasoning focused on the possible similarity between markets and OSS projects based on the self-organization principle (absence of central planners). In this case overall network metrics like degree distribution, average distance and clustering coefficient, have been used to depict collaboration structures as 'economic' exchange (Lerner and Tirole, 2002; Weiss et al, 2006; von Hippel, 2007). In the previous section I showed that implementing institutional arrangements affects the structure of communication networks. When considering such institutional arrangements like a formalization of a shared view on community members merit, rather than crude systems of rules for transactions, an issue arise about the impact of social evaluation structures on communication networks.

Even if recognizing the importance of self-organization in virtual communities, O'Mahoney and Ferraro (2004, 2007) argued that OSS projects design governance arrangements formalizing institutionalized (Berger and Luckman, 1966) status positions in communities. They used actors centrality in communication networks as a dependent variable estimating the likelihood of being appointed into such formal governance positions. Grewal et al (2006) tested the effect of both programmers and projects embeddedness (Uzzi, 1996) in communication networks as an estimator for their performance.

Very few attention has been payed to the role of social networks dyadic properties as 'dependent variables' affecting the emergence of communication networks. In particular when considering weighted (number of exchanged emails) communication networks I expect that (strong) communication ties are more likely to emerge among core developers (Mokus et al., 2002; Lee and Cole, 2003). More specifically: (i) communication among programmers with high differential in the number of others they talk with, will be not so frequent (weak ties); (ii) developers linking high communicative others will provide an integration function over the time.

Then differentials in such an integration attribute for programmers in a time period *t* will persist over the time negatively affecting the emergence of valued communication ties. (iii) because community building is a valuable activity for core developers, communication mediators will affect the shape of communication over the time. Then A second hypothesis follows from the above:

Hypothesis 2a: *actors 'distance' in both Degree and Bonachich Power has a negative effect on the emergence of strong communication links among programmer dyads.*
Hypothesis 2b: *Actors 'distance' in betweenness centrality has a positive effect on the emergence of strong communication links among programmer dyads.*

Some literature on OSS development pointed out that open projects are socio-technical interaction systems (Mokus et al. 2002; O'Mahoney and Ferraro, 2007) governed by meritocratic rules. O'Mahoney and Ferraro (2007) showed how the shared understanding of 'merit' in the Debian community shifted from a technical (writing code) domain to an organizational (community building) one. They also found that actors conformity to meritocratic shared expectations affects the shape of organizational structure. Kuk (2006) found, in a case study on the KDE developers mailing list, that information sharing by means of communication is correlated with differentials in activity levels for dyads of agents.

Following this line of reasoning I expect that programmer levels of contribution to community discussions (not necessarily direct communication) affects the likelihood for the emergence of communication ties among programmers. As pointed in the previous section direct communication networks do not allow for an accurate discrimination of 'weights' in individual contribution to the development process. A lot of actions (like committing new code) are just reflected in not directed communication networks (affiliation programmer-mailing list). In particular, I ask what is the effect of programmers global (not only direct communication) activity level to the emergence of communication networks? An empirically testable hypothesis follows from the above:

Hypothesis 3a: *programmers 'distance' in productivity has a negative effect on the emergence of strong communication links among programmer dyads in the short term.*
Hypothesis 3b: *programmers 'distance' in productivity has a negative effect on the emergence of strong communication links among programmer dyads in the medium term.*

Literature on governance structures in OSS projects (Mokus et al., 2002; Lee and Cole, 2003; Howison and Crowston, 2005) pointed out that community members organize production structures in nested levels with different roles and responsibilities. As pointed out in section one, there are no contribution in OSS literature addressing the problem of relations among roles at the overall network level. The issue of social roles (as relational patterns) and positions (as systems of connected roles) has been very central in structuralist sociology since the seventies (White et al, 1976). The idea of structural equivalence, developed by Harrison

White, occurs when two actors occupy similar positions in a social system by having structurally comparable network ties (White, Boorman and Breiger, 1974; Lorrain and White, 1971). Consider two American universities, each with active ties to different corporate benefactors, student loan providers, and state governments. The universities are structurally equivalent, that is, they occupy a similar position by having the same kinds of relationships, even though their ties are not to the same organizational partners (Smith-Doerr and Powell, 2003).

Following this insightful example and my own reflections on functional specialization of communication (see section two pp. 63-65), I expect that actors structural (White et al., 1976) and regular equivalence (which is a relaxation of the structural eq. hypothesis) could predict the emergence of communication networks. If the analyzed communication network hold a core periphery structure (as showed in section three for the httpd project) where the core is denser than the periphery, then the more two agents are 'equivalent' the more they should communicate each other. From the above an hypothesis follows:

Hypothesis 4a: *The more two productive agents in a dyad are structurally equivalent the more a communication link is likely to emerge among them.*
Hypothesis 4b: *The more two productive agents in a dyad are regularly equivalent the more a communication link is likely to emerge among them.*

The effect of past interaction has been proved to affect trust among programmers in virtual communities (Stewart, 2005; O'Mahoney and Ferraro, 2004, 2007). In effect when collaboration is open to access social evaluation of programmer reliability could be very important to reduce development uncertainty.

Hypothesis 5: *the effect of all the causal factors introduced in former hypothesis has an impact when such factors are measured at a precedent time period.*

Another exploratory hypothesis will follow from the idea of co-evolution between individual 'attributes' and social networks (Snijders, 2005). Inverting the logic adopted until now I would ask: what is the effect of institutionalized network structures on individual productivity?

Hypothesis 6: *social structures at time t-1 predict activity level differentials among programmer dyads at a subsequent time t.*

Research Strategy

Data organization

I introduced in section two the idea of a functional specialization (pp. 63-65) for communication networks along with evolutionary paths of product design. We can think that communication on different interfaces addresses different tasks for the overall project. Based on this assumption I attributed a “task kind” attribute - that is, bug, cvs, dev(elopment) and doc(umentation) - to links among programmer nodes. This idea (see figure 16 below) will be particularly useful for the computation of structural and regular equivalence.

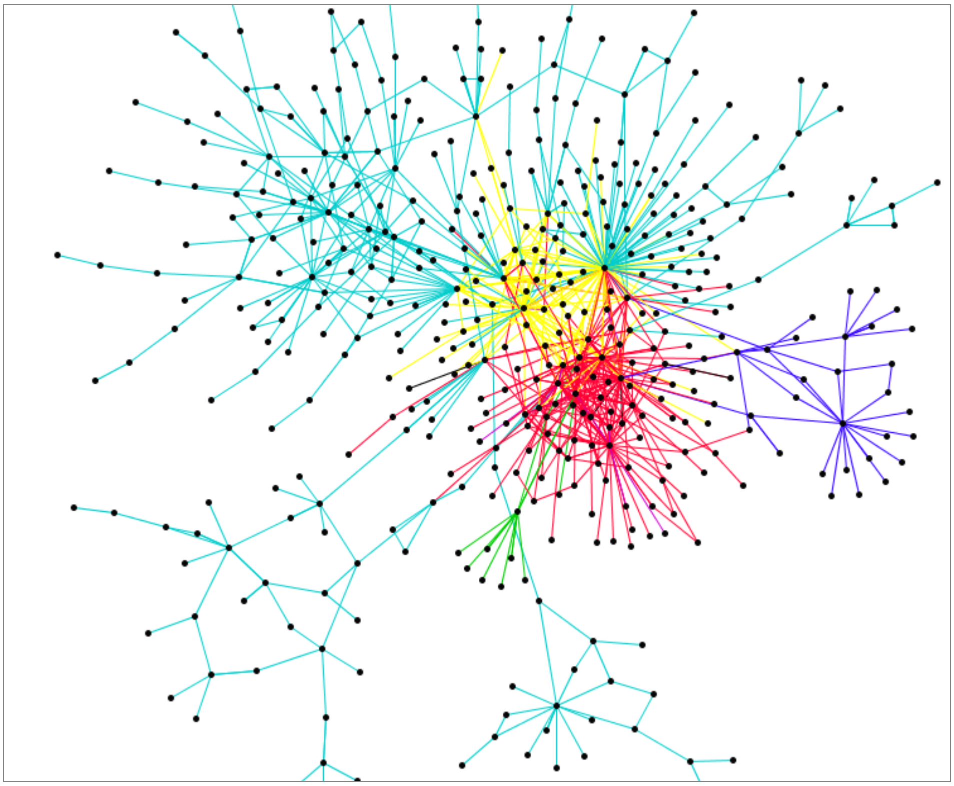


Figure 16. Direct communication network for the 'platform' project in 2002. Tie colors reflect the different mailing lists on where emails have been exchanged.

Variables

In the first regression analysis *direct communication networks*, as the likelihood of weighted communication dyads in a time point *t* have been used as a main dependent variable to be estimated with MRQUAP models.

The main independent variable is the direct *communication network* itself *at a t-1 time period*. It has been used as a proxy for reproduced strategic communication ties embedding high levels of trust over the time. The existence of a tie between two nodes at time *t-1* will be the strongest estimator for the emergence of a tie between the same two nodes at a subsequent time period *t.*

The *total* (input+output) *degree centrality* for programmers in communication networks, after being transformed in a square matrix of absolute distances, is assumed to reflect the absolute connectivity of nodes. These degree distances matrices had as both row and column labels the programmers id and, as cells, the absolute value of the degree-difference (segment) between programmers dyads.

Following the same procedure for transformation, *Bonachich Power* (Bonachic, 1983) has been used as a proxy for the integration provided among high degree communicators. Developers similarity in its values is expected to affect positively the emergence of communication ties.

I used *Betweenness Centrality* (Freeman, 1977) for agents in direct communication networks as a measure of non-redundant information channels. Increasing differentials in this measure for programmers dyads should then positively affect the emergence of direct communication ties between the same nodes.

Agent *contribution levels* introduced in hypothesis 3 are measured by the total number of email a programmers wrote in a given time period on whatever mailing list. This measure also (not reflected in direct communication) capture the activity of writing code and contributing to bug fixing because bug databases and control version systems automatically generate emails when some significant activity is performed into them.

Structural and regular equivalence have been computed directly in relational terms as the first correlation matrices generated by the UCNET procedure for this algorithm.

MRQUAP Dyadic Regression Models for Networks

Traditional inferential statistics encounters a fundamental problem when dealing with network data (Proctor, 1969), that is: should we consider the observations as independent ones when the unit of analysis is the dyad? Krackhardt (1998) proposed to frame this problem as econometricians do – as one of autocorrelation –, for example when dealing with time series. Both GLS and OLS methods hold problems in estimating the autocorrelation parameters (Judge et al. 1985: 174). A non parametric answer to this problem of testing the null hypothesis that two network variables (or two networks) are uncorrelated has been proposed (Mantel, 1967) and developed at length (Hubert, 1985).

By generating all correlations that result from permuting the rows and columns of one of the structural matrices, one can determine the distribution of all possible correlations given the structure of the two matrices (Krackhardt, 1988). Thus, it builds into the test statistic the kind of row/columns interdependence that is assumed in network data. This permutation procedure, referred to as the quadratic assignment procedure (QAP), is one answer to the aforementioned autocorrelation question. The superior robustness (standard errors) of this methods when compared to OLS ones, in presence of autocorrelated data, has been tested with Monte Carlo simulations (Krackhardt, 1988; Dekker et al. 2007).

The basic linear model for square matrix data considered here is:

$$Y = \beta X + Z\gamma + E, \quad (1)$$

where Y , X, and E are $n \times n$ matrices, β is a scalar, Z is an $n \times n \times q$ array, and γ is $q \times 1$. The diagonals of the matrices are always ignored. The null hypothesis is

$$H0 : \beta = 0. \quad (2)$$

The variables Z and X are not assumed to be independent. Specifically, we assume between these variables the linear model

$$X = \delta Z + V\,, \tag{3}$$

where V is an n × n matrix. The situation $\delta = 0$ will be called collinearity. The nonparametric approach to square matrix data means here that the residuals associated with the n objects are exchangeable or, equivalently, the matrices E and V are invariant under permutations of rows and columns simultaneously by the same permutation. Whenever the term permutation is used, it will be assumed that this permutation acts on rows and columns simultaneously and in the same way, as described above for the QAP.

In particular, I used a recently proposed method called Double Semi-Partialing (Dekker et al., 2007) defining

$$\hat{\epsilon}_{XZ} = X - \hat{\delta} Z, \tag{4}$$

where $\hat{\delta}$ is the OLS estimate for the model

$$X = \delta Z + V\,. \tag{5}$$

In the new method these residuals are permuted and the model

$$Y = \beta\pi(\hat{\epsilon}_{XZ}) + \gamma Z + E \tag{6}$$

is used to obtain reference values for the test statistic. The rationale here is that under the null hypothesis $\beta = 0$, the reference model (6) for Y is the same as the original model in (1), and if the estimation error $\hat{\delta} - \delta$ is negligible, the permutational invariance assumption for V implies that

$$\pi(\hat{\epsilon}_{XZ}) = \pi((\delta - \hat{\delta}) Z + V) \tag{7}$$

has the same distribution as V .

This new approach is referred to as Double-Semi-Partialing (DSP) regression since the effect of Z is partialed out of X and then the resulting residuals $\hat{\epsilon}_{XZ}$ are permuted and entered in a regression of Y on both $\hat{\epsilon}_{XZ}$ and Z. As such Z enters the regression twice, hence the “double”. The DSP approach minimizes the correlation between the focal variable and the control variables under permutation.

In order to test my hypothesis I used MRQUAP regression models (Krackhardt, 1988) with DSP test (Dekker et al., 2007) using network variables at three time periods – 2000, 2001 and 2002 – . As modeled in figure 17 below, I separately tested for: (1) the effect of network variables at both time period *t-1* and *t* on the communication network at time *t* (figure XX a); and (2) the effect of network variables at both *t* and *t+1* time periods on activity (differential) networks at t+1 time period (Figure XX b).

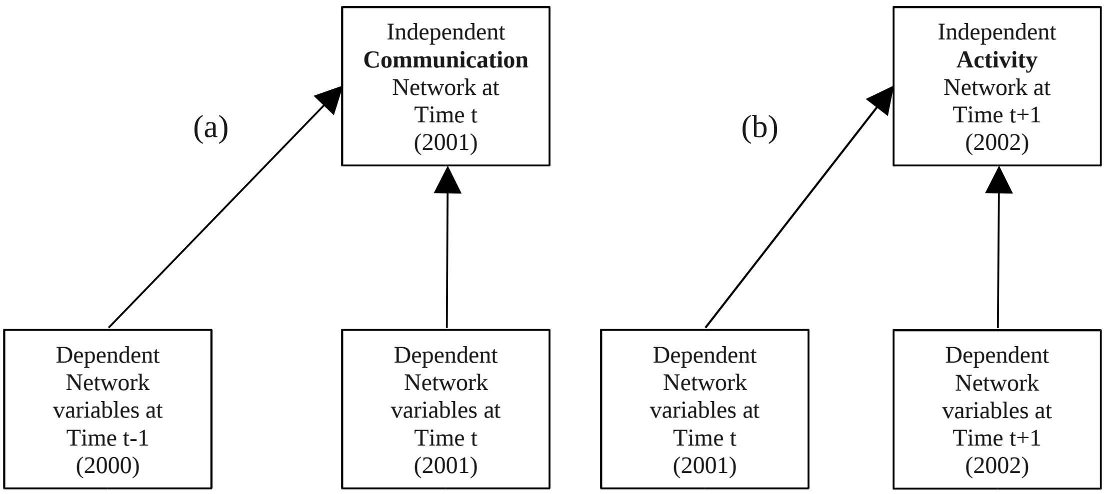


Figure 17. (a) Regression one, the dependent variable is the communication network in 2001. Independent variables are entered in subsequent models for both 2000 and 2001 ; (b) regression two, the dependent variable is the activity differential network in 2002. Independent variables are entered in subsequent models for both 2001 and 2002.

In the second regression analysis Programmer levels of activity, as dyadic differentials, have also been used as a dependent variable with the aim of exploring complex interplays with communication networks for further research.

Analysis

Here below I introduce the results of MRQUAP regression models. All data have been transformed in relational terms calculating the absolute differences among dyadic nodal values.

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**
Variable	Coef	Coef	Coef	Coef	Coef	Coef	Coef
Constant	0.0130	0.0004	0.0022	-0.0022	-0.0052	-0.0030	-0.0030
2001_coord_req.	0.0001•	0.0001•	- 0.0001•	- 0.0001†	- 0.0001†	- 0.0001••	- 0.0001••
	[-0.0093]	[-0.0093]	[-0.0092]	[-0.0067]	[-0.0041]	[-0.0224]	[-0.0223]
2001_agent_module	0.0041•	0.0031•	0.0038•	- 0.0014†	- 0.0001	- 0.0015••	- 0.0015•
	[0.0186]	[0.0140]	[0.0175]	[-0.0066]	[-0.0013]	[-0.0069]	[-0.0070]
2001_betweennes		0.0001•••	0.0001•••	0.0001•••	0.0001•••	0.0001•••	0.0001•••
		[-0.2272]	[0.2228]	[0.2339]	[0.2141]	[0.2083]	[-0.2083]
2001_bonachic		0.0001	- 0.0001	0.0001•	- 0.0001	- 0.0001	- 0.0001
		[0.0036]	[-0.0040]	[-0.0083]	[-0.0055]	[-0.0041]	[-0.0042]
2001_Freeman_deg		0.0001	- 0.0007••	0.0006•	0.0004••	- 0.0001	- 0.0001
		[0.0137]	[-0.0613]	[0.0520]	[0.0358]	[-0.0017]	[0.0001]
2001_productivity			0.0003•••	- 0.0004•••	- 0.0003•••	- 0.0002•••	- 0.0002•••
			[0.0578]	[-0.0659]	[-0.0624]	[-0.0452]	[-0.0427]
2001_regular_eq				0.0735•••	0.0721•••	0.06••	0.0613•••
				[0.0326]	[0.0321]	[0.0267]	[0.0273]
2001_structural_eq				0.0881•••	0.0876•••	0.0707•••	0.0740•••
				[0.0444]	[0.0438]	[0.0357]	[0.0373]
2000_betweenness					0.0005•••	0.0003•••	0.0003•••
					[0.1235]	[0.0861]	[0.0859]
2000_Bonachic power					- 0.0020••	- 0.0024•••	- 0.0023•••
					[-0.0372]	[-0.045]	[-0.0424]
2000_Freeman_deg					0.0009••	- 0.0001	-0.0001
					[0.0391]	[-0.0037]	[-0.0045]
2000_all						0.4345•••	0.4343•••
						[0.3681]	[0.3680]
2000_productivity						0.0001	0.0001
						[0.0109]	[0.0083]
2000_regular_eq							- 0.0025
							[-0.0009]
2000_structural_eq							0.0117
							[0.0043]
2000_cord_req							0.4323•••
							[0.3680]
2000_agent_module							0.0003•••
							[0.0859]
R-Squared:	**0.0001**	**0.0560**	**0.0575**	**0.0624**	**0.0820**	**0.2082**	**0.2082**

† p<0.10; • p<0.05; •• p<0.01; ••• p<001

Regression 1. Dependent Variable is the Communication Network in 2001.

	Model 1	Model 2	Model 3	Model 4	Model 5
Variable	Coef	Coef	Coef	Coef	Coef
Constant	1.5405	1.6750	1.8410	1.9004	1.7459
2002_betweennes	0.0041••	0.0041••	0.0042••	0.0041••	0.0040••
	[0.1012]	[0.0999]	[0.1022]	[0.1001]	0.0991
2002_bonachic	- 0.3899•••	- 0.3894•••	- 0.3909•••	- 0.3922•••	- 0.3866•••
	[-0.1350]	[-0.1348]	[-0.1353]	[-0.1358]	-0.1339
2002_freeman_deg	2.4908•••	2.4902•••	2.4827•••	2.4919•••	2.4863•••
	[0.8483]	[0.8481]	[0.8456]	[0.8487]	0.8468
2002_all		4.0734•••	4.0918•••	4.4294•••	4.9699•••
		[0.0112]	[0.0113]	[0.0124]	0.0137
2002_regular			- 4.0282•	- 4.1517•	- 3.8397••
			[-0.0162]	[-0.0167]	-0.0155
2002_structural_eq.			3.5606	3.7742	3.4423†
			[0.0057]	[0.0060]	[0.0055]
2001_betweennes				0.0059†	0.0083•
				[0.0237]	[0.0389]
2001_bonachic				0.0089	0.0053
				[0.0051]	[0.0031]
2001_Freeman_deg				- 0.1886•	- 0.5427•••
				[-0.0414]	[-0.1193]
2001_productivity					0.1941•••
					[0.0882]
2001_regular_eq.					1.6686
					[0.0019]
2001_structural_eq.					-3.6179
					[-0.0055]
2001_all					- 2.1329••
					[-0.0051]
R-Squared:	**0.7670**	**0.7671**	**0.7677**	**0.7679**	**0.7707**

† p<0.10; • p<0.05; •• p<0.01; ••• p<001

Regression 2. Dependent Variable is Activity differentials among programmers in 2002.

Discussion

The models I presented here are only exploratory ones but at least I should comment that for the first regression in each model the network at one precedent time period resulted always resulted significant estimator for the dependent network. This is the only one result which had a significant impact on the emergence of communication networks. This result could be interpreted as confirming the institutional effect of building trust based on precedent collaboration. In particular there is a consolidated literature in management about strategic alliances and trust over the time (Gulati, 1995; Gulati, 2001).

There are Possible extensions to this first regression model: (i) adding the belonging to partition clusters of structural equivalent actors based on degree centrality, or other attributes. (ii) to add attributes (in terms of distances) to MRQUAP models; (iii) estimation of similar models for networks whose nodes are mailing lists instead of programmers.

The second regression, was computed just to control for the possibility of a further influence of institutionalized communication on community structure. Surprisingly the R coefficient resulted extremely high, then a further specification of this causal relation could be the next step of future developments of this research.

SECTION IV

DISTRIBUTED GOVERNANCE FROM DISCUSSION NETWORKS IN VIRTUAL COMMUNITIES OF PRODUCTION

Introduction

In this section I investigate how discussion enables and constrains distributed systems of governance in virtual communities of production. In particular I want to cast light on how communication practices provide a coherent domain for distributed decision making linking different components of production dynamics (agents, tasks and materials).

As augmented in section one, I believe that discussion represents the fine-grained domain in which the socio-cultural artifact comes into being, arising from the practices of different communication *genres* (Yates and Orlikowski, 2002; Im et al., 2005). In this section I will empirically consolidate my point of view on distributed systems of governance claiming that communication could be conceived as a process of reciprocal influence between the material and the cultural domain of production (Giddens, 1984; Bourdieu, 1990).

Open Source Software (OSS) development is an emblematic domain in which the coordination of production process is managed in absence of centralized planning. It emerges out of an ongoing conversation among process participants. The literature on governance of OSS projects highlights that the problem of communication has been studied separating the material/technical interaction domain from the socio/symbolic one.

The discussions are the blood of the coordination process and the source of both technological and symbolic structures. By means of discussion development, management practices are produced, reproduced and selected resulting in institutions. E-mail-based discussion is itself a kind of social institution in distributed social systems. Like other institutions it evolves by means of the use of social rules.

In the following of this section I will present the overall evolutionary dynamic of communication *genres*, identified by [tags] in email subjects, over a ten year time period (1995-2004). Then, I will develop a small case studies using semantic network analytical tools. The aim of extracting semantic networks from texts is to explore how discussions enable and constraint distributed systems of governance in a virtual community of production. More precisely, I will use e-mails' contents to depict how communication is the means of coordination which shapes both technological (material) and symbolic (cultural) structures. I will analyze the communication flows, at a micro-interaction level, in detail therefor I will use email data from a project over a two weeks time period to cast light on the process of structuration of both technological and symbolic domains.

Method and Research Strategy

Virtual communities are challenging contexts for traditional research methods. Because of the distributed nature of development, virtual communities of production use email communication as the main means of coordination.

> *"Mailing lists are the life blood of Apache communities. They are the primary mode of discourse and constitute a public and historic record of the project.*

Other forms of communication (P2P, F2F, personal emails and so on) are secondary." ... "The reason is that communications on other than the public mail aliases exclude parts of the community. Even publicly advertised IRC chats can be exclusionary due to time zone constraints or conflicting time commitments by community members who might want to participate[12]*".*

We look at communication in OSS projects as a process of interaction by means of which the social structure provided by genres is enacted. Because in this environment most of developers never meet face-to-face, we consider communication in public lists as the only available reality for development practices synthesizing historic traceability, the scope of the process and thematic coherence. One way of understanding discursive genres is to examine the socially recognized or sanctioned expectations around key aspects of communication: purpose, content, participants, form, time, and location (Yates and Orlikowski, 2002). In this view, genres reveal what communities do or do not do (purpose), what they do and do not value (content), what different roles members of the community may or may not play (participants), and the conditions (time, place, form) under which interactions should and should not occur (Im et. al., 2005).

Setting: The Apache Project

The Apache project started in February 1995 when Rob McCool stopped developing his httpd-server program at NCSA[13] and then a small group of users, the so called Apache Group (AG), began a combined effort to coordinate existing fixes to the existing code. After several months of adding features and small fixes, the AG replaced the old server code base in July

[12] Quotation from Apache community building guidelines.
[13] National Center for Supercomputing Applications (NCSA).

1995 with a new architecture designed by Robert Thau.

As the core developers were distributed around the world and all of them were working at the project on a totally voluntary base, both leadership and coordination mechanism were distributed as well to take in account the limited amount of time that each programmer could devote to the project. As Roy Fielding (1999), one of the founding members, pointed out:

> *"Unlike most open-source projects, Apache has not been organized around a single person or primary contributor"..."There was no Apache CEO, president, or manager to turn to for making decisions. Instead, we needed to determine group consensus, without using synchronous communication, and in a way that would interfere as little as possible with the project progress. What we devised was a system of voting via email that was based on minimal quorum consensus. Each independent developer could vote on any issue facing the project by sending mail to the mailing list with a "+1" (yes) or "-1" (no) vote".*

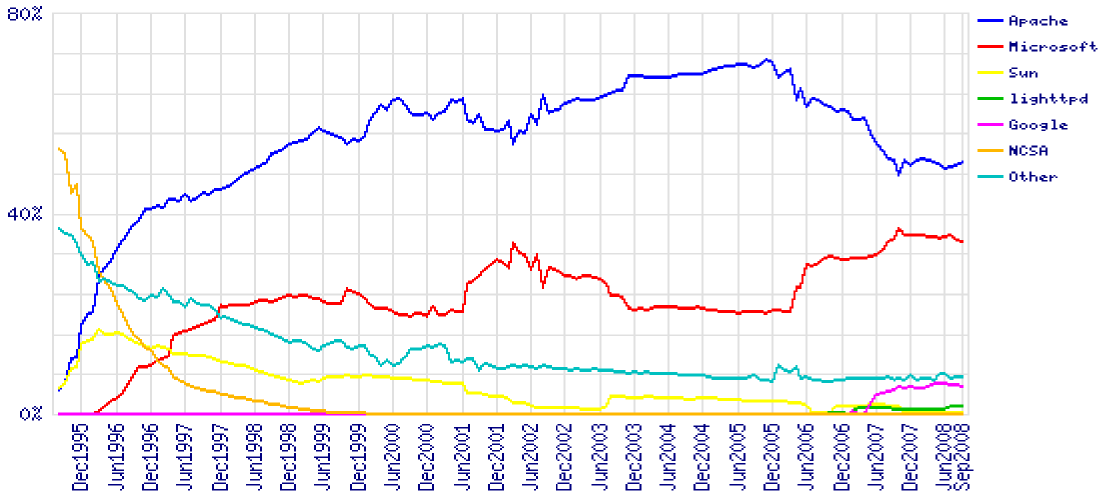


Figure 18. Servers market share across all domains. Number of domains (in percent) adopting the Apache server over time is marked by the blue-line. The main competitor is Microsoft whose market share is marked by the red line.

According to the Netcraft survey[14] in few months Apache (blu line in chart 1) became the

[14] Netcraft survey: http://news.netcraft.com/archives/web_server_survey.html. In 09-2008 the survey traced

most used server software in the world and it still is today. Microsoft which is Apache's main competitor (red line in chart 1) also became 'involved' in Apache's with a platinum sponsorship in 2008. The amount of work to be coordinated in order to maintain the software over the firs four years of development grew along with the increasing number of users. Then the Apache Group made an important step toward a more formal system of governance.

In 1999 The Apache Software Foundation (ASF) was created to provide: (i) hardware, communication and business infrastructures; (ii) a legal entity for code donations assuring that those resources will be used in the public interest; (iii) legal assistance and legitimation to new projects admitted under its identity umbrella; (iv) protection to the Apache brand from being abused by other organizations. The Apache 'software code' from this point will belong to the foundation which aims to maintain it public:

> *The Apache Software Foundation creates and maintains open source software products for the public benefit utilizing a collaborative, meritocratic approach to software development. Our products are developed by a diverse community of volunteers, a large number of whom use our software products in the course of their own daily lives. Our development discussions are held on public mailing lists. Everyone is invited to join the discussion so long as the usual courtesies of email netiquette are observed. (from the Apache guidelines[15]).*

The Apache meritocratic system of governance, also called “Apache Way” became over the time an institution for the OSS movement as whole. Literally the govern of merit means that who writes the Apache's (software) code also hold the power in institutional collective

181,277,835 sites. Of the 4.5 million sites that have been gained this month, more than 3/4 are using Apache.

15 Apache 'code of conduct' is readable on-line: http://wiki.apache.org/incubator/CodeOfConduct

decision making concerning the overall direction of development. This create a distinction between committers, who write the code, and non-committers contributing in different ways ranging from documentation, signaling bugs, providing patches, to the discussion of organizational topics.

The Apache Way is a challenging governance system for researchers studying coordination practices. In a very simplified picture, decisions are taken in two steps: (i) generating the consensus/dissensions around a proposal; (ii) vote the emergent/structuring proposal when no consensus is achieved by means of 'simple conversation'.

Data Collection

We gathered data for our study from an infrastructural mailing list belonging to the ASF where topics regarding community building are discussed[16]. The community@ mailing list was created after a period (1999-2002) of institutional re-organization. We selected an email discussion concerning the decision on how open the mailing list on community building issues should be, because of our focus on distributed governance (O'Mahoney, 2007).

The discussion is composed by 155 single e-mails. The temporal extension of the discussion (2 threads) is: 22 October 2002 – 6 November 2002. All the selected emails had the tag [vote]. Following Im et al. (2005) this means that all the selected emails belonged to the same communication genre and then those are supposed to obey to the same institutionalized rules of interaction. In order to make the distributed decisional process working, each Apache voting session should not go ahead for more than 72 hours. Because

[16] The Apache-Community list is available on-line: http://mail-archives.apache.org/mod_mbox/www-community/

we looked for a flexible use of genres as 'violation' of codes according to situations, we selected these two threads as concerning the same decision expecting that something 'did not work properly' in the first session.

Semantic Analysis

Our analysis is articulated in two main steps: (a) text pre-processing; and (b) semantic network analysis. We parsed the email text to extract single 'concepts' and used those concepts to build network representations of the decisional process to be further analyzed (Diesner et al., 2005).

Pre-Processing

As we are going to explain, text pre-processing is a fundamental requisite for semantic network analysis. We have followed three steps that we call (i) redundant information, (ii) frequency and (iii) thesaurus.

Redundant information. We have performed our analysis on a flow of 155 email; we have deleted redundant text arising from communications 'in replay to' and 'forwarded'. The presence of automatic copy in those email could be a source of biases. The deletion of redundant information is a fundamental step to preserve the text it has been intentionally communicated by individuals.

Frequency. A text is characterized by a number of words that we call concepts; each concept is characterized by a frequency. The distribution of the words in a text follows a

Power Law distribution; high-frequency words (trivial concepts) are those represented by commonly used significants such as 'the', 'a', 'have' etc. Low frequency words (idiosyncratic concepts) are those which are not relevant in the domain of discourse A fundamental step of processing text is the deletion of both tails of the distribution; in other words we have deleted concept with very high (trivial concepts) or very low frequency (idiosyncratic concepts).

Thesaurus. We have made the grain of the text coarser by bringing back similar concepts to a single concept; the loss of information represents a gain in terms of synthesis.

Semantic Network Analysis

The outcome of the pre-processing step is a set of semantic networks (one for each email). These semantic networks are made by nodes and ties, where nodes are concepts and ties are the relationships between such concepts which occur in the discourse. The weight of a tie between two nodes/concepts has been set as the co-occurrence frequency of those concepts in the same sentence. The outcome of the pre-processing step was a set of 155 semantic networks, one for each analyzed email, according to this principle. Here we introduce our semantic analytical techniques.

General Statistics and Symbols. After computing some synthetic statistics for all individual email networks, we analyzed those semantic networks looking for concepts with high measures of centrality (Wasserman and Faust, 1994). When betweenness centrality, degree centrality and 'consensus' were high the corresponding concept was labeled as a *symbol* (Carley and Kaufer, 1993).

In order to make easier the interpretation of results addressing our first research question, we consolidated all those single-mail networks in to a synthetic one. In the resulting *Consolidated Semantic Networks* the weight of ties is given by the occurrence of that tie across multiple messages.

Network Complexity. in order to address in a more 'quantitative' way our second research question we computed some standard indicator of network structure for both thread_1 and thread_2. Here we try to show how the application of a more 'structured genre' simplify the decisional process.

Semantic Networks

General Statistics. General statistics computed across all 155 semantic networks (table 1. and table 2.) show that, on average, emails in Thread_1 had a major number of concepts 37.3068 than Tread_2 (20.4478). Thread_1 is also characterized by an higher concepts' standard deviation (21.8891) when compared to Thread_2 (7.87263). Semantic networks in thread_2 are more densely connected 0.0602428 than networks from Thread_1 0.045165. At the same time the average diameter is lower in email belonging to Tread_2 (19.9254) than in email belonging to Thread_ (135.7386). Taken together these two results tell us that emails in thread_1 are, on average, more cohesive in terms of linked concepts than emails in Thread_2. Finally, semantic networks from thread_1 displayed a higher clustering coefficient than those from thread_2. This means that emails of the first set have more concepts' sub-aggregations relatively independent among them.

Table1. Statistics Across all Semantic Networks from Thread_1. Total number of semantic networks = 88

Measure	**Min**	**Mean**	**Max**	**Std.dev**
Number of concepts	5	37.3068	113	21.8891
Number of isolated concepts	0	0	0	0
Number of links	4	46.3295	162	32.1418
Density	0.0128003	0.045165	0.2	0.0310637
Diameter	5	35.7386	113	21.9689
Clustering Coefficient	0	0.0251622	0.0839026	0.0224173

Table 2. Statistics Across all Semantic Networks from Thread_2. Total number of semantic networks = 67

Measure	**Min**	**Mean**	**Max**	**Std.dev**
Number of concepts	5	20.4478	54	7.87263
Number of isolated concepts	0	0.970149	4	0.869869
Number of links	2	21.8657	76	10.4242
Density	0.0260244	0.0602428	0.166667	0.0208344
Diameter	5	19.9254	54	8.17853
Clustering Coefficient	0	0.0167306	0.166667	0.0304111

Symbols. Confronting most ranked symbols (high degree, high betweenness and high consensus), we are able to draw some additional qualitative results. In particular even if both Thread_1 and Thread_2 are of the same communication genre [vote] and the same subject 'openness', we observe that concept ranking is slightly different (see table 3. and table 4.).

In particular it seems to us that: in Thread_1, there is a sort of call for committers to express their personal views and, at the same time, a call for voting; while in Tread_2, there is a voting behavior concerning two clear proposals (vote1 and vote2). In this case it is also interesting to note that the negative vote (-1) is very highly ranked. This is because the most of committers voted negatively to the proposal of a complete openness (read and write) of the list.

Table 3. Symbols (high degree, high betweenness, high consensus) in thread_1. There are 74 concepts in this class.

Rank	Concept	Consensus	Degree Centrality	Betweenness Centrality
1	apache	0.0127932	0.086351	0.109545
2	committers	0.0149254	0.0877437	0.0998514
3	vote	0.00852878	0.0835655	0.107672
4	view	0.0140116	0.0738162	0.0930322
5	archive	0.0134024	0.0793872	0.0818668
6	org	0.0130978	0.0584958	0.0887219
7	community	0.0103564	0.0598886	0.0589193
8	sam	0.0124886	0.0529248	0.0597323
9	more	0.00761499	0.051532	0.0473578
10	do	0.00761499	0.0557103	0.0424875

Table 4. Symbols (high degree, high betweenness, high consensus) in thread_2. There are 50 concepts in this class.

Rank	Concept	Consensus	Degree Centrality	Betweenness Centrality
1	vote	0.0080292	0.0597484	0.240226
2	vote1	0.0416058	0.0880503	0.128177
3	need	0.00291971	0.0220126	0.204971
4	committers	0.029927	0.0613208	0.119199
5	vote2	0.0416058	0.0833333	0.0739063
6	-1	0.0321168	0.0204403	0.128142
7	let's	0.0423358	0.0157233	0.118811
8	community	0.00437956	0.0361635	0.129095
9	roy	0.00218978	0.0283019	0.126322
10	no	0.0306569	0.0157233	0.100321

Consolidated Semantic Networks (CSN)

CSN (see Figure 2. and Figure 3.) helped us to deeper understand the earlier findings coming from individual emails networks' analysis. In order to better 'see' in to these networks we removed links with weight (based on co-occurrence) less than 1 and then removed all the 'isolated' (disconnected individual concepts) nodes. Looking at the CSN for thread_1 (Figure 2.) the most evident result is that, after such a reduction, the concept 'view' previously ranked as 4^{th} in Table 3. (individual networks) is now the most central in terms of weighted degree (more incident high weight lines).

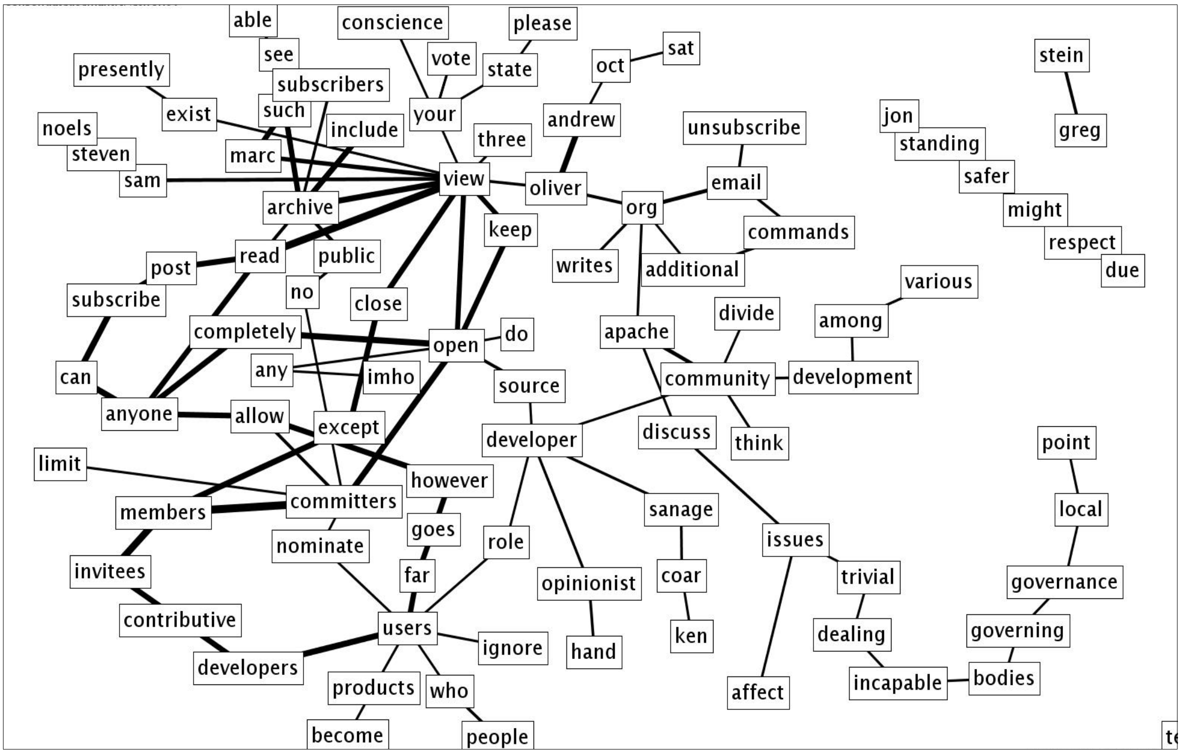


Figure 19. Consolidated Semantic Network from email thread_1. Both links with weigh less than 5.1 and isolate nodes have been recursively removed from the original network in order to offer a clearer representation.

Departing from that node we can find at least two paths corresponding to different proposals about the openness of the community@ mailing list, for example: (i) view → close → except → committers → members → invitees; (ii) view → open → completely → anyone → can → subscribe → post → read.

Another interesting path for understanding distributed governance is the one in the bottom right side of Figure 2.: local → governance/governing → bodies → incapable → dealing → trivial → issues → affect. This path expresses the major threat coming with the potential scenario where everybody is allowed to read and write on the mailing list.

So what is missing in this (Thread_1) representation? As a [vote] thread we expected a very structured communication (few densely connected concepts) and most important we expected to see as very frequent/central concepts those expressing the voting action [+1], [0] and [-1].

All these just mentioned features are actually present in the consolidated semantic network representing Thread_2. In this representation (see Figure 3.) the central sub-group of nodes is clearly expressing an email-mediated voting behavior. Concepts like 'vote1' and 'vote2' are kind of formalized proposals to be voted; '1', '0' and '-1' are the voting actions. Departing from those concepts/nodes we can find again, even if in a more stylized representation, the elements of different proposals: vote1 → 1 → yes → let's → open → everyone; or vote2 → committers → keep → private.

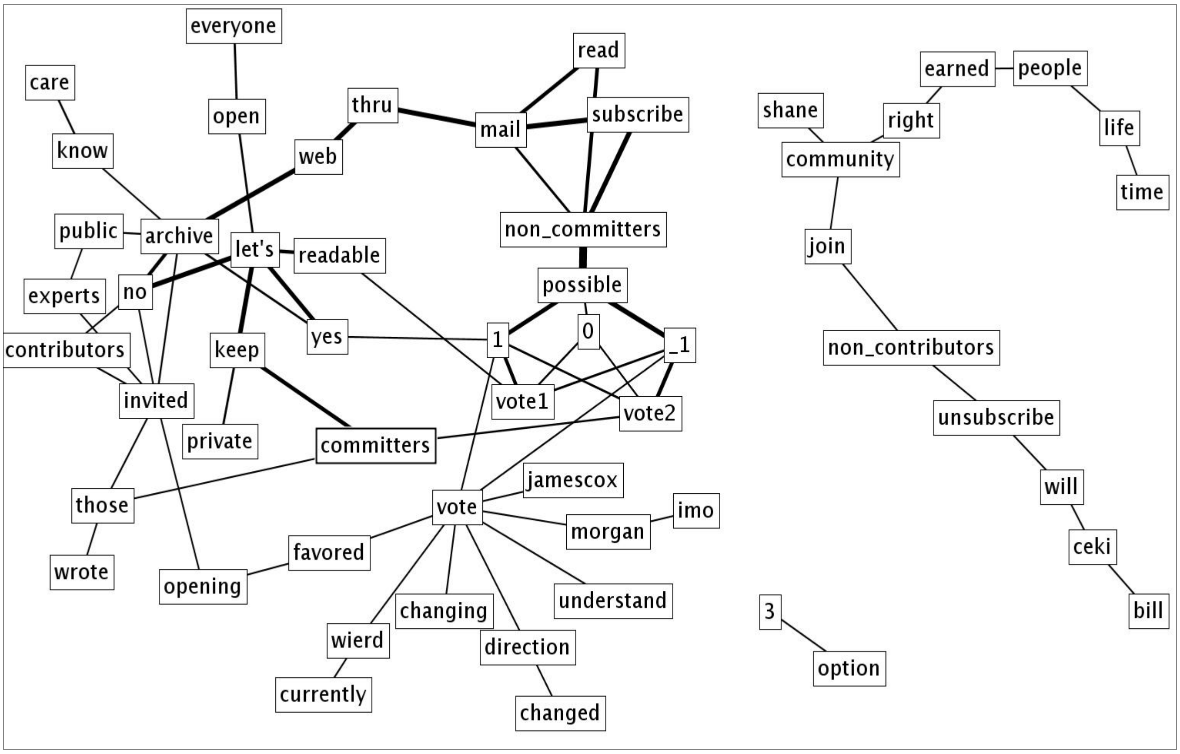


Figure 20. Consolidated Semantic Network from email thread_2. Both links with weigh less than 1.1 and isolate nodes have been recursively removed from th original network in order to offer a clearer representation.

The general governance issue for the analyzed decision was: how open the community mailing should be. In major detail the decision concerned the 'who' should be allowed to do 'what'. The options for the 'who' issue were 'committers' and 'non-committers', while the options for the 'what' issue to be decided were 'write' and 'read'. Different configurations of these elements have been formalized in a proposal with alternative 'scenarios' to be voted by community members in Thread_1.

As a second voting session have been required for the same decision we expected that some kind of re-alignment of action (voting) toward a collective decision (how-open) should have been taken. So we checked for an explanation in the first email of the Thread_2 whose body text is reported below (Figure 4.). We see here that 'voting' is thinkable as a process leaded by at least two flexible practice: (ii) dialogic consensus formation; and (ii) voting as it.

The mail in Figure 4. confirms that in Thread_1 the conversational practice used for consensus generating took a very complicated path to be interpreted as a collective decision and in doing so it obscured the 'voting as it'. This mail (Figure 4.) also confirms the flexibility of the 'vote' genre according to the situation stating that: first, “*Note: there is no need to indicate the reason for your votes, either for negative ones.*”. This happens because the individual points of view already emerged in the precedent session.; second, “*Also, please, don't vote 0.5 or other numbers, let's keep it simple for the final count.*”.

```
Please, allow me to restart the voting in order to make it easier to reach some
consensus since it's hard to interpret the results of the previous one.
                                   - o -
There are two different concerns for openness:
  - open to read
  - open to write
Let's try to keep them separate.
NOTE: there is no need to indicate the reasons for your votes, either
for negative ones.
Also, please, don't vote 0.5 or other numbers, let's keep it simple for
the final count.
                                   - o -
VOTE 1:  would you like to make it possible for non-committers to read
this mail list through a web archive?
  [ ] +1 yes, let's make it readable
  [ ]  0 don't know/don't care
  [ ] -1 no, let's keep it private
                                   - o -
VOTE 2:  would you like to make it possible for non-committers to fully
subscribe to this mail list?
  [ ] +1 yes, let's open it to everyone
  [ ]  0 don't know/don't care
  [ ] -1 no, let's keep it for committers only
                                   - o -
Please, place your vote even if you already voted in the previous poll. We'll reset
the clock and give 78 hours for the vote. I volunteer to count the results and post
them here.Thanks in advance and sorry for the double poll noise.
```

Figure 21. Body of text of the first (in chronological order) email in Thread_2.

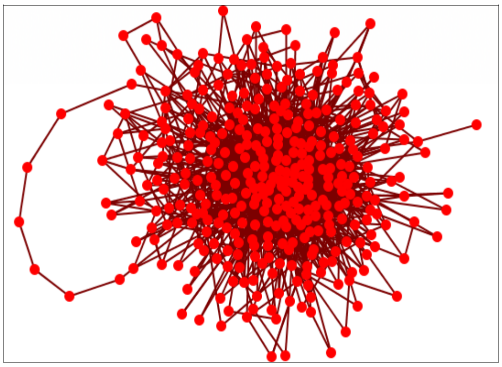
Network Complexity

As showed in the previous paragraph, conversation (as low-structured dialogic practice) is a good way to express diverse opinions, but at the same time, it could be very difficult to take a decision relying just on conversational practices. We expected that voting is a potential solution for such collective decision making because it synthesizes discussion complexity, as an high number of interrelated concepts, in a smaller number of high structured ones.

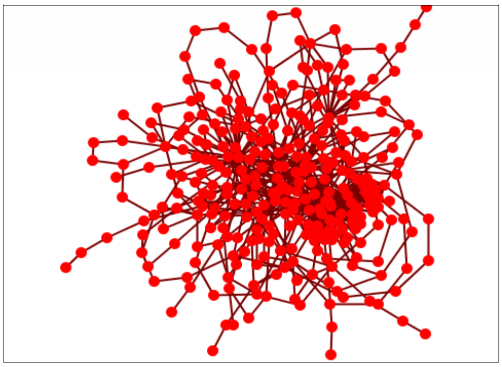
In Table 5. we report some structural indicator of overall complexity (number of concepts and connections among them) for consolidated semantic networks representing both Thread_1 (left column) and Thread_2 (right column). Indicators are computed on both the 'original' consolidated networks, that is considering all the existing concepts, and the 'reduced' networks, that is considering the same networks after removing links with weight less than 1 (two concepts that co-occurred just in one email).

Table 5. Summary of network measures for the *original* networks and the *reduced* networks. In reduced networks both links with weight less than 1 and isolates nodes are recursively removed. In the left column measures refer to Thread_1; in the right column measures refer to Thread_2.

(t1) Thread_1_original



(t2) Thread_2_original



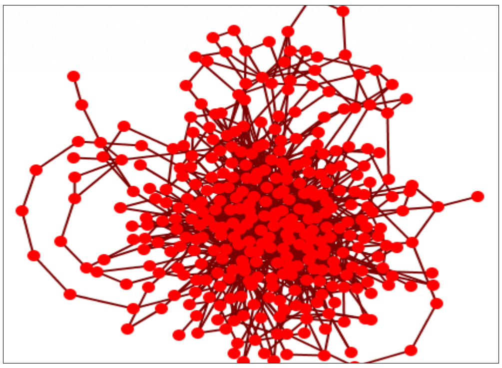
Measure	Value (original)	Value (reduction)
Row count	360	345
Column count	360	345
Link count	١۶٣٩	887
Density	0.0127	0.0075
Characteristic path length	٣.٨١٣٠	۵.٠٨۶٠
Clustering coefficient	0.0759	0.0466
Network levels (diameter)	14	١٧
Degree centralization	0.0755	0.0495
Betweenness centralization	0.1020	٠.١۶١٠
Closeness centralization*	0.2370	0.0187

* only connected component

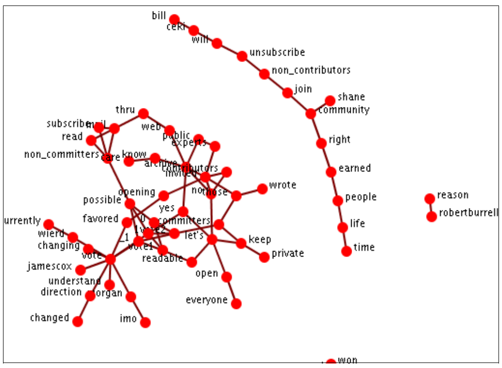
Measure	Value (original)	Value (reduction)
Row count	٣٠۶	61
Column count	306	61
Link count	۶٣١	٧٢
Density	0.0068	0.0197
Characteristic path length	6.2960	5.2960
Clustering coefficient	0.0482	0.0344
Network levels (diameter)	17	12
Degree centralization	٠.٠٨٨٩	٠.٠۵٧٢
Betweenness centralization	0.2370	0.1171
Closeness centralization*	٠.٠٣٧٩	٠.٠٢۴٨

* only connected component

(t1r) Thread_1_reduction



(t2r) Thread_2_reduction



Comparing 'original' networks we see that Thread_1 (t1) is more complex than Thread_2 (t2) in terms of: concepts/nodes 360 (t1) vs 306 (t2); and number of ties (t1) 1639 vs 631 (t2). Moreover when observing the 'reduced' networks (t1r) and (t2r) this difference is dramatically accentuated. In fact, thread_1 (t1r) has 345 nodes and 887 ties, while tread_2 (t2r) has 61 nodes and 72 ties. So when we did not consider concepts with a very low co-occurrence frequency across all the emails (which are more likely to be 'noise'), we see that the number of nodes just slightly lowers in thread_1 (-4,2%), but it dramatically falls in thread_2 (80,1%). In a similar direction links in thread_1 decrease almost of a half of the total (-45%), but links in thread _2 decrease almost as twice (-88,6%).

Discussion and Further Research

Organization and System Design

In both commercial and non commercial environments, collective decision making is gaining visibility, when compared to centralized management practices, as a low-cost/high-quality solution to manage complex and fast changing problems. Today, a lot of productive 'situations' could be characterized by a (on-line) community of user-developers determining its product life cycles. For example research and development, knowledge production and management, product customization, brand communities and so going forth.

Our study reinforce the suggestion by Faraj and Xiao (2006) that In rapidly changing knowledge-intensive environments the 'lens of practice' are more suitable to understand coordination than traditional contingent approaches. Practices have as their organizing principle not a set of conscious, constant and formal rules, but practical schemes, opaque to their possessors and varying according to the logic of situation.

Indeed the email voting practice is institutionalized in the Apache community. The Apache Group have been using it from the very beginning of its development processes in order to establish both which patches and which new features should have been applied to the existing software code. Even if email voting is institutionalized as a communication genre (Im et al, 2005), in this paper, we have found a flexible use of such a genre as consensus generating and voting occurring in emails with the same tag - [vote] -.

The set of practices for organizing, also known as the *Apache Way*, is by now a very influential one in the world of OSS communities. One of its key principles is the promotion of diversity in contributions as a way to enhance continuous innovation. According to this principle, decision are taken debating new proposals in order to leave both the consensus and the dissension emerge.

Even if work practices are very much more institutionalized in Apache than in other communities, we found that mashing voting with consensus generation, as a sort of violation of voting genre's rules, is quite well tolerated. This could be interpreted as reinforcing the general argument of practice theorists that practices are flexible to the situation (Bourdieu, 1990). In those cases the adoption of a voting genre could have been used as short-cut toward a decisions. However, our results show that mashing consensus generating and voting could result in a not so clear frame for collective decision making.

Communication and Coordination Related Research

The use of different communication genres like that we identified through email subjects (tags), but also genres linked to other communication forms like blogs, and social networking websites, is a fundamental coordination-communication practice for distributed (virtual) teams and community management.

We have found that communication genres, like 'voting', came in handy redirecting discussions toward an objective (a decision) and synthesizing the 'core' elements of the decisional process. It seems to us that the distributed governance in Apache is an ongoing synthesis between 'conversation' and 'situation' over time. Mische and White (1998) proposed that *conversation* is a discussion form in which the 'story' does not tend to a precise final, while *situation* (Mische and White, 1998) is more about the possibility of an unexpected or problematic final (which in our case is the absence of a clear policy for community management). The switching dynamic, from conversation to situation and vice versa, could be a second topic for further research.

Our contribution to the research on distributed governance systems consists of a new representation of organizational action which brings together the domain of action with the domain of culture. In the same direction Mohr et al. (2004) proposed the use of 'content analysis' and 'Galois lattices' aiming to show how institutional action is linked to symbolic categories of recipients for that action. Our research strategy is similar to Mohr's in its general intent, however we used different analytical tools. Since the use of semantic analytical tools (Carley and Kaufer, 1993) is still a relatively unexplored research strategy for decisional processes, we would suggest that an eventual further analysis in this direction could positively contribute the growing research on distributed systems of governance.

REFERENCES

Ancona, D., H. Bresman, Kaeufer K. (2002). The comparative advantage of X-teams. *MIT Sloan Management Review*, 43: 33.

Baldwin, C., B. Clark. 2006. The architecture of participation: Does code architecture mitigate free riding in the open source development model? *Management Science*, 52, 1116–1127.

Berger, B.L., Luckmann, T. (1966). *The Social Construction of Reality*. New York: Doubleday.

Bourdieu P. (1990). *The logic of Practice*, Stanford University Press, Stanford University.

Chen, K.K., O'Mahoney S. (2007). The Selective Synthesis of Competing Logics. *W. Paper*.

Carley K.M., Kaufer. D.S. (1993). Semantic Connectivity: An Approach for Analyzing Symbols in Semantic Networks. *Communication Theory*, 3, 183-213.

Daft, R.L., Lengel, R.H. (1986). Organizational information requirements, media richness and structural design. *Management Science*, 32, 554 – 571.

Diesner, J., T.L. Frantz, Carley M.K. (2005). Communication Networks from the Enron Email Corpus "It's Always About the People. Enron is no Different". *Computational & Mathematical Organization Theory*, 11, 201-228.

DiMaggio, P.J., W.W. Powell. (1983). The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields. *American Sociological Review*, 48, 147-160.

Faraj, S., Xiao, Y. (2006). Coordination in Fast-Response Organizations. *Management Science*, 52, 1155-1169.

Fielding, R.T. (1999). Shared leadership in the apache project. *Communication ACM*, 42, 42-43.

Galbraith, J. (1973). *Designing Complex Organizations*. Reading MA: Addison Wesley.

Giddens, A. (1984). *The Constitution of Society. Outline of the Theory of Structuration*. Cambridge: Polity.

Granovetter, M. 1985. Economic action and social structure: The problem of embeddedness. *American Sociological Review*, 15, 478-499.

Grewal, R. G., L. Lilien, Mallapragada G. (2006) Location, Location, Location: How Net work Embeddedness Affects Project Success in Open Source Systems. *Management Science*, 2, 1043-1056.

Im, H.G., W. Orlikowsky, Yates J. (2005). Temporal Coordination through communication: using genres in a virtual start-up organization. *Information Technology & People*, 18, 89-119.

Kuk, G. (2006). Strategic Interaction and Knowledge Sharing in the KDE Developer Mailing-List *Management Science*, 52: 1031-1042.

Lawrence, P.R., Lorsch, J.W. (1967). *Organization and Environment*. Cambridge: Harvard University Press.

Lee, G., R.E. Cole. (2003). From a Firm-Based to a Community-Based Model of Knowledge Creation: The Case of the Linux Kernel Development. *Organization Science,* 14, 633-49.

Levinthal, D.A., Warglien, M. (1999). Landscape Design: Designing for Local Action in Complex Worlds. *Organization Science,* 3, 342-357.

March, J.G., Simon, H.A. (1958). *Organizations*. New York: Wiley.

Mische, A. White, H. (1998). Between conversation and situation: Public switching dynamics across network domains. *Social Research,* 65(3), 695-72

Mockus, A., Fielding, R.T., & Herbsleb, J. (2002). Two Case Studies of Open Source Software Development: Apache and Mozilla. *ACM Transactions on Software Engineering and Methodology*, 11, 309-346.

Mohr, J.W., Duquenne, V. (1997). The Duality of Culture and Practice: Poverty Relief in New York City, 1888-1917. *Theory and Society*, 26, 305-356.

Mohr, J.W., M. Bourgeois, Duquenne, V. (2004). The Logic of Opportunity: a Formal Analisis of the University of California's Outreach and Diversity Discourse. Research & Paper Series: University of California, Berkeley http://ishi.lib.berkeley.edu/cshe/

O'Mahony, S., (2007). The governance of open source initiatives: what does it mean to be community managed?. *Journal of Management and Governance,* 11, 139-150.

O'Mahony, S., Ferraro, F. (2007). The Emergence of Governance in an Open Source Community. *Academy of Management Journal,* 50, 1079-1106.

Powell W. Di Maggio P. (1991). *The New Institutionalism in Organizational Analysis*, Chicago: University of Chicago Press.

Raymond, E. S. 1999. *The cathedral and the bazaar: Musings on Linux and open source by an accidental revolutionary*. Sebastopol, CA: O'Reilly & Associates.

Sanchez, R. Mahoney, J.T. (1996). Modularity, flexibility, and knowledge management in product and organization design. *Strategic Management Journal*, 17, 63-76.

Simon, H.A. (1957). *"A Behavioral Model of Rational Choice", in Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting*. New York: Wiley.

Simon, H.A. (1962). The Architecture of Complexity. *Proceedings of the American Philosophical Society*, 106, 467-482.

Simon, Herbert A. (1969). *The Sciences of the Artificial*. Cambridge MA: MIT Press.

Strauss, A. L. (1993). *Continual Permutations of Action*. Aldyne de Gruyter, New York.

Thompson J. (1967). *Organizations in Action*. New York: McGraw-Hill.

Uzzi, B. 1996. The sources and consequences of embeddedness for the economic performance of organizations: The network effect. *American Sociological Review*, 61, 674-698.

von Hippel, E. (1990). Task Partitioning: An Innovation Process Variable. *Research Policy*, 19, 407-418.

von Hippel, E., G. von Krogh. 2003. The private-collective innovation model in open source software development. *Organization Science*, 14, 209-223.

Von Hippel, E. (2007). Horizontal innovation networks by and for users. *Industrial and Corporate Change*, 16, 2.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications.* Cambridge: Cambridge University Press.

Weick, Karl E. 1995. *Sense making in Organizations*. Thousand Oaks, Calif.: Sage.

Yates, J. Orlikowski, W. (1992). Genres of Organizational Communication: A Structurational Approach to Studying Communication and Media. *Academy of Management Review*, 17, 299-326.

Yates, J. Orlikowski, W. (2002). Genre Systems: Structuring Interaction through Commu nicative Norms. *Journal of Business Communication*, 39, 13-35.

Sraffa, P. (1960). Production of Commodities by Means of Commodities: Prelude to a Critique of Economic Theory. Cambridge: Cambridge University Press.

Solow, Robert M. 1955–56. “The Production Function and the Theory of Capital.” Review of Economic Studies. 23:2, pp. 101– 08.

Solow, Robert M. 1957. “Technical Change and the Aggregate Production Function.” Review of Economics and Statistics. 39:3, pp. 312-20.

Solow, Robert M. 1963. Capital Theory and the Rate of Return. Amsterdam: North-Holland.

Lakhani, K., Wolf, B., Bates, J., DiBona, C., 2002. The Boston Consulting Group Hacker Survey, Release 0.73. Located at: http://www.bcg.com/opensource/BCGHackerSurveyOSCON24July02v073.pdf.

Krackhardt, D. (1988). Predicting with networks—Nonparametric multiple-regression analysis of dyadic data. Social Networks, 10(4), 359–381.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. Cancer Research, 27(2), 209–220.

Proctor, C.H. (1969). Analyzing pair data and point data on social relationships, attitudes and background characteristics of Costa Rican Census Bureau employees. Social Statistics Section, Proceedings of the American Statistical Association, pp. 457–465.

APPENDIX I: NEOCLASSICAL PRODUCTION FUNCTIONS PROPERTIES

The following assumptions are often imposed on any generic production function:

$$f : \mathbb{R}_+^m \to \mathbb{R} :$$

1. $f(x)$ is finite, nonnegative, real-valued and single-valued for all non-negative and finite x;
2. $f(0, 0, .., 0) = 0$ (no inputs implies no output);
3. If $x \geq x'$, then $f(x) \geq f(x')$ (monotonicity... an increase in inputs does not decrease output);
4. f is *continuous* and continuously *differentiable* everywhere in the interior of the prod. set;
5. The set $V(y) = \{x \mid f(x) \geq y\}$ is a convex set (quasi-concavity of f);
6. The set V(y) is closed and non-empty for any $y > 0$.

Marginal productivity. The assumptions given earlier imply that, for any given production function $y = f(x_1, x_2, .., x_m)$, it is a generally the case that, at least up to some maximum point:

$$\partial y / \partial x_i = f_i \geq 0$$

for all factor inputs i = 1, 2, ..., m. In other words, adding more units of any factor input will increase output (or at least not reduce it). This is the heart of assumption (A.3). However, it is also common in Neoclassical theory to also impose (A.5), i.e. to assume "quasi-concavity" of the production function. It is often the case in economics that the quasi-concavity assumption implies that:

$$\partial^2 y/\partial x_i^2 = f_{ii} < 0$$

for all i = 1, .., m, i.e. diminishing marginal productivity of the i^{th} factors. This means that more we add of a particular factor input, *all others factors remaining* , the less the employment of an additional unit of that factor input contributes to output as a whole.

The law of diminishing returns. Let us be clear about the definition of the *marginal productivity* of a factor. Letting Δx_i denote a unit increase in factor x_i, then the marginal product of that factor is $\Delta y/\Delta x_i$, i.e. the change in output arising from an increase in factor i by a unit. Mathematically, however, it is more convenient to assume that Δ x is infinitesimal. This permits us to express the marginal product of the factor x_i as the first partial derivative of the production function with respect to that factor, thus the marginal product of the i^{th} factor is simply:

$$\partial y/\partial x_i = f_i.$$

If we do not wish to assume that factor units are infinitely divisible or if we do not assume that the production function is differentiable, we cannot express the marginal product mathematically as a derivative. The implication, then, is that as we increase the amount of labor applied to a particular fixed amount of land, each additional unit will increase total output but by *smaller and smaller* increments.

When the field is empty, the first laborer has absolutely free range and produces as much as his body can reasonably do, say ten bushels of corn. When you add a second laborer to the same field, total output may increase, say to eighteen bushels of corn. Thus, the marginal product is eight. The basic idea is that by adding the second man, the field gets "crowded" and the men begin to get in each other's way. If that explanation does not seem credible, think of the units of labor in terms of labor-hours for a single man: in the first hour, a particular man produces ten; in the second hour he produces eight, etc. The diminution can be explained in this case as an "exhaustion" effect.

The law of variable proportions. Marginal productivity is *not* obvious in the production function $Y = f\ (L, K)$ in Figure 2 as both inputs are varying there. We must first fix one of the factors and let the other factor vary. This is shown in Figure 2, by the "reduced" production function $Y = f\ (L, K_0)$, where only labor (L) varies while capital is held fixed at K_0. To obtain this from the former, we must figuratively "slice" the hill in Figure 1 vertically at the level K_0. Thus, Figure 2, which represents the reduced production function $Y = f\ (L, K_0)$, is a vertical section of the hill in Figure 1. A reduced production function where all factors but one are held constant are often referred to as the "*total product*" curve.

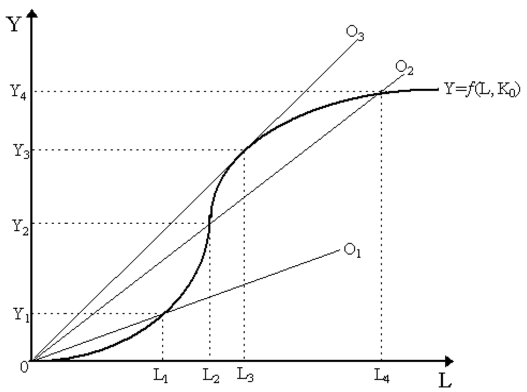


Figure 2 - Total Product Curve

The total product curve in Figure 2 can be read in conjunction with the average and marginal product curves in Figure 3. The particular shape of the total product curve shown in Figure 2 exhibits what has been baptized by John M. Cassels (1936) as *the Law of Variable Proportions.* The *marginal product* of the factor L is given by the *slope* of the total product curve, thus:

$$MP_L = \partial Y/\partial L = df(L, K_0)/dL \ .$$

As we see, at low levels of L up to L_2 in Figure 2, we have *rising* marginal productivity of the factor. At levels of L above L_2 we have *diminishing* marginal productivity of that factor. Thus, marginal productivity of L reaches its maximum at L_2. We can thus trace out a marginal product of L curve, MP_L, in Figure 3. The labels there correspond to those of Figure 2. Thus the MP_L curve in Figure 3 rises until the inflection point L_2, and falls after it. It becomes negative after L_5 - which would be equivalent to the "top" of the reduced production function. The slope of the different *rays* through the origin (O_1, O_2, O_3, etc.) in Figure 2 reflect *average products* of the factor L,

$$\text{i.e. } AP_L = Y/L \ .$$

The steeper the ray, the higher the average product. Thus, at low levels of output such as Y_1, the average product represented by the slope of O_1 is rather low, while at some levels of output such as Y_3, the average product (here the slope of O_3) is much higher. Indeed, as we can see, average product is at its highest at Y_3, what is sometimes called the *extensive margin* of production. Notice that at Y_2 and Y_4 we have the same average product (i.e. the ray O_2 passes through both points). The average product curve AP_L corresponding to Figure 2 is also drawn in Figure 3.

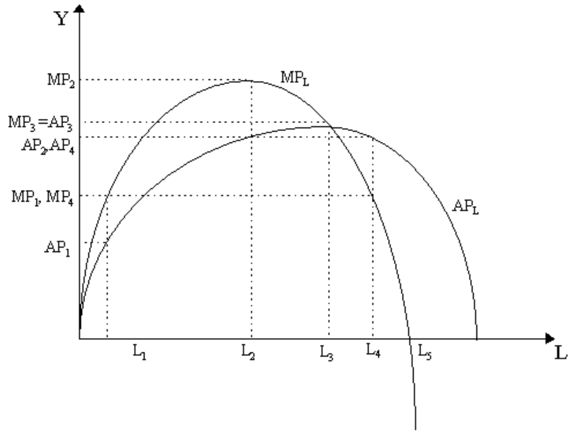


Figure 3. Marginal Product and Average Product curves

As we can see in Figure 2, the slope of the total product curve is equal to the slope of the ray from the origin at L_3, thus average product and marginal product are equal at this point (as shown in Figure 3). We also know that as the ray from the origin associated with L_3 is the highest, thus average product curve intersects the marginal product curve, $MP_L = AP_L$, exactly where the average product curve is at its maximum. Notice that at values below L_3, $MP_L > AP_L$, marginal product is greater than average product whereas above L_3, we have the reverse, $MP_L < AP_L$.

Isoquant analysis. The contours along the production "hill" in Figure 1 are the isoquants shown in Figure 4. A particular isoquant denotes the combinations of factors K and L which produce the same quantity of output. As we are assuming factors K and L are continuously substitutable (on which we will have more to say later), then every point on a particular isoquant represents a particular feasible *technique*, or factor combination, that can be used to produce a *particular* level of output. The isoquants play the same topographic role to the production "hill" as indifference curves played in the the "utility hill". As the isoquants ascend to the northeast, the amount of output produced increases, thus $Y' < Y^* < Y''$.

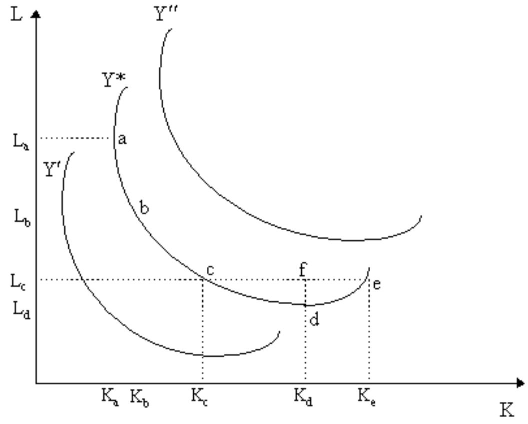


Figure 4 – Isoquants.

It is an elementary matter to derive the slope of an isoquant. For a two-factor case, we had a production function $Y = f(L, K)$. For the production of a given fixed quantity of output (call it Y*), it follows that $Y^* = f(L, K)$. This is the formula for a particular isoquant. Totally differentiating this:

$$dY^* = f_L dL + f_K dK$$

where $f_L = \partial Y/\partial L$ and $f_K = \partial Y/\partial K$ are the marginal products of labor and capital respectively, evaluated around Y*. Since on any isoquant, output is fixed at Y*, then $dY^* = 0$. This implies that:

$$f_L dL = -f_K dK, \text{ or simply: } -dL/dK|_{Y^*} = f_K/f_L .$$

The term on the left is the negative of the slope of the isoquant corresponding to output level Y*. This is known as the *marginal rate of technical substitution* (MRTS), i.e. the rate at which capital can be susbstituted for labor while holding output constant along an isoquant.

(note that dL/dK by itself is already negative, thus the MRTS will be a positive number). Provided our isoquants are smoothly differentiable, we will be able to define the MRTS at any point in Figure 4. Thus, the MRTS depends not only on the level of output (which isoquant we are on), but also the amounts of capital and labor (where on the isoquant we are).

The equality of the MRTS with the ratio of marginal products of capital and labor, f_K/f_L, is a fundamental feature of production theory and helps us capture the concept of diminishing marginal productivity to a factor. In Figure 4, on isoquant Y*, as we move from point a to b to c to d, we are moving towards greater employment of K and less employment of L to produce a given level of output Y*, thus we are moving from *labor-intensive* techniques (i.e. low capital-labor ratios) towards *capital-intensive* techniques (high capital-labor ratios). Notice also that the isoquant becomes flatter as we move from a to d, thus the marginal rate of technical substitution is *higher* at a than at d, i.e. $MRTS_a > MRTS_b > MRTS_c > MRTS_d$. Thus, there is *diminishing* marginal rates of technical substitution as we move from a towards d.

We should note, however, that *not* every point along the isoquant is relevant. The isoquants, after all, are contours of our "production" hill and thus are actually "circular". This is captured in Figure 2.5, where we show the isoquants in their full topographic glory as a horizontal section of the production hill of our earlier Figure 1. Notice that the isoquant labels represent increasing output levels, $Y < Y' < Y'' < Y'''$, etc. The "top of the hill", the highest output achievable, is represented by point M in the center, achieved by factor combination L_M and K_M. Notice that if we are the top of the hill, *if* we increase factor inputs (above K_M or L_M), output will actually *decline*.

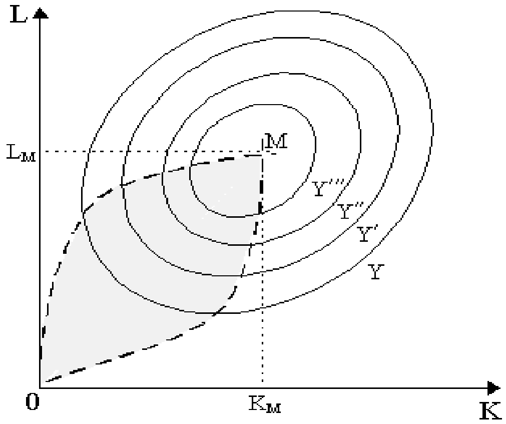


Figure 5 - Isoquants with Ridge Lines.

We have also added dashed "ridge lines" to the topographic map in Figure 5. Only those points *within* the ridge lines, in the lightly shaded region, are of economic relevance. To see why, return to Figure 4 and notice that at point a, the isoquant has a vertical slope and a point d, the isoquant has a horizontal slope. Thus:

$$MRTS_a = f_K/f_L|_a = \infty \text{ ; and } MRTS_d = f_K/f_L|_d = 0 \text{ .}$$

But our isoquants seem to continue beyond them, yet we assert that points beyond them are economically irrelevant. Why? Consider a factor combination such as at point e in Figure 4. Obviously, here, the slope of the isoquant is positive,

i.e. $dL/dK|_e > 0$, which implies, in turn, that $MRTS_e = f_K/f_L < 0$,

thus the marginal product of one of the factors is negative. This violates the first assumption we made about the production function: namely, that $f_i > 0$ for all i, i.e. increasing the employment of any factor in a production process will always increase output. Thus, we ought to exclude all regions where marginal products are negative. Is this assumption reasonable? Well, notice that at point e, we are employing factors K_e and L_e to produce output level Y*. Yet, we could *decrease* the amount of capital employed to K_d and leave labor at L_e in order to achieve a combination at point f. But notice that as point f is *above* the isoquant Y*, it effectively represents a *higher* level of output. Thus, if we are at a point such as e, then by *reducing* factor inputs we can *increase* output: such factor combinations are therefore not "economical". Consequently we can rule out point e - and, indeed, all factor combinations on the isoquant Y* beyond d in Figure 4. Similarly, we exclude points on the isoquant beyond point a for the same reason.

The "ridge lines" drawn in Figure 5, pass through limiting points of the various isoquants akin to points a and d in Figure 4. In other words, at any point on the upper ridge line, MRTS = ∞ for the relevant isoquant, while at any point on the lower ridge line, MRTS = 0 for the relevant isoquant. Thus, we *exclude* all regions above the upper ridge line and below the lower ridge line as economically irrelevant. Only the lightly shaded area in Figure 5 is "relevent". The ridge lines meet at point M, the "top" of the production "hill". The definition of quasi-concavity we used in (A.5) states that:

$$V(y) = \{x \mid f(x) \geq y\} \text{ is convex.}$$

In other words, a function is quasi-concave if the *upper contour set* V(y) is convex. As we see, this "upper contour set" V(y) is merely the isoquant defined by y and the area above that isoquant. This is illustrated in Figure 6. Y* is the isoquant of relevance, thus V(Y*), the shaded area, is the upper contour set.

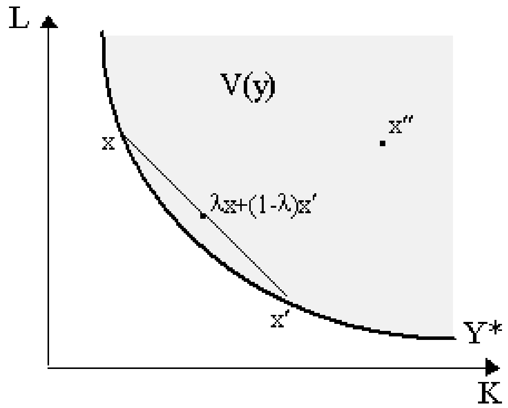


Fig. 6 - Upper Contour Set and Convexity.

It is worth pointing out now what the assumption (A.6) on production meant. Effectively, if V(y) has an interior point, then (A.6) intuitively states that for any upper contour set V(y), there is a factor bundle "inside" it (e.g. x′ ′ in Figure 6). In order to guarantee (A.6), it is common practice in to impose the assumption of *free disposibility* in production. What this assumption states, effectively, is that if one can produce a particular level of output y with input bundle x, then one can produce an amount of output which is *less* than y with that same input bundle, x or, equivalently, one can produce the *same* level of output y with *more* inputs.

Thus, in Figure 6, x′ ′ can be used to produce y if the extra amount normally produced by x′ ′ is "thrown away" or if the extra factors are just left unused by the producer. The free disposal assumption, then, is meant to guarantee at least the technical possibility of "inefficient" production, and thus interior points to the production set and the input requirement sets. However extensively used in production theory, it might be regarded nonetheless as somewhat stronger assumption than one might wish for and, thus many economists have endeavored to dispose of it.