# The Pattern of Knowledge Flows between Technology Fields: modularity, and autocatalytic sets. (Preliminary and incomplete draft)

Mauro Caminati*, Arsenio Stabile**
University of Siena
*Department of Ec. Policy, Finance, and Development
**Faculty of Economics

November 2007
Keywords: patent citations, modularity, autocatalytic set
JEL classification: O3

***Abstract*** **:**

This paper exploits recent contributions to the notions of modularity and autocatalytic sets, to identify the functional and structural units of an empirical knowledge pattern that define the strongest systematic and self sustaining mechanisms of knowledge transfer and accumulation within the network. These 'core' structures are defined by the connectivity property that every node (technology field) in the core is connected to every other node in the same core by a circular self-sustaining information flow. Our analysis reconstructs the architecture of the empirical knowledge pattern based on USPTO patent citation data at the level of resolution of 3-digits technology classes, for the period 1975-1999. Based on this fine grained analysis, the changes through time in the cross-field architecture of knowledge transfer are investigated. Our results are consistent with the idea that the information and communication technologies (ICT), although representing the core of knowledge creation throughout the period, only in the second half became fully integrated with the other sectors.

## 1 Knowledge pattern and innovations

We think of the disciplinary technological knowledge available in the economy as subdivided into different disciplinary fields, or simply fields, which only partly correspond to application domains. A field identifies a set of funcional or '*phenotipic*' traits in the broad technology domain. The set of technology fields existing at a given date and the relations between them are understood to be the outcome of the way in which the human society has historically explored and exploited the set of possibilities offered by the physical world. Technological determinism is therefore inappropriate, in that there is a social element inherent to the architecture of technology. In this paper, innovations, that is, additions to the stock of technological knowledge, are distinguished between incremental, radical, and network. The scope of a disciplinary-field definition is sufficiently narrow that radical innovations are understood as exogenous events affecting the number and quality of technology fields. This may occur through the disclosure of a new application domain, such as wireless communication, or through the provision of a new type of input, such as electric power. Network innovations are changes in the social relations affecting the cross-field organization of R&D activity. They amount to changes in the matrix describing the active cross-field learning interfaces and the strength of these knowledge connections. Incremental innovations are those additions to the knowledge stock originating in a specific field that do not affect the set of available fields or the design of cross field interactions.

The way in which the structure of the knowledge pattern evolves through time is shaped by radical and network innovations. This paper proceeds on the bold hypothesis that the organization of a knowledge pattern in a given historical period, say the last decades of the 20th century, reflects not only the key technological interfaces that are dominant in the period, for instance those concerning the information and communication technologies (ICT). We expect that the organization of these interfaces will partly reflect more general principles, bearing upon the way in which the accumulation of new ideas over time affects the difficulty of innovation activity.

The difficulty in question tends to increase with the complexity of interactions between technological knowledge components (subsection 2.1 ). Thus, evolvability in the technological knowledge domain requires, much like in other domains, that complexity does not grow in propotion with the inevitable growth in the scale of the system (as measured for instance by the number of technology fields). As we shall see, there is convincing evidence that this is achieved through a selection for modularity in the organization of the learning interfaces between technology fields.

We follow Newman and Girivan [25] method of defining the community structure within a network based on their notion and measure of modularity. This is more general than Simon's [35][37] near decomposability, in that 1st order magnitude links between modules are permitted. Beside the structural viewpoint of partitioning the network of technology fields into modules, we adopt also a functional viewpoint. Through the separation of first-order from lower-order magnitude links and by exploiting the notion of autocatalytic sets, we identify the functional units that define the strongest systematic and self sustaining mechanisms of knowledge transfer and accumulation within the network. We expect that in an empirical knowledge pattern the dominant core units so identified correspond to the subset of technology fields and knowledge interfaces that together define the dominant technology paradigm of the period. As will turn out, the prediction is fully corroborated in the empirical analysis to follow. This explores the architecture of the empirical knowledge pattern in the period 1975 - 1979, as derived from the USPTO patent citation data [14] at the level of resolution of 3-digits technology classes. Based on this fine grained analysis, directions of structural change within the period are also investigated. They are broadly consistent with the idea that the information and communication technologies (ICT), although representing the core of knowledge creation throughout the period, only in the second half became fully integrated with the other sectors.

The paper is organized as follows. In the next section we introduce our formal description of a knowlegde pattern and the precise notions of ruggedness, modularity, near-decomposability, and core structures that will be used in the rest of the paper. Section 3 relates our reconstruction of the empirical knowledge pattern from the NBER files of patent-citation data to the growing literature on knowledge spillovers based on patent citations. Section 4 exploits the notions of modularity and ACS to analyze the architecture of the re-constructed empirical knowledge pattern in the periods 1975-1986 and 1987-1999. Section 5 concludes.

# 2 Knowledge pattern, modularity and autocatalytic sets

## 2.1 Technological landscapes are rugged

To clarify this point, it is best to think of an idea as a specific configuration of a set of basic codifiable knowledge components. Ideas discovered in different fields may share some of their basic components and for this reason R&D in one field may be relevant to R&D in others. Our premise is that exploiting the 'relevance' of a knowledge input to the discovery of a knowledge output requires that the configuration of the latter conforms to a number of constraints imposed by the configuration of the former. The reason is that the relative fitness of ideas is strongly affected by relations of interdependence or complementarity, that produce many and possibly poor local optima in performance space. As a result, finding a better position in this space may require the simuoltaneous change in the configuration of many knowledge components. This makes the problem of finding the best configuration of a *given* set of knowledge components difficult. In other words, technological fitness landscapes are rugged [19]. [1]

Ruggedness does not fully define the complicatedness of R&D in one field. The reason is that, the *shape* of the fitness landscape and possibly its dimension, changes through time as a result of deformations induced by discoveries in R&D laboratories operating in other fields (there is strategic interdependence between R&D choices). In other words, R&D landscapes of different disciplinary fields

[1] In the landscape metaphor, the output of incremental R&D is an expansion in the known surface of the given landscape. At a given date, the dimension of the landscape depends on the number of basic knowledge components available, which we assume a strictly increasing function of the number of known fields. The dimension of disciplinary R&D landscapes tends to increase through time together with the number of fields, as a result of radical innovations.

are more or less tightly coupled[2]. Taken together, ruggedness and degree of coupling are the sources of the '*complicatedness*' facing R&D activity.

## 2.2 The connection matrix $C$

We consider an economy with a finite set $S = \{1, ..., n\}$ of known technology fields. A field $j$ is here understood as a (possibly infinite) set $T_j$ of potential configurations, or designs. The technological state of the economy is defined by $\{G(S, L, C), A\}$. $A_i$ $i = 1, ..., n$, is the number of useful ideas cumulatively produced by R&D in field $i$. $G(S, L, C)$ is a weighted directed graph, with a set $S$ of nodes, that are here interpreted as technology fields, a set $L$ of directed knowledge links between these nodes, and a connection matrix $C$ of weights, or intensity coefficients, attached to the links in question. $c_{ij}$ is the strength of the directed link from $j$ to $i$. It is a measure of the extent to which ideas developed in sector $j$ are relevant to R&D in sector $i$, in the sense that $A_j$ expands the knowledge base of the latter.

The average number of incremental innovations per unit of time in a given field depends on two main factors: in the first place, the set of innovation opportunities available in that field, in the second place the innovation effort in the same field. In this paper we assume that innovation opportunities are primarily determined by the progressive local knowledge base. This consists of the subset of ideas that are known by R&D laboratories currently operating in the given technology field and that are potentially conducive to useful recombinations and developments leading to new disciplinary knowledge. Under a recombinant interpretation of knowledge growth (Reiter [31], Weitzman [39]), the progressive knowledge base can be regarded as the repertoire of recombination possibilities from which innovations will originate. A larger knowledge base provides a larger repertoire of recombination possibilities, but, simultaneously, makes R&D activity more difficult, because search spaces are more complex. On this ground, we hold to Aghion and Howitt's idea ([1]) that when the effect of innovation activity on innovation output is at stake, R&D activity in field $i$ is best measured by effective R&D effort $Q_i/A_i$, where $Q$ is unweighted R&D investment.

So defined, the progressive local knowledge base partly consists of ideas originated from past innovations in the same technology field, but will also partly consist of ideas originated from past innovations in other fields that are made available to the field in question by the knowledge interfaces that are currently active across technologies. The knowledge pattern is the set of knowledge interfaces that are active across fields, together with their degree of activation. More precisely, the intensity $c_{ij}$ of knowledge transfer from field $j$ to field $i$ is the average frequency with which one innovation in field $j$ gives rise to ideas that are

[2] S. Page [27] associates '*difficulty*' of search with the fact that fitness lanscapes are rugged, so that difficulty increases with ruggedness, and associates '*complexity*' of search with the fact that fitness landscapes are coupled in a way that a search step in one induces a deformation in the others. In Page's definition, complexity is the a measure of how tight is the coupling between the landscapes. In the sequel we shall not exploit Page's distinction.

relevant to one unit of effective R&D effort in field $i$.[3] We may note, in passing, that the same idea discovered in one field, may be relevant to many other fields; hence, there is no implication that the elements in the columns of $C$ add up to 1[4]. Knowledge produced by past innovations in one field is always relevant to R&D activity in the same field, that is, $c_{ii} > 0$, $i = 1, ..., n$. By definition, $C$ satisfies the condition: $c_{ij} = 0$ if and only if the directed link $(j \to i) \notin L$. This justifies the definition:

**Definition 1** *$G(S, L)$ is the unweighted directed graph associated with the weighted directed graph $G(S, L, C)$, or, more sinthetically, with $C$.*

The discovery which brings $j$ in the set $S$ of known technologies, brings also the knowledge stock $A_j$ to its lower bound $A_j = 1$; after that, $A_j$ grows as a result of the cumulative flow of incremental-innovation arrivals in the technology field $j$. Let $a_i = A_i / \sum_j A_j$. A companion paper written with Serena Sordi [6] builds a dynamics of the column vector $a$ of share distributions $a_i, i = 1, ..., n$, driven by the flows of knowledge inputs across fields. The flow of useful innovations in sector $i$ depends on two factors, the effective R&D effort in this field, $Q_i/A_i$, and the repertoire of available ideas that are the 'building blocks' of R&D in field $i$; this repertoire corresponds to the knowledge flows $\sum_j c_{ij} A_j$ received by $i$ through the active interfaces described by $C$. The stock $A_i$, $j = 1, ...n$, evolves according to the differential equation:

$$\dot{A}_i = \sigma \frac{Q_i}{A_i} \sum\nolimits_j c_{ij} A_j \tag{1}$$

where $\sigma$ is a uniform productivity parameter. It can be redily verified that in the long-run condition such that the effective R&D effort $Q_i/A_i$ is uniform across fields, every right eigenvector of $C$ is a dynamic equilibrium of the differential equation above. Under the hypothesis that relative R&D effort in field $i$, $Q_i / \sum_i Q_i$, increases (decreases) depending on the extent in which innovation opportunities in this field, $\sum_j c_{ij} A_j$ is higher (lower) than average, it is proved[5] that the dynamics of $a$ converges to a fixed point $a^*$ which is the right eigenvector of $C$ associated with the Perron-Frobenius eigenvalue $\lambda^*$. (Using a genericity argument, $\lambda^*$ is assumed to have multiplicity 1). The suggested interpretation is that $\lambda^*$ is the highest long-term sustainable rate at which ideas are produced within the subset of leading technology fields. These fields correspond to the positive entries in the (right) eigenvector of $C$ associated to $\lambda^*$. Taken together, they define the dominant technology paradigm.

[3] $c_{ij}$ will not fully capture the occasional transfer of a radically new idea from $j$ to $i$, unless we have a reliable way of weighing the importance of ideas. The absence of such reliable weights is made less dramatic by the fact that a radical innovation is normally followed by a swarm of incremental innovations, so that the frequency $c_{ij}$ of systematic knowledge transfer will at least partly reflect the importance of ideas.

[4] Indeed, it follows from our definitions that the case $c_{ij} > 1$ can not be ruled out.

[5] The proof is given in [6] for low dimensional $n$. By way of simulations, the result is conjectured to hold for any given $n$.

## 2.3 Modularity of $C$

Under the above interpretation, the exploitation by field $i$ of a relevant input idea discovered in field $j$ requires meeting the constraints carried by such input. A relatively high (low) value of $c_{ij}$ is also an indication that the landscape of field $i$ is tightly (weakly) coupled to that of field $j$; the present directions of useful discovery in the former are strongly (weakly) conditioned by the direction in which the configuration of useful ideas has been developing in the latter.

Mutually high values of $c_{ij}$ and $c_{ji}$ signal a coevolution of the directions of discovery in fields $i$ and $j$. The constraints imposed by such a coevolution bring in a trade-off. The price to pay for the oppotunity of a faster progress through the formation of compatible knowledge standards in fields $i$ and $j$, and the opening of active interfaces between them, is a reduced capability to move away from the direction specified by those standards. This creates the danger of a technological lock in, because trajectories traced by local search procedures are path dependent.

It is worth stressing that under our interpretation, the search 'complicatedness' faced by R&D in field $i$ is not linearly additive in the parameters $c_{ij}$ composing the *ith* row of $C$ . The reason is that the exploitation of spillovers from a larger number of diverse fields requires compliance to a wider set of qualitatively different constraints. The complexity of the search space facing R&D in field $i$ comes to depend not only on the total sum $\sum_j c_{ij}$ , but, more importantly, on the distribution $\left[\frac{c_{ij}}{\sum_j c_{ij}}, i = 1, ..., n\right]$ and on the technological diversity between the fields from which field $i$ draws its knowledge inputs. Ceteris paribus, R&D receiving its knowledge inputs from a smaller number of qualitatively more similar technology fields is expected to face a less complicated search space.

The observation above identifies a strong incentive for field $i$ to concentrate the incoming knowledge links of total intensity $\sum_j c_{ij}$ across a restricted number of technologically similar source fields. In other words, we expect a selection for modularity in the structure of $C$. Intuitively, the set of $n$ fields can be partioned into $m < n$ disjoint groups, such that , *on average*, and in ways that will be specified below, the within group links are stronger than the between group links. It is also worth observing that the argument can be replicated at different hierachic levels; but to the extent that there is qualitative variation in the nature of technological constraints, activities and functions at different levels of the hierarchy, there is no direct implication that the organization of knowledge patterns is necessarily self-similar across modules and at every scale of resolution.

## 2.4 Modularity and dynamics: near-decomposition, aggregation and core structures

The above intuitive and quite genaral idea of modularity of the connection matrix $C$ admits a quantitative expression, based on recent contributions in network theory and applications. Suppose that the set $N = \{1, ..., n\}$ of technology fields is partitioned into $m$ disjoint subsets, or groups, so that $N = N_1 \bigcup N_2 \bigcup ... \bigcup N_m$, where $N_h$is the set of fields belonging to group $h$. The total intensity of an outword link from group $h$ directed to itself or to other grpups is $\hat{a}_h = \sum_i \sum_j c_{ij}, j \in N_h, i = 1, ..., n$. The corresponding total intensity of an inword link to gruop $h$ from itself or from other groups is $\check{a}_h = \sum_j \sum_i c_{ij}, i \in N_h, j = 1, ..., n$. If the total intensity of links in $C$ is $T = \sum_i \sum_j c_{ij}, i, j = 1, ..., n$, then the average relative frequency with which an outword link in $C$ originates from, and arrives to, group $h$ is $\hat{e}_h = \frac{\hat{a}_h}{T}$ and $\check{e}_h = \frac{\check{a}_h}{T}$, respectively. The modularity measure $Q_h$ of the links from and to gruop $h$ in the context of the given network $C$, is then expressed by the extent in which the frequency of within-group links exceeds the frequency which would be expected from the hypothesis of a random wiring.

$$Q_h = \left[ \sum_{i \in N_h} \sum_{j \in N_h} c_{ij} \right] - \hat{e}_h \check{e}_h \tag{2}$$

The modularity of $C$ according to the partition $\{N_1, ..., N_m\}$is then expressed by the sum $Q = \sum_h Q_h$, which may be negative, if the partition is ill-chosen. Indeed, the relative goodness of two alternative partitions of $N$ is evaluated by choosing the partition yielding a higher value of $Q$. In this spirit, the modularity of $C$ is defined by selecting the $Q$-maximizing partition ([25]). Since the $Q$ modularity of the null partition $\{N\}$ is zero, the $Q$ modularity of $C$ takes values in the interval $[0, 1]$.

A strong form of modularity corresponds to Simon and Ando [37] near-decomposability, which applies if there exists a permutation of rows and columns such that $C$ takes the form:

$$\begin{bmatrix} C & C & c & c & {\scriptstyle c} & {\scriptstyle c} & {\scriptstyle c} & {\scriptstyle c} \\ C & C & c & c & {\scriptstyle c} & {\scriptstyle c} & {\scriptstyle c} & {\scriptstyle c} \\ c & c & C & C & {\scriptstyle c} & {\scriptstyle c} & {\scriptstyle c} & {\scriptstyle c} \\ c & c & C & C & {\scriptstyle c} & {\scriptstyle c} & {\scriptstyle c} & {\scriptstyle c} \\ {\scriptstyle c} & {\scriptstyle c} & {\scriptstyle c} & {\scriptstyle c} & C & C & c & c \\ {\scriptstyle c} & {\scriptstyle c} & {\scriptstyle c} & {\scriptstyle c} & C & C & c & c \\ {\scriptstyle c} & {\scriptstyle c} & {\scriptstyle c} & {\scriptstyle c} & c & c & C & C \\ {\scriptstyle c} & {\scriptstyle c} & {\scriptstyle c} & {\scriptstyle c} & c & c & C & C \end{bmatrix}$$

In this case, if the ratio

$$\frac{c}{C}$$

is sufficiently close to zero, the linear operator $C$ is decomposable into the sum:

$$C = C^* + \varepsilon D$$

where $C^*$ is block diagonal and $\varepsilon$ is sufficiently small. If a square non negative matrix like $C$, decomposable as above, describes the equations of motion of a (locally) linear dinamical system of $n$ variables $X = (X_1, ..., X_n)$ of the form[6]

$$\dot{X}_t = CX_t \tag{3}$$

the diagonal operators $C_1^*, ..., C_m^*$ act on the components of $X_t$ corresponding to the partition $\{N_1, ..., N_m\}$, respectively. During the short run, the dynamical behavior is dominated by the diagonal block operators acting on the relevant components of $X_t$, that is, it is almost completely determined by the within partition relations. During this interval the dinamical system is decomposable, but is not aggregable, because the within-partition components of $X_t$ have not yet completed their convergence to the dominant eigenvectors $a_1^*, ..., a_m^*$ of the diagonal block operators. This convergence marks the inception of the medium run. During this interval the within partition dynamics approximated by $C^*$ still dominates, so the system is still decomposable, but to the extent that the within partition distributions are closely approximated by $a_1^*, ..., a_m^*$, the system is also aggregable. In the long run, the between partition relations become relevant and for this reason $C^*$ does not offer a good approximation of the dynamics any longer. During this interval the changes in $X_t$ induced by the between partition relations are weighted by the equilibria of the within-partition distributions. For this reason the system is still aggregable, even though it is no longer decomposable.

A system like 3 may lend itself to forms of aggregation and decomposition of variables even in situations where Simon's near-decomposability fails. To clarify this point, which plays an important role in the sequel, it is worth considering an example concerning the dynamics induced by 3 on the share distribution variables:

$$x_{it} = \frac{X_{it}}{\sum_{i=1}^{n} X_{it}}$$

$$\dot{x}_i = \sum_{j=1}^{n} c_{ij}x_j - x_i \sum_{j,k=1}^{n} c_{kj}x_j \tag{4}$$

In this example the operator $C$ takes the form:

$$C = C^* + \varepsilon D = \begin{bmatrix} c_{11}^* & c_{12}^* & 0 & 0 \\ c_{21}^* & 0 & 0 & c_{24}^* \\ c_{31}^* & 0 & c_{33}^* & 0 \\ 0 & 0 & 0 & c_{44}^* \end{bmatrix} + \varepsilon D$$

where $D$ , $C^*$ are $n \times n$ non negative matrices, $n = 4$ and $\varepsilon$ is 'sufficiently small'. Here near-decomposability fails, because $C^*$ is not block diagonal. The short run dynamics of the relative share distributions of $(x_{1,t}, .., x_{4,t})$ converges to the share distribution of the right eigenvector of $C^*$ associated to its dominant eigenvalue $\lambda^*$. The aggregation referred to above is induced by the dominant eigenvector properties of $C^*$. For the sake of later reference we introduce the following definitions and remark.

[6] [37] refers to the corresponding 1st order difference equation.

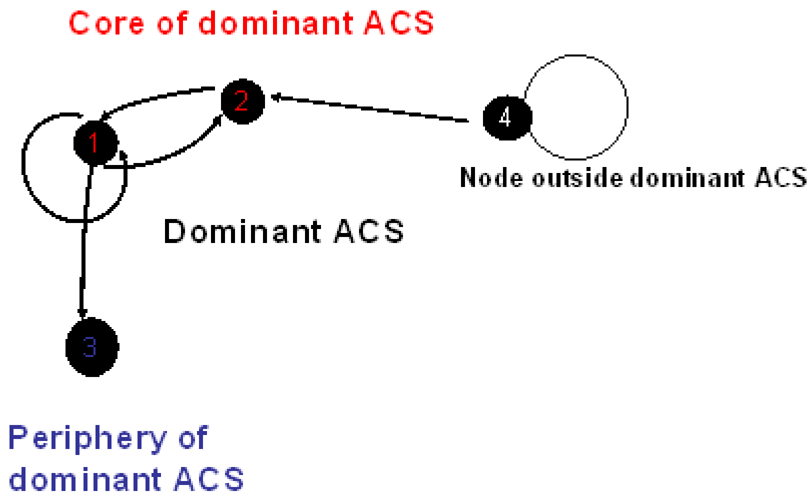
**Definition 2** *For the graph $G(S, L)$ associated to a connection matrix $C$, a autocatalytic set (ACS) is a subgraph of $G(S, L)$ such that each vertex in the subgraph has at least one incoming link from some vertex of the subgraph (Jain and Krishna [17]). Notice that our assumption $c_{ii} > 0$, $i = 1, ..., n$, implies that $G(S, L)$ has $n$ trivial ACSs. The* dominant ACS *of $G(S, L)$ is its largest subgraph with the property that the associated connection matrix $C_a$ satisfies $\lambda^*(C) = \lambda^*(C_a)$.*

**Definition 3** *For the dominant eigenvector $a^*$ of the connection matrix $C$ in the weighted directed graph $G(S, L, C)$, consider the subset $S_a \subseteq S$ of the vertices corresponding to the positive components of $a^*$, together with the subset $L_a \subseteq L$ of the links between them. $G(S_a, L_a)$ is the subgraph corresponding to $a^*$ of the unweighted directed graph $G(S, L)$ associated to $G(S, L, C)$.*

**Remark 4** *$G(S_a, L_a)$ is the dominant ACS of $G(S, L)$.*

With the tools above we can now look at the graph $G(S^*, L^*)$ induced by $C^*$. $G(S^*, L^*)$ is itself a ACS, because node 4 sends a link to itself, but provided that $c_{44}$ is sufficiently small, node 4 does not belong to the dominant ACS of $G(S^*, L^*)$, which consists of the nodes 1, 2, 3 and the links between them. The reason is that node 4 does not receive links from the others; as a result, to the extent that $c_{44}$ is strictly lower than the dominant eigenvalue of $C^*$, the fourth component in the dominant eigenvector of $C^*$ is zero.

In the dominant ACS of $G(S^*, L^*)$, and in every ACS more generally, we distinguish a 'core' and a 'periphery'. The *core* subgraph of the ACS, which in our example is formed by vertices 1 and 2, and by the links between them, has the defining property that starting from any vertex of the core, any other vertex of the autocatalytic set can be reached following a sequence of directed links. This defining property of the core is labelled *closed path connectivity.* Vertices in the autocatalytic set that do not belong to its core, belong to its *periphery.* In our example the periphery of $G(S^*, L^*)$ consists of vertex 3, together with the link from vertex 3 to itself.




The form of aggregation enabled by the dominant eigenvector propereties of $C^*$is now revealed by the fact that the relative size $(x_1/x_2)$ of the variables 1 and 2 in the core of the dominant ACS is independent of the measure of connections $c_{ij}$ outside the core. In spite of the fact that the system is not decomposable in Simon's sense, the variables $x_1, x_2$ exert an aggregate influence on the medim-run convergence of variables outside the core. Moreover, the aggregation dominating the short-run equilibria will be also, if only approximately, felt in the long run, because $\varepsilon$ is small.

The core of the dominant ACS of a knowledge pattern is the centre of the strongest self-sustaining mechanisms of knowledge creation and transmission within that pattern. In a relevant sense, the links connecting the core to the other fields in the dominant ACS disseminate building blocks ([16]) that are the aggregate outcome of the relations within the core. Mathematically, this corresponds ton the fact that the Perron-Frobenius eigenvalue of the dominant ACS is affected by any quantitative change of a connection coefficient within the core, independently of the wiring and intensity of the links from the core to the periphery. Euristically, this form of aggregation reflects the combinatorial view of knowledge creation adopted in this paper. A stronger link $c_{ij}$ signals the increased capacity of the target field $j$ to creatively recombine knowledge from the source field $i$ with building blocks directly or indirectly received from other fields. Such building blocks are of course the outcome of previous creative recombinations. The closed path connectivity of the core is crucial, in this respect, in that it signals that the mechanisms in question are self sustaining. The Perron-Frobenius eigenvalue provides the aggregate measure of self-sustainingness.

The only reason to avoid the tighter coupling of the fields in a knowledge pattern through pervasive strong links is to avoid the corresponding growth in complexity carried by the need to set the technology standards. in one field in tune with those that are simultaneously evolving in other fields. The setting in tune will be easier, if the first order-size links that give rise to the closed path connectivity within $C^*$ are relatively few in number. The point here is that the coordination between technology fields is more complicated if the relations between them are not strictly hierarchical (one -way), but contemplate a multiplicity of feed-back loops. such loops are characterisctic of the relations within the core, which are circular, with a multiplicity in the measure of closed path connectivity, which tends to grow with the number of first order links within $C$.

We conjecture that the relatively low dimension of the (first-order) dominant core facilitates the formation of well defined standards and smooth learning interfaces within the core. The incentives for a low-dimensional dominant core will be strongest during the early phase of design-standard formation within the core, because in this phase the process of knowledge creation is more turbulent ([2], [7]). The diffusion of the aggregate knowledge produced by the core to the periphery of the ACS is made less complex by the fact that the relations between the core and the periphery are hierarchical. To this extent, we expect that the ratio between the size of core and periphery is lower during the early turbulent phase of design standard formation.

# 3 Reconstructing knowledge spillovers from patent-citation data: a brief overview (missing)

# 4 The pattern of knowledge flows and innovation dynamics: 1975-1999

The data source for our exercise is the NBER Patent-Citations data file, as made available in Jaffe and Trajtenberg [18]. The main data set PAT63_99 contains all utility patents[7] granted by the U.S. Patent and Trademark Office (PTO) between January 1, 1963 and December 30, 1999. Among the variables that the PTO originally assigns to each patent, most relevant for us, in addition

[7] Utility patents constitute the overwhelming majority of patents, which include, in addition, design, reissue and plant patents. Cfr. Hall, Jaffe and Trajtenberg [14, p. 407, n. 4].

to the grant year, is the main U.S. patent class.[8] There were 417 patent classes in the classification in use in 1999. The 'original' variables assigned by the PTO to the various patents are enriched by the authors of the dataset with a number of 'constructed variables'. In particular, the 417 classes are aggregated by the authors into 36 technological subcategories and these further aggregated into 6 categories ('Chemical', 'Computers & Communications', 'Drugs & Medical', 'Electrical & Electronic', 'Mechanical', and 'Others'). The data set PAT63_99 can be profitably matched with a second data set, namely, CITE75_99, which contains all citations made to patents in PAT63_99 by patents issued between January 1, 1975 and December 30, 1999.

The first aim of our exercise is to obtain from the citations data just described, a computationally viable description of the knowledge flows between technology fields, and of the changes thereof. In the companion paper [6] the analysis was carried out resorting to a simplified description of technology fields according to their partition into 36 subcategories. This paper extends the analysis to the technological classification according to the 418 3-digit classes. To evaluate the intensity of knowledge spillovers across technology fields, we studied how far patenting in a class $xy$ in a time interval $[t, t+z]$ was followed by citations to $xy$ by patents issued in every other class in the time interval $[t+s, t+z]$. In this way, for each class $xy$, we obtained a 418-dimensional vector of citations to $xy$. The corresponding vector of spillover intensity from $xy$ to the other classes was obtained by dividing the citations vector by the number of patents issued in $xy$ in the period $[t, t+z]$. Proceeding in this way for each $xy$ in the set of 418 classes, we arrived at a matrix of spillover intensity which is the empirical analogue of the matrix $C$ in our model. To detect structural change, if any, in the pattern of knowledge spillovers in the period under study, we divided the latter into two sub-periods and obtained a corresponding analogue of matrix $C$ for each sub-period.

The actual procedure followed was complicated by two types of considerations that have to do with those characteristics of the available data set, that are most relevant to our exercise.

The first relevant characteristic is that the number of citations in a finite time interval is affected by truncation effects related to backward and forward citation lags (Hall, Jaffe and Trajtenberg [14, pp. 421-424]). This imposed a choice of the subperiods in a way that comparisons between them were least affected by the unavoidable distortions introduced by truncation effects. In particular, the parameter $s$ was held constant between the subperiods ($s = 12$) and differences in $z$ were negligible ($z = 23$ in the first subperiod, $z = 24$, in the second). The corresponding choices for $t$ were $t = 1963$ and $t = 1975$, respectively. For the sake of later reference, the intervals $[t+s, t+z] = [1975 - 1986]$ and

[8] The reason for the qualification 'main' is that each patent is assigned by the PTO to a 3-digit patent class and to a subclass, but also to any number of 'subsidiary' classes and subclasses that seem appropriate. Moreover, the system is continuously updated with new classes being added and others being reclassified or discarded. In this case, the PTO retroactively assigns patents to patent classes, according to the most recent clasification system. Cfr. Hall, Jaffe and Trajtenberg [14, p. 415.]

$[t+s, t+z] = [1987 - 1999]$ are referred to below as first window ($W1$) and second window ($W2$), respectively.

The second relevant characteristic is that there is a sharp rising trend, largely common across categories, in the mean number of citations, per patent. This trend reflects, to a large extent, an increasing propensity to cite by PTO officers, as a result of the easier access to larger data sources brought about by computerisation of the PTO during the 1980's. Although the rising citations trend may not be *entirely* a pure artifact of the changed PTO practices, in the absence of a better alternative, the construction of the connection matrix for the second window was carried out using discounted citations data. In particular, the number of citations made by patents issued in class $xy$ in the second window, was discounted by the $xy$ growth rate of citations-made per patent between the first and second window.

There is a third potentially distorting characteristic in the data set, namely, the rising trend in the yearly number of patents issued since 1983. This feature is at least partly taken care of by our procedure, since according to our estimate of the connection matrix, the number of citations made by class $xy$ patents, issued in window $[t+s, t+z]$, to class $hk$ patents issued in $[t, t+z]$, is divided by the number of $hk$ patents granted in $[t, t+z]$.

## 4.1 Modularity of the empirical connection matrix

In this and the next sub-section we focus on a number of structural and functional properties in the pattern of knowledge flows between technology fields, that are particularly worth of attention, and are also qualitatively similar across the two windows $W1$ and $W2$.For greater ease of exposition, these properties will be illustrated in the main text with data referring to the first period 1975-1986. The corresponding data referring to 1987-1999 are shown in appendix A and subsection 4.3. The latter is devoted to the analysis of structural change, if any, between $W1$ and $W2$.

Fig. 1 reports a visual representation of the connection matrix $C(W1)$ . The colours identify different orders of magnitude of the connection coefficients.

The bright colour blocks and stripes depicted in fig. 1 is partly revealing. For instance, the red main diagonal results from the fact every class tends to be more tightly connected with itself than with other classes; the same should apply to 'well chosen groups' of technology classes. The problem revealed by fig. 1 is that the ordering of rows and columns is not particularly well chosen; it simply reflects the NBER original ordering of 3-digit classes, which is strongly influenced by temporal sequence in which the classes were first introduced. As a result, these figures do not offer an adequate visual representation of the quasi-modular structureof the matrix. A far better candidate in this respect appears to be an endogenous permutation of the ordering that groups together the classes showing a similar structural relationship with the other classes. To this end, we generated for each period a 32 groups partition[9] of the 418 classes,

[9] The number of groups in this partition was fixed exogenously.

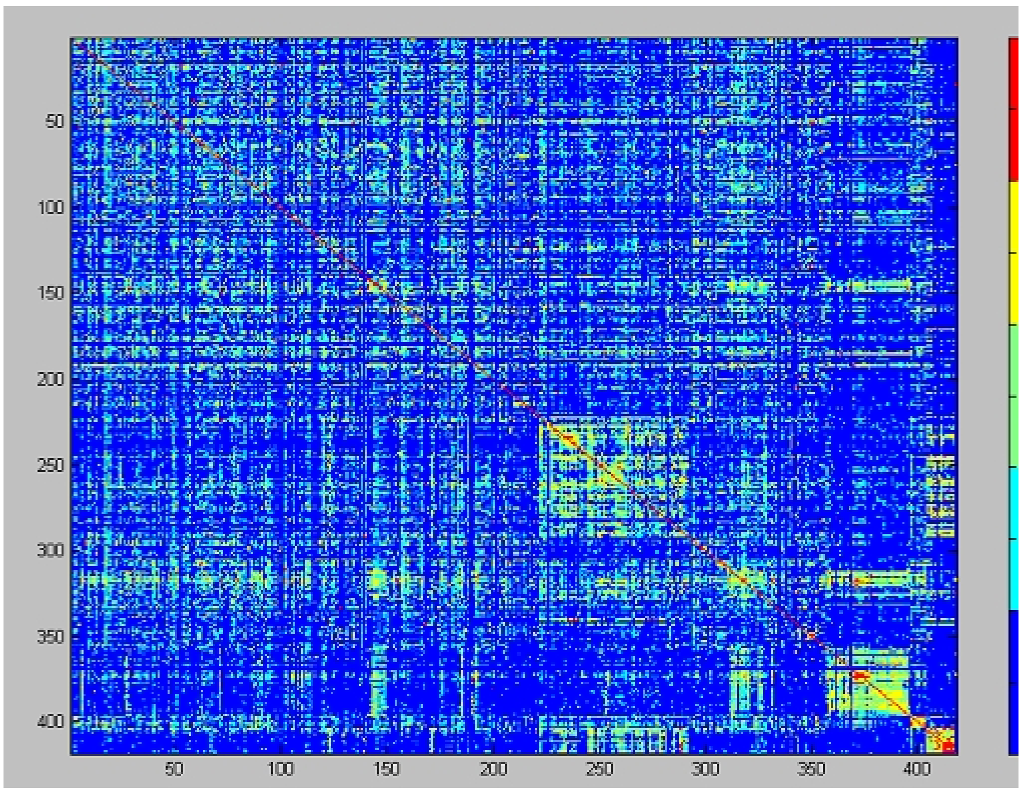



Figure 1: Average citation flows of 1 patent issued in a column technology-class, by patents issued in the row technology-class: 1975-1986. Representation based the NBER ordering of 418 3-digit classes. The colour sequence blue, light-blue, light green, yellow, red identifies progressively higher orders of magnitude of link intensity.

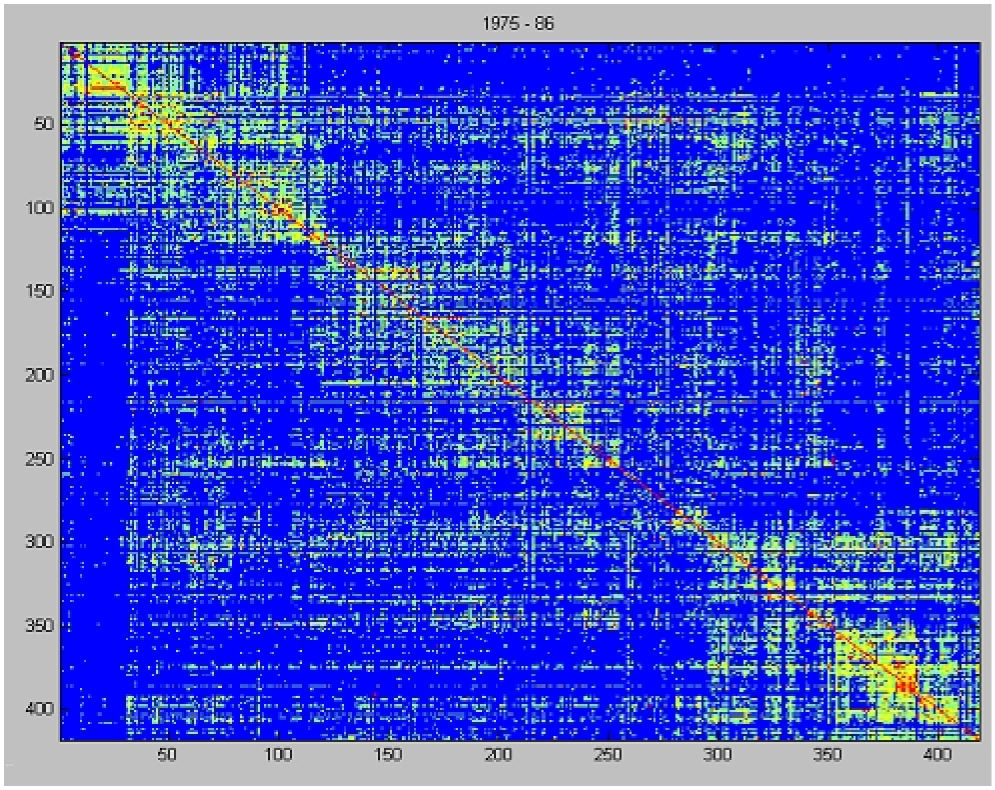



Figure 2: Average citation flows of 1 patent issued in a column technology class, by patents issued in the row technology-class: 1975-1986. Representation based the permutation of $C(W1)$ generated by the algorithm CONCOR. Colours identify link intensity as in Fig. 1 .

and a corresponding permutation of $C$, using the algorithm CONCOR, which partitions iteratively the set of technology classes into two groups, such that the inward and outword connections of two classes in the same (opposite) groupare positively (negatively) correlated[10]. The effects on the visual representation of connection strengths and their quasi-modular organization is quite sharp.

On the ground that the mathematical underpinnings of the algorithm CON-

[10] CONCOR proceeds by first building a sequence of correlation matrices $\{\varrho(1), \varrho(2), ..., \varrho(z)\}$ based on the connection matrix $C$ of a given network. $\varrho_{ij}(s)$ is the Pearson correlation coefficient between the *profile vectors* of nodes $i$ and $j$ at step $s$. The profile vector of node $i$ at step $s$ is obtained by concatenating the $i$th row and column of $\varrho(s-1)$. $\varrho(0) = C$. The iteration proceeds until convergence of the sequence to a matrix $\varrho(z)$ such that $\varrho_{ij}(z) = +1$ or $\varrho_{ij}(z) = -1$. This matrix is then used to split the nodes into two groups, such that members of the same group are positively correlated. Members of different groups are negatively correlated. CONCOR uses the above procedure to first split the set of nodes into two groups. Successive splits are then applied to the separate groups. ([4])

COR are questionable ([3], pp.270-271), we performed an alternative partition of the network, following the approach recently suggested by Newman and Girivan [25]. These authors propose that the appropriateness of any two community-structure partitions of a given network are evaluated using their proposed measure of modularity $Q$ that was described in section 2 above. In this spirit, they propose that the best community structure identification in a given network is the $Q$ maximizing partition.

Figure 3 offers a visual representation of the effect induced on $C(W1)$ by class re-ordering corresponding to such $Q$ maximizing partition. To generate this network-community structure, we followed the algorithm suggested in a still more recent contribution by M. Newman [24], which was adapted to our purpose of dealing with weighted directed networks. The MATLAB routine carrying out the computation is in itself an output of the present research[11]. The routine devides the knowledge network of period $W1$ into $n = 31$ technological communities, where $n$ was endogenously determined. According to the $Q$ criterion adopted, the resulting Newman community strucure identification ($Q = 0.6764$) performs better than the CONCOR partition ($Q = 670135$), which is in turn better than the NBER technological partition in 36 subcategories ($Q = 647874$)[12].

Figures 4 and 5 specify the class composition of the CONCOR and main Newman groups for the window $W1$. The colours emphasise the correspondence between the endogenously generated groups and the NBER technological categories. For ease of later reference, and for reasons that will be clarified in the sequel, group 28 in figure 4 and group 31 in figure 5 are the *Core-Group* (CG) within their respective lists. By far and large, the CG is composed of classes in the information and communication technologies (ITC); most, if not all, classes belong in the *Computer and communications* technological category, with some significant participation to the CG of the Newman partition by classes in *Electrical and electronics*. Most notably, in period $W1$ the CONCOR CG is a strict subset of the Newman CG, and indeed the size of the former (13 classes) is much smaller than the size of the latter.(44 classes). As we shall see in the sequel, both CGs provide meaningful approximations to well defined functional components within the network.

The group-contributions to modularity, weighted and unweighted by the number of group members, is reported in fig. 6. What is most relevant there is that in both partitions the maximum per-class contribution to modularity comes from the blue coloured 'Core Group'. It is therefore far from surprising that in the list of the contributions to modularity induced by the exogenous partition of classes into 36 NBER subcategories (shown in Appendix A, fig. 17), the blue coloured *Computer and communications* subcategories rank very high[13].

[11] It was obtained with the programming assistance of Emiliano Sparacino, now PhD student at the Department of Information Engineering of Siena University.

[12] It may be worth observing that this partition gives rise to a class ordering which is not the NBER 'historical' ordering embedded in Fig. 1.

[13] The correspondence between colour and technological category is reported in previous figures.

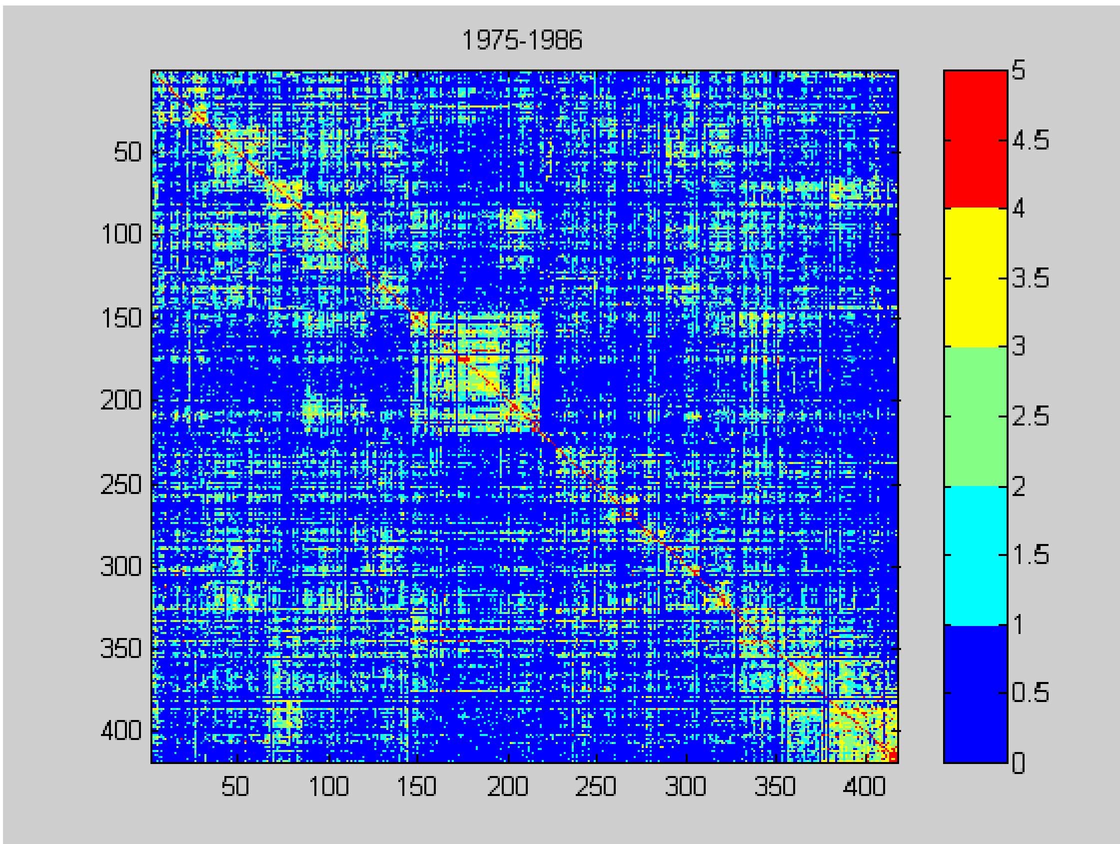



Figure 3: Average citation flows of 1 patent issued in a column technology class, by patents issued in the row technology-class: 1975-1986. Representation based the permutation of $C(W1)$ generated by our adaptation of the algorithm proposed in [24]. Colours identify link intensity as in Fig. 1 .

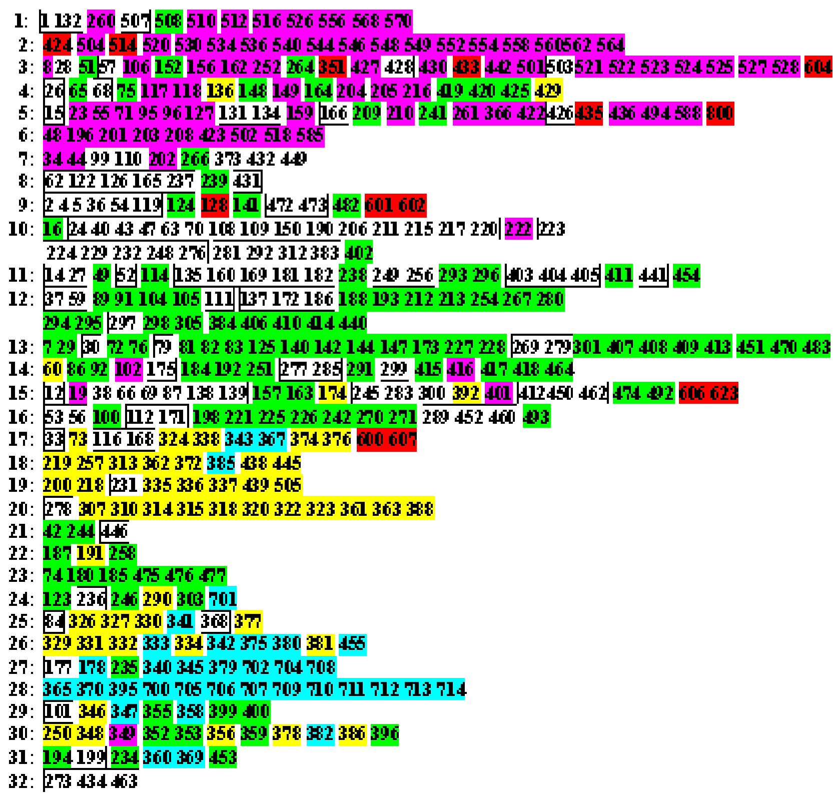



Figure 4: Partition of the set of 3-digits classes into 32 structurally 'similar' groups performed by the algorithm CONCOR on the connection matrix $C(W1)$. Blue = Computer and communic., Green = Mechanical, Pink = Chemical, Red = Drugs and medical, Yellow = Electrical and electronics, White = Others.

**1:** 73,116,338
**2:** 381
**4:** 2,4,5,7,27,54,128,132,141,215,222,227,231,383,401,417,424,433,450,482,
600,601,602,604,606,607,623
**5:** 30,56,59,70,74,91,92,173,180,184,185,188,192,244,254,267,277,280,292,
296,298,303,305,384,403,416,440,464,474,475,476,477
**8:** 200,218,307,310,315,318,320,322,323,330,331,334,335,336,361,363,388
**9:** 48,55,60,62,95,96,99,110,122,123,126,137,138,160,165,169,181,219,236,
237,239,251,261,278,285,290,291,337,376,392,406,415,418,431,432,454
**10:** 177
**11:** 24,63,86,108,112,124,150,190,211,224,248,273,297,312,343,410,446,472,473
**15:** 23,75,148,204,205,266,419,420,423,429,588
**16:** 1,8,106,149,162,260,442,503,504,507,508,510,512,514,516,520,521,523,524,525,
526,527,528,534,536,540,544,546,548,549,552,554,556,558,560,562,564,568,570
**17:** 34,44,71,127,134,159,196,201,202,203,208,210,366,422,426,435,436,494,502,518,530,585
**23:** 12,19,26,28,38,53,66,68,69,100,101,140,164,191,223,226,242,249,270,271,281,289,
294,402,412,425,452,462,492,493
**24:** 15,72,82,83,125,142,147,163,279,300,407,408,409,413,451,470,483
**25:** 131,171,186,193,194,198,209,221,232,241,258,414,453,460
**26:** 16,49,52,105,109,135,182,212,213,256,293,411
**27:** 43,47,119,206,217,220,229,245,800
**28:** 14,37,51,76,81,89,102,111,114,166,172,175,299,404,405,441
**29:** 29,57,65,79,87,117,136,156,174,216,225,228,252,257,264,269,283,333,373,427,428,438,439,445,501,505,52
**30:** 40,118,250,313,314,342,346,347,349,351,352,353,355,356,359,362,369,372,374,378,385,396,399,430
**31:** 178,199,234,235,276,324,326,327,329,332,340,341,345,348,358,360,365,367,370,375,377,
379,380,382,386,395,400,455,463,700,701,702,704,705,706,707,708,709,710,711,712,713,714

Figure 5: Class composition of the 'main' 20 out of 31 Newman groups ('main' groups have 9 or more classes, or share classes with the leading ACS specified in subsection 4.2) generated by our MATLAB implementation and extension of the algorithm in [24]. Only 18 out of 418 classes are excluded from the list in this figure; they belong to the remaining 11 groups (mostly, of 1 component). Colours identify NBER technological categories as in Fig. 4.

| CONCOR | | | | | NEWMAN | | |
|---|---|---|---|---|---|---|---|
| 0,015216 | 0,08497 | 12 | | 1 | 0,0033939 | 0.1131 | 3 |
| 0,026971 | 0,10041 | 18 | | 2 | 0,0017304 | 0.1730 | 1 |
| 0,055041 | 0,14186 | 26 | | 3 | 0,0012589 | 0.1259 | 1 |
| 0,026505 | 0,10448 | 17 | | 4 | 0,047212 | 0.1749 | 27 |
| 0,039946 | 0,11638 | 23 | | 5 | 0,046039 | 0.1439 | 32 |
| 0,019552 | 0,14558 | 9 | | 6 | 0,0021318 | 0.1066 | 2 |
| 0,011816 | 0,08798 | 9 | | 7 | 0,0015527 | 0.1553 | 1 |
| 0,012821 | 0,12274 | 7 | | 8 | 0,026372 | 0.1551 | 17 |
| 0,022328 | 0,10687 | 14 | | 9 | 0,059423 | 0.1651 | 36 |
| 0,03595 | 0,08604 | 28 | | 10 | 0,0016064 | 0.1606 | 1 |
| 0,030036 | 0,09584 | 21 | | 11 | 0,025299 | 0.1332 | 19 |
| 0,033801 | 0,08389 | 27 | | 12 | 0,0026465 | 0.1323 | 2 |
| 0,035748 | 0,08872 | 27 | | 13 | 0,0046462 | 0.2323 | 2 |
| 0,024742 | 0,09753 | 17 | | 14 | 0,0013398 | 0.1340 | 1 |
| 0,025293 | 0,07369 | 23 | | 15 | 0,018589 | 0.1690 | 11 |
| 0,023255 | 0,09739 | 16 | | 16 | 0,063889 | 0.1638 | 39 |
| 0,021061 | 0,11761 | 12 | | 17 | 0,042016 | 0.1910 | 22 |
| 0,016614 | 0,13917 | 8 | | 18 | 0,0015718 | 0.1572 | 1 |
| 0,008697 | 0,07285 | 8 | | 19 | 0,0013097 | 0.1310 | 1 |
| 0,018281 | 0,10208 | 12 | | 20 | 0,006442 | 0.1288 | 5 |
| 0,003402 | 0,07598 | 3 | | 21 | 0,0011466 | 0.1147 | 1 |
| 0,002382 | 0,05321 | 3 | | 22 | 0,0010225 | 0.1022 | 1 |
| 0,009408 | 0,10507 | 6 | | 23 | 0,034995 | 0.1166 | 30 |
| 0,011247 | 0,12561 | 6 | | 24 | 0,019751 | 0.1162 | 17 |
| 0,012458 | 0,11926 | 7 | | 25 | 0,015868 | 0.1133 | 14 |
| 0,017534 | 0,11750 | 10 | | 26 | 0,015358 | 0.1280 | 12 |
| 0,016086 | 0,11977 | 9 | | 27 | 0,011999 | 0.1333 | 9 |
| 0,041107 | 0,21190 | 13 | | 28 | 0,027013 | 0.1688 | 16 |
| 0,017668 | 0,16914 | 7 | | 29 | 0,0422 | 0.1563 | 27 |
| 0,02227 | 0,13567 | 11 | | 30 | 0,043874 | 0.1828 | 24 |
| 0,007173 | 0,08011 | 6 | | 31 | 0,1047 | 0.2435 | 43 |
| 0,005727 | 0,12791 | 3 | | 32 | | | |
| **0,670135** | | 418 | TOTAL | | **0,676400** | | 418 |
| ***Q(i)*** | ***100Q(i)/n(i)*** | ***n(i)*** | | ***i*** | ***Q(i)*** | ***100Q(i)/n(i)*** | ***n(i)*** |

Figure 6: Group contributions to $Q$ modularity, weighted and unweighted by number of classes per group. Partitions generated by CONCOR and Newman algorithms on $C(W1)$. Blue entries refer to the 'Core Groups'.

The analysis above, suggests that the technology classes in the Computer and communication technological subcategories, and even more the ICT (Information and communication technology) classes belonging to the 'Core Group' are not only most active in R&D, but receive and send a much higher than average share of their citations from and to classes belonging to the same subcategory or group. Apparently this conclusion marks a sharp contrast with the finding in [14], based on the Herfindahl concentration of the *class* distributions of patent citations made (input) and received (output). Hall, Jaffe and Trajtenberg [14] find that, on average, and throughout the period 1975-1999, patents in the *Computer and communications* category, have the *lowest concentration indexes of the input and output patent citations by class*. On this account, they argue that patents in *Computer and communications* are most 'original' because they creatively exploit knowledge from a wider set of technology classes, and produce also the most 'general' knowledge, because knowledge created by them disseminates to a wider set of classes  According to [14],the highest generality score makes the label 'general purpose technologies' most appropriate for the classes belonging to the *Computer and communications* category.

The solution to the apparent paradox is that the modularity measures considered in this paper are based on the grouping together of classes into 'similar' technological communities. The fact that the ICT classes exhibit a relatively high modularity measure, based on this partition, does not contradict the further fact that they also rank lowest in the Herfindhal concentration index based on the class distribution (of inward and outward citations). It simply means that a relatively large share in the 'wider sets of technology classes' sending knowledge connections to and receiving knowledge connections from the ICT classes, belong in the same technology group or subcategory. The denomination of 'general purpose technologies'given by Hall, Jaffe and Trajtenberg [14] to the ICT classes is therefore inappropriate, if it is simply based on their finding concerning the Herfindahl index. To corroborate this conclusion, we must spel the doubt that the apparent clash between our modularity measures and the cited results of [14] may have to with the fact that the former, unlike the latter, are based on our definition of a connection matrix, in which each column distribution of the citations receaved by one class is normalized by the number of patents issued in that class in the corresponding period. To this end, we report below the Herfindahl concentration indexes concerning the distribution of the absolute number of inward and outward citations, by group and subcategory, for the period $W1$ (fig. **??**). The findings qualify our modularity result, confirming that the CONCOR Core Group *ranks relatively high in the ordering of concentration indexes*; for the wider set of ITC classes in *Computer and communications* and also in *Electrical and electronics* the situation is more mixed, but in any case they do not rank lowest in the list.

Our findings do not necessarily contradict the idea that the new knowledge embodied in ICT innovations was 'general purpose' and, as such, could be exploited in a wide set of diverse technology classes. The corroboration of this idea can not simply rest on concentration indexes of citation distributions by class. As will be shown, it requires a much more elaborate analysis of some

| H* on unweighted 32x32 matrix | | |
|---|---|---|
| block | input distr. | output distr. |
| 1 | 0,38717 | 0,35993 |
| 2 | 0,7087 | 0,64151 |
| 3 | 0,5593 | 0,59829 |
| 4 | 0,48597 | 0,46876 |
| 5 | 0,50805 | 0,53951 |
| 6 | 0,56613 | 0,55953 |
| 7 | 0,3368 | 0,37148 |
| 8 | 0,50963 | 0,52178 |
| 9 | 0,52957 | 0,58593 |
| 10 | 0,55445 | 0,51401 |
| 11 | 0,51822 | 0,51927 |
| 12 | 0,50702 | 0,46333 |
| 13 | 0,47776 | 0,50606 |
| 14 | 0,48874 | 0,45484 |
| 15 | 0,38458 | 0,35475 |
| 16 | 0,5528 | 0,52312 |
| 17 | 0,48688 | 0,50091 |
| 18 | 0,47909 | 0,51145 |
| 19 | 0,60956 | 0,53109 |
| 20 | 0,47802 | 0,44215 |
| 21 | 0,41756 | 0,44173 |
| 22 | 0,37798 | 0,46299 |
| 23 | 0,4235 | 0,3848 |
| 24 | 0,50749 | 0,62645 |
| 25 | 0,5 | 0,43053 |
| 26 | 0,49958 | 0,46926 |
| 27 | 0,33754 | 0,34883 |
| 28 | 0,47518 | 0,52802 |
| 29 | 0,49896 | 0,61149 |
| 30 | 0,54692 | 0,5122 |
| 31 | 0,52156 | 0,53522 |
| 32 | 0,50171 | 0,57316 |

| H* on 36x36 unweighted matrix | | |
|---|---|---|
| subcat. | input distr. | output distr. |
| 11 | 0,2405 | 0,28089 |
| 12 | 0,19727 | 0,20308 |
| 13 | 0,38906 | 0,38557 |
| 14 | 0,57121 | 0,35733 |
| 15 | 0,52319 | 0,48689 |
| 19 | 0,49983 | 0,51596 |
| 21 | 0,49434 | 0,49226 |
| 22 | 0,33026 | 0,45034 |
| 23 | 0,31244 | 0,36196 |
| 24 | 0,50297 | 0,45282 |
| 31 | 0,35729 | 0,48437 |
| 32 | 0,53607 | 0,63031 |
| 33 | 0,38477 | 0,42907 |
| 39 | 0,51528 | 0,55276 |
| 41 | 0,50032 | 0,3742 |
| 42 | 0,51939 | 0,47803 |
| 43 | 0,41635 | 0,39916 |
| 44 | 0,42619 | 0,37802 |
| 45 | 0,46774 | 0,44189 |
| 46 | 0,52496 | 0,52765 |
| 49 | 0,33447 | 0,35676 |
| 51 | 0,46315 | 0,44653 |
| 52 | 0,39081 | 0,44235 |
| 53 | 0,56713 | 0,56178 |
| 54 | 0,50677 | 0,55653 |
| 55 | 0,50329 | 0,5281 |
| 59 | 0,41762 | 0,37756 |
| 61 | 0,57993 | 0,61724 |
| 62 | 0,57877 | 0,66926 |
| 63 | 0,59548 | 0,603 |
| 64 | 0,56644 | 0,64103 |
| 65 | 0,46747 | 0,50109 |
| 66 | 0,48288 | 0,52523 |
| 67 | 0,3641 | 0,28828 |
| 68 | 0,50343 | 0,49064 |
| 69 | 0,36531 | 0,37321 |

Figure 7: Normalized Herfindahl concentration indexes concerning the distributions of patent citations made (input distribution) and received (output distribution) by each group in the 32 and 36 group partitions for the period 1975-1986.

structural properties of the connection matrices, focused on the notions of near-decomposability, and autocatalytic sets.

It is to this analysis that we now turn.

## 4.2 Near decomposability and the Core properties of ICT

The CONCOR and Newman partitions of the knowledge network described by $C(W1)$ share the property that the technological community, or group, exhibiting the highest $Q$ modularity measure is composed (almost) exclusively by ICT classes. It is now time to justify the claim that each of the two communities thus identified represents the 'Core group' (CG) in its partition, and to discuss the relevance of this claim.

The first step in the argument is to see to what extent each CG represents a module in the sense of Simon [36], that is, in the sense that the size of the connection links between the classes within the group are at least one order of magnitude larger than those sent to, or receaved from, the classes that do not belong in the group. To this end, we produced a dichotomized connection matrix $C_A(W1)$[14] with the defining property that all connection links of the original matrix $C(W1)$, that are larger than or equal to 0.1 are set equal to 1 and all the others are set to 0. The exercise shows that in period $W1$[15] both the relatively small CONCOR CG and the relatively large Newman CG are connected by first-order-magnitude links with classes belonging to other groups in their respective partition. Most notably, in spite of the fact that the groups in the Newman partition provide the best approximation to a modular division of the network, they do not meet the strong requirements imposed by Simon's definition of modularity.

### 4.2.1 The dominant autocatalytic set of $C_A$

Simon's idea of separating first-oder-magnitude from lower-order magnitude links brings to the fore interesting functional and structural properties of the connection matrix $C$. It turns out that the dominant ACS of the dichotomized matrix $C_A(W1)$, which is associated to the dominant eigernvalue $\lambda_1(W1_A) =$ 5.34, has a Core consisting of 8 classes, all of which are in the *Computer and communication* category, and all belonging to the CONCOR group CG($W1_{CONCOR}$)[16] and, a fortiori, to the Newman community CG($W1_{Newman}$) which draw their name from this finding (see fig. 8). They are:

[14] $C_A(W2)$ for period 1987-1999.

[15] The same holds for period $W2$.

[16] Three out of the five classes in CG($W1_{A\_CONCOR}$), which do not belong to the core of the dominant ACS($C_A(W1)$), belong to its periphery. They are:

365 Static information storage and retrieval

370 Multiplex communications

700 Data processing: generic control systems or specific applications.

Finally, the remaining two classes of CG($W1_{A\_CONCOR}$), namely, class 706 (artificial intelligence) and 395, do not belong to the dominant ACS($C_A(W1)$), but send first-order magnitude links to members of this set.

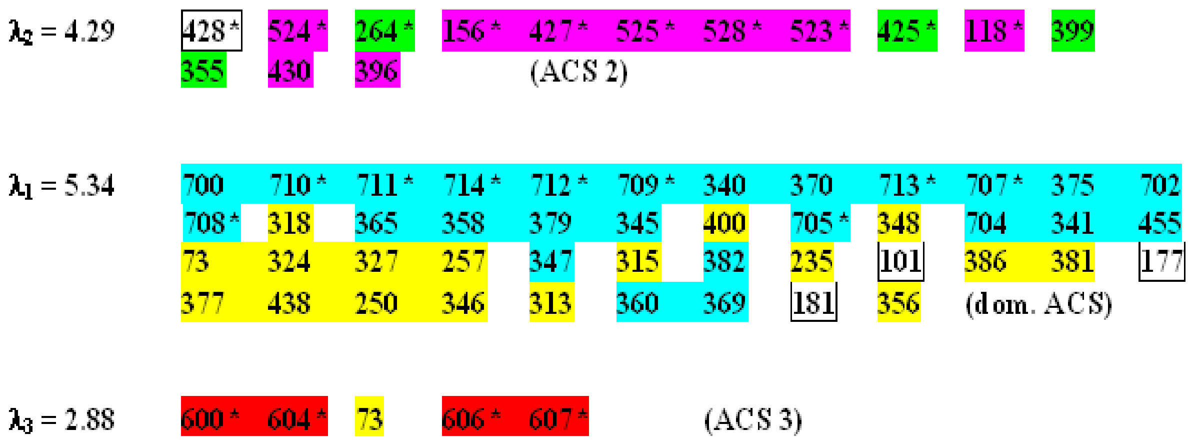



Figure 8: Class composition of the 3 autocatalytic sets of the matrix $C_A(W1)$ associated with the first 3 highest positive eigenvalues. Autocatalytic sets embedded in dom ACS have not been considered. Classes with the asterisc are in the core of the ACS.

705 Data processing: financial, business practice, management, or cost/price determination

707 Data processing: database and file management or data structures

709 Electrical computers and digital processing systems: multicomputer data transferring

710 Electrical computers and digital data processing systems: input/output

711 Electrical computers and digital processing systems: memory

712 Electrical computers and digital processing systems: processing architectures and instruction processing (e.g., processors)

713 Electrical computers and digital processing systems: support

714 Error detection/correction and fault detection/recovery

The periphery of the dominant ACS($C_A(W1)$) is almost five times larger than the core. It consists of 37 classes, 17 in the Computer & communications category, 15 in Electrical & electronics, 3 in Others and 2 in Mechanical; 22 of these classes belong to the core community of the Newman partition, which comprises 30 out of the 45 nodes that, together with their mutual (first order) links build the dominant ACS($C_A(W1)$). Of these 45 classes, only five do not obviously belong to the ITC group. Figure **??** offers a visual representation of the link architecture of this autocatalytic set. The 8 blue nodes of the core send first order magnitude links not only to nodes allined on a one-way path, but also to 8 other loops of strongly connected components that are core structures in smaller ACS embedded in the periphery of the dominant ACS($C_A(W1)$) (4 components with 2 members, 2 with 4 members, 2 with 5 members).

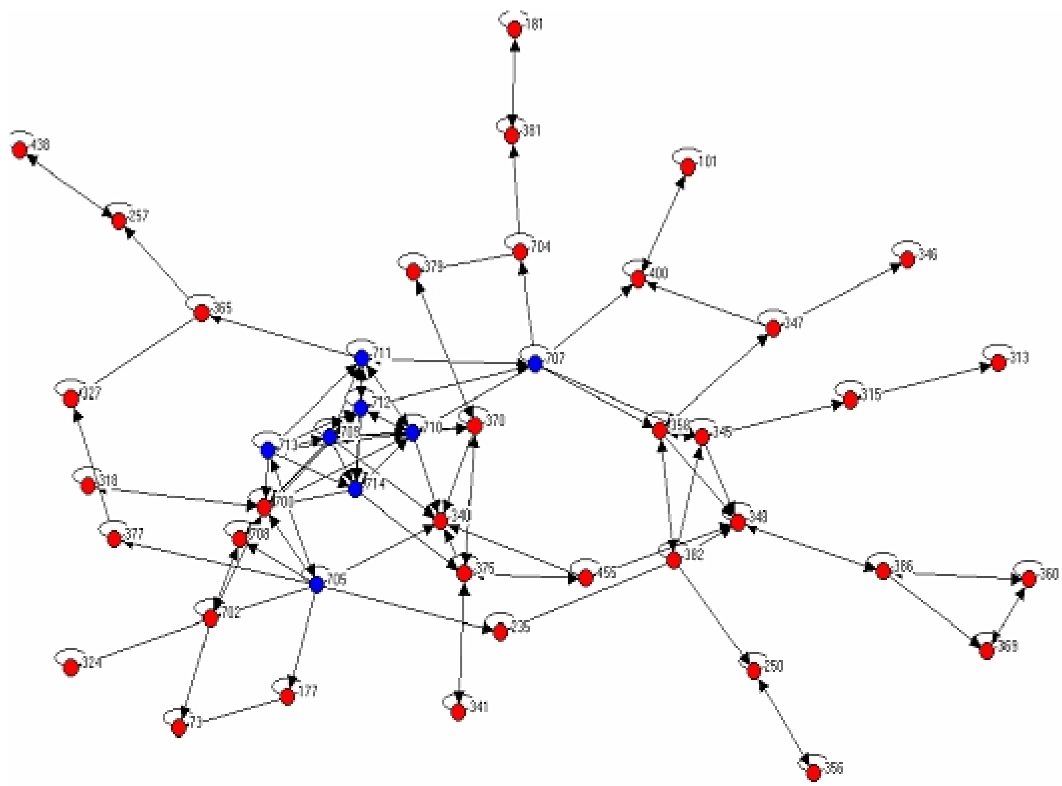


Figure 9: Dominant ACS of the dichotomized matrix $C_A(W1)$. The blue and red nodes correspond to core and periphery nodes, respectively.

### 4.2.2 Non dominant Autocatalytic sets of $C_A(W1)$

The directed network corresponding to the dichotomized matrix $C_A(W1)$ embeds two other prominent, but much smaller autocatalytic sets, beside the dominant one. Their class composition in the first period is shown in figure 8, where they are labelled $ACS_2$ and $ACS_3$, respectively. They are associated to positive eigenvalues of the matrix $C_A(W1)$, namely $\lambda_2 = 4.29$ and $\lambda_3 = 2.88$. The functional significance of the ACS concept in the context of patent analysis is confirmed by the fact that, as was the case for the dominant ACS, both $ACS_2(W1_A)$ and $ACS_3(W1_A)$ correspond to functional groups of classes that lend themselves to a clear interpretation.

In particular, $ACS_2(W1_A)$ comprises 14 classes that bear a close relation with the category *Chemicals*. With the possible sole exception of class 418 (Stock material or miscellaneous articles), the 10 classes in the core of $ACS_2(W1_A)$ either belong in this category, or identify processes that use plastics as their material support.

$ACS_3(W1_A)$ is considerably smaller and refers to more specialized classes of activity related to surgery, in the category *Drugs & medicals*. It may be worth observing that $ACS_3(W1_A)$ and the dominant $ACS_1(W1_A)$ have one node of their respective peripheries in common, namely the node corresponding to class 73 (Measuring and testing) in the category *Electrical & electronics*. Thus, not even the concept of ACS identifies functional modules that fully correspond to Simon's definition of a module.

As it turns out, the approximation based on first-order-magnitude links, standardized to weight size 1, detects structural properties of the empirical connection matrix $C(W)$ that do not lose their significance after all links of every magnitude have been re-introduced and all links bear their proper weight. From the weighted directed graph $G(S, L, C)$, corresponding to the empirical connection matrix $C$, we extract the subgraph $G(S_i, L_i, C_i)$ where $S_i$ is the set of nodes (classes) in the $ACS_i$ of the dichotomized matrix $C_A$ defined above (the dominant ACS is $ACS_1$), $L_i \subset L$ is the subset of links connecting the nodes in $S_i$, and $C_i$ is the connection matrix specifying the intensity of the links in $L_i$, as reported in $C$. Our model suggests that the relative degree of participation of a subset $S_i$ of nodes (classes) to the long-term self sustaining mechanisms of knowledge creation and transmission within $C$ can be evaluated by comparing the Perron-Frobenius eigenvalue of the connection matrix $C_i$ of the subgraph $G(S_i, L_i, C_i)$ with the dominant eigenvalue of $C$. The results of this analysis, carried out for the periods $W1$ and $W2$ are as follows.

| | $\lambda^*(C)$ | $\lambda^*(C_1)$ | $\lambda^*(C_2)$ | $\lambda^*(C_3)$ |
|---|---|---|---|---|
| $W1$ | 3.5825 | 3.5603 | 2.5081 | 2.8827 |
| $W2$ | 6.6492 | 6.6442 | 3.3182 | 3.4045 |

## 4.3 Structural change in the pattern of knowledge flows

Most of the properties discussed in subsection 4.2 are qualitatively unchanged between period $W1$ and $W2$.

The group contribution to modularity (weighted and unweighted by the number of group members, appendix A, fig. 16) continues to be highest for the ITC core group (CG) in the CONCOR, as well as in the Newman partition[17]. The above modularity result is confirmed by the finding that, again, in the $W2$ distribution by technology 'group' of inward and outward patent citations the core group CG($CW2_{CONCOR}$) ranks very high in the rank order list of Herfindahl concentration indexes (appendix A, fig. 18). Thus,the same strong qualifications, discussed in subsection 4.1, that are required by the identification of the ITC technologies as 'general purpose' apply also to period 1987-1999.

There is finally a high degree of continuity between the two periods in the nature and composition of the 3 leading ACS of the dichotomized matrix $C_A$. As before, if the dominant $ACS_1(W2_A)$ corresponds to the ITC group, the other two, $ACS_2(W2_A)$ and $ACS_3(W2_A)$ are characterized by core components, that are firmly embedded in *Chemicals* and in *Measuring and testing*, respectively (fig. 12).

The main features of structural change in the pattern of knowledge flows in period $W2$, as compared to period $W1$ are:

(i) The global increase in the connectivity of the knowledge network, signalled by the strong increase in the eigenvalue measure of connectivity shown at the end of subsection 4.2, indicates a higher monitored activity[18] of knowledge transfer.

(ii) The lower degree of modularity of the network is signalled by the fact that, for each type of partition considered, NBER subcategory, CONCOR or Newman, the $Q$ measure of modularity is lower in the second period[19].

(iii) The composition between core and perifery in the dominant ACS of the dichotomized matrix $C_A$ changes dramatically between $W1$ and $W2$.

(iv) Further changes occur in the other leading ACS of the matrix $C_A$.

Below we expand on points (iii) and (iv).The number of classes in the dominant ACS($C_A(W2)$) is not much larger than the corresponding number in the dominant ACS($C_A(W1)$): 52 in the former against 45 in the latter. It is the relative composition of the dominant ACS between core and periphery to change dramatically from period $W1$ to $W2$. The ratio between the size of core and perifery is reversed: $\frac{\#Core}{\#Periphery}$ is 0.216 in $W1$ and 4.2 in $W2$.In 1975-99 the 8

[17] In the former, CG($CW2_{CONCOR}$) is now group 1, in the latter, CG($CW2_{Newman}$) is group 44. The class composition of the CONCOR and main Newman groups are shown in appendix A, figures 14 and 15.

[18] As explained before, the number of patent citations was discounted in the period $W2$ in the attempt of taking into account the changed USPTO practices. It is hard to know to what extent the higher activity of knowledge transfer is over-estimated or under-estimated by the discounted data.

[19] The Newman Girivan $Q$ measures of modularity for period $W1$ are:
$Q_{NBER} = 0.6478$, $Q_{CONCOR} = 0.6701$, $Q_{Newman} = 0.6764$;
for period $W2$ they are:
$Q_{NBER} = 0.6107$, $Q_{CONCOR} = 0.6145$, $Q_{Newman} = 0.6278$.

**Autocatalytic set of C(W1.A)**

***Core : 8 members***

705 707 709 710 711 712 713 714

***Periphery: 37 members***

73 101 177 181 235 250 257 313 315 318 324 327 340 341 345
346 347 348 356 358 360 365 369 370 375 377 379 380 381 382
386 400 438 455 700 702 704

**Autocatalytic set of C(W2.A)**

***Core : 42 members***

180 192 235 303 318 324 326 327 340 341 345 347 348 358 360
365 369 370 375 379 380 382 386 395 399 400 455 475 477 700
701 702 705 706 707 708 709 710 711 712 713 714

***Periphery: 10 members***

29 73 74 123 257 280 310 361 438 439

Figure 10: Core and Periphery of the Autocatalytic sets of the connection matrices $C(W1)$ and $C(W2)$. Blue = Computer and Communications, Yellow = Electrical and electronics, Green = Mechanical.

core classes all belong to the NBER category *Computer & communications*, and more specifically to *Electrical computers and digital processing systems*, *Data processing*, and *Error detection/correction*. In 1987-99 the core of the dominant ACS($C_A(W2)$) contains 42 member classes, of which 28 belong to *Computer & communications*, 8 to *Mechanical* and 6 to *Electrical & electronics*. The sharp absolute and relative increase in the number of core members in the second period is illustrated in fig. 10.

The increase in core size and its more differentiated composition by technological category signals a higher degree of integration of the ITC dominant paradigm with the rest of the economy. A much larger number of classes belonging to more heterogeneous technologies, is participating beside the core ITC classes in the first-order size self-sustaining mechanisms of knowledge creation and transmission in period $W2$ as compared to $W1$. Figure **??** shows the changed structure of the dominant ACS ($C_A(W2)$). Most loops of strong com-

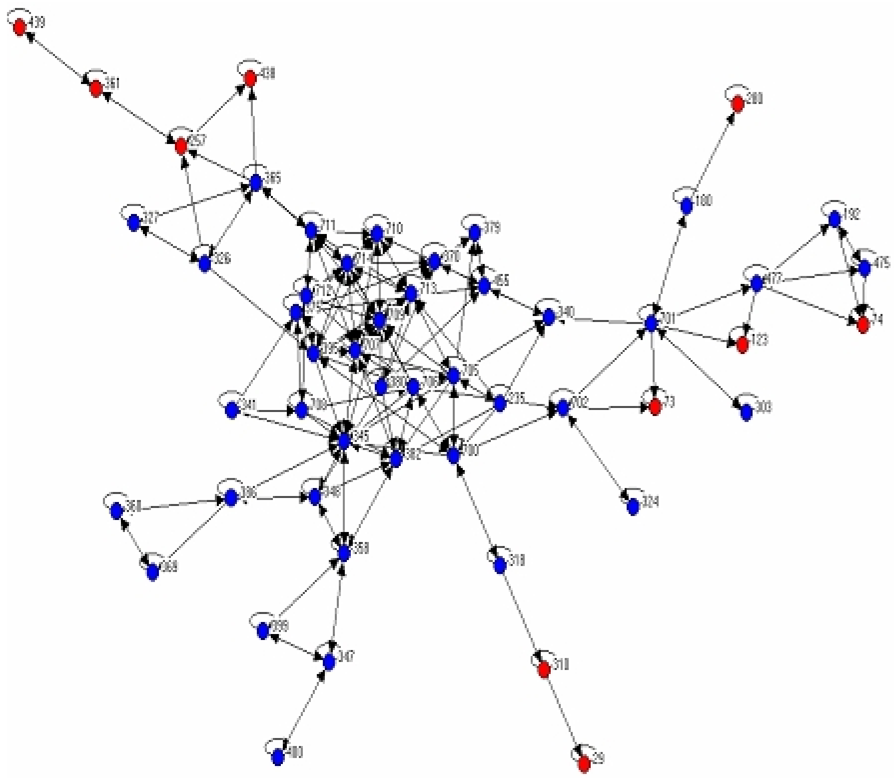


Figure 11: Dominant ACS of the dichotomized matrix $C_A(W2)$. The blue and red nodes correspond to core and periphery nodes, respectively.

ponents previously embedded in the periphery of the dominant ACS have now been included in the new expanded core. There are only 10 nodes in the periphery, all belonging to the categories *Mechanical* and *Electrical & electronics*; of them, 4 form a strongly connected component.

The expansion of first order magnitude links in the second period is not circumscribed to the dominant ACS, but is a more general trend. $ACS_2$ expands from 14 to 20 members, with some change also in its composition.by class (4 classes exit and 10 new classes enter, mostly belonging in the categories *Chemicals* and *Electrical & electronics*). The outcome is also a change in the composition of the core of $ACS_2(W2_A)$, which in the second period consists of 9 classes, all of which in *Chemicals*, and only 4 of them already part of the core in the first period.

The expansion of $ACS_3$ from the first period to the second is much sharper than for the other autocatalytic sets. The number of members increases from 5 to 24 nodes, with new classes entering the core and the periphery, while no exits

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\lambda_1 = 7.7526$ | 709* | 714* | 707* | 395* | 710* | 711* | 345* | 713* | 712* | 370* | 706* | |
| | 375* | 382* | 700* | 365* | 358* | 455* | 705* | 340* | 380* | 379* | 708* | |
| | 701* | 702* | 257 | 438 | 326* | 327* | 318* | 347* | 399* | 386* | 235* | 73 |
| | 341* | 439 | 324* | 361 | 310 | 360* | 369* | 400* | 123 | 477* | 180* | |
| | 29 | 303* | 74 | 192* | 475* | 280 | 348* | (dom. ACS) | | | | |
| $\lambda_3 = 4.0205$ | 600* | 606* | 604* | 428 | 424 | 128* | 607* | 623* | 264 | 156 | 427 | 435 |
| | 514 | 438 | 425 | 436 | 118 | 530 | 17 | 257 | 422 | 361 | 210 | 439 |
| | (ACS 3) | | | | | | | | | | | |
| $\lambda_2 = 4.0564$ | 428 | 264 | 156 | 427 | 524* | 525* | 528* | 523* | 438 | 425 | 526* | 118 |
| | 502* | 257 | 208* | 585* | 361 | 423* | 439 | 18 | | (ACS 2) | | |

Figure 12: Class composition of the 3 autocatalytic sets of the matrix $C_A(W2)$ associated with the first 3 highest positive eigenvalues. Autocatalytic sets embedded in dom ACS have not been considered. Classes with the asterisc are in the core of the ACS.

occur. The expanded core now contains 6 classes, again highly specialized and all related to the activity of surgery or fabrication of prosthesis. In spite of the highly specialized core, all technological categories except *Computer & communications* are represented in this autocatalytic set in period $W2$. The member nodes include classes in the semiconductor technology, in molecular biology, in chemistery and in mechanical processes using plastics as material support. It is also remarkable that the $ACS_2(C_A(W2))$ shares with the $ACS_3(C_A(W2))$ 10 members of its perifery (belonging in the categories *Chemicals*, *Electr. & electronics*, *Mechanicals* and *'Others'*) and that both the $ACS_2(C_A(W2))$ and the $ACS_3(C_A(W2))$ share 4 nodes with the dominant $ACS(C_A(W2))$, all of which in *Electrical & electronics.* This finding strongly reinforces the conclusion that autocatalytic sets are functional structures that do not correspond to modules in Simon's sense.

# 5 Conclusions: ICT fields, general purpose technology, modularity, and ACS

Since our narrative is focused on the role played by the ICT fields during the period of analysis, it is worth summarising our findings in this perspective, and bring them within a coherent framework.

A first order of considerations refers to structural features of the knowledge transfer between technology fields that are common to the periods under consideration, 1975-1986 and 1987-1999.

If the organization of the learning interfaces connecting technology fields is regarded through the lenses of the Newman Girivan [25] $Q$ measure of modularity, and if connection links of every size are considered, the organization concerning the ICT fields reveals, compared to that of the other technology groups, a relatively high degree of modularity. The statement must be interpreted in the sense that this technological community shows a maximum, or at least relatively high, propensity to be more tightly connected with itself than with other groups. This finding needs to be reconciled with the statement [14] and with the common wisdom that the ITC represented general purpose technologies in the period under consideration.

In our interpretation, the label of 'general purpose' for the ICT technologies in the whole period 75-99 is based on the finding that in both $W1$ and $W2$ a subset of ICT fields is a crucial part of a core sub-network connected by a circular path of strong links that together build a self-sustaining mechanism of knowledge transfer. What is special about the core structure pertaining to the critical ICT fields is: (i) It obtains, among all such structures, characterized by closed-path connectivity, the highest rate of knowledge transfer, as evaluated through a dominant eigenvalue measure. (ii) It reaches out, through first order links, the by far largest set of technology classes, which build up the dominant ACS($C_A$) of the strong-links network. We emphasize that the closed path connectivity of the core implies that the systematic and persistent flow of knowledge transfer from the core of the dominant ACS($C_A$) to the periphery and to the network at large, through second order links, contains a form of knowledge aggregation.

When we ask what technologies may be or may not be regarded as general purpose in a given historical period, we are posing questions concerning the absolute pervasiveness of such technologies. The judgement cannot abstract from the absolute overall intensity of the links connecting the technology in question to the others. An abstraction of this type is precisely what is implied by measures (in terms of $Q$ modularity) of how community oriented is the organisation of a connection structure,independently of its overall intensity; or by measures (in terms of Herfindahl indexes) of how concentrated is a distribution of connections.

A second order of considerations refers to the euristic interpretation of the directions of structural change between the first and the second period. To some extent, information flows were more intense, diffused and less community oriented in the second period, as is suggested by a mild decline of the $Q$ modularity measures from $W1$ to $W2$. In this respect the ICT group, no matter if identified by the NBER classification, by 'structural similarity', or by Newman community structure, simply follows the general trend.

Additional features of structural change were obtained by focusing on first-order magnitude links. In particular, the absolute and *relative* increase in the size of the core, compared to periphery, and its more differentiated composition by technological category, signals a higher degree of integration of the ITC dominant paradigm with the rest of the economy. This finding is prima facie consistent with a evolutionary interpretation of design standard formation ([2], [7]). During the early phase of design standard formation technological change is more turbulent and there is a marked trade off between the knowledge gains that can be obtained through tighter links with R&D in other fields, on the one hand, and the constraints imposed by conformity with the standars prevailing in such fields, on the other. An excess of connectedness makes finding a 'fit' solution or design on a technological landscape more difficult, because landscape are constantly deforming as a result of the fact that convergence to a stable technology standard is not yet complete and the solutions provisionally identified in different fields may not be compatible. The nature of the trade off is drastically altered in favour of connectedness, after a set of mutually compatible standards has emerged. Now the relation between connectedness and landscape deformation is much waker. Correspondingly, there are higher incentives for a relatively large set of strong mutual interactions, as we observe in the core of the dominant ACS($C_A$).

# References


[1] Aghion P., Howitt P.(2005):"Growth with Quality-Improving Innovations: An Integrated Framework", in Durlauf S. N., Howitt P., *Handbook of Economic Growth*, Amsterdam, Elsevier, pp. 67-110.

[2] Anderson P., Thusman M. (1990): "Technological Discontinuities and Dominant Designs: A Cyclical Model of technological Change", *Administartive Science Quarterly*, vol. 35, pp. 604-633.

[3] Brandes U., Erlebach T. (eds.) (2005): *Network Analysis*, Berlin, Springer Verlag.

[4] Breiger R., Boorman S., Arabie P. (1975): "An Algorithm for Clustering Relational Data, with Applications to Social Network Analysis and Comparison with Multi-Dimensional Scaling", Journal of Mathematical Psychology, 12, 328-383.

[5] Caminati M. (2006): "Knowledge Growth, Complexity, and Returns to R&D", in *Journal of Evolutionary Economics,* vol. 16, n. 3, pp. 207-229.

[6] Caminati M., Sordi S., Stabile A. (2006): "Patterns of Discovery", *Quaderni del Dipartimento di Economia Politica*, n. , Università di Siena.

[7] Chesbrough H. (2003): "Towards a Dynamics of Modularity. A Cyclical Model of Technical Advance", in [30], pp. 174-198.

[8] Cowan R. W. (2004): "Network Models of Innovation and Knowledge Diffusion", *MERIT Research Memorandum* RM2004-016, www.merit.unimaas.nl.

[9] Crutchfield P. J. (2001): "When Evolution is Revolution- Origins of Innovation", in [10].

[10] Crutchfield P. J. and Schuster P. (eds.) (2001): *Evolutionary Dynamics-Exploring the Interplay of Selection, Neutrality, Accident and Function*, Santa Fe Institute Series in the Sciences of Complexity, New York, Oxford University Press.

[11] Danon L, Diaz-Guilera A., Duch J. and Arenas A. (2005): "Comparing Community Structure Identification", *Journal of Statistical Mechanics*, P09008.

[12] David P. (1990): "The Dynamo and the Computer: An Historical Perspective on the modern Productivity Paradox", *American Economic Review, Papers and Proceedings*, 80, 355-361.

[13] Devine W. (1983): "From Shafts to Wires: Historical Perspective on Electrification", *The Journal of Economic History*, XLIII, 347-372.

[14] Hall B. H., Jaffe A. B. and Trajtenberg M. (2002): "The NBER Patent-Citations Data File: Lessons, Insights and Methodological Tools", in Jaffe and Trajtenberg [18].

[15] Hirsch M. W. and Smale S. (1974): *Differential Equations, Dynamical Systems, and Linear Algebra*, New York, Academic Press.

[16] Holland J. H. (1998): *Emergence: From Chaos to Order*, Reading, Addison Wesley.

[17] Jain S. and Krishna S. (2003): "Graph Theory and the Evolution of Autocatalytic Networks", in Bornholdt S. and Schuster H. G. (eds.), *Handbook of Graphs and Networks*, Weinheim, GE, WILEY-VCH, pp. 355-395.

[18] Jaffe A. B. and Trajtenberg M. (eds.) (2002): *Patents, Citations & Innovations, a Window on the Knowledge Economy*, Cambridge, MA, The MIT Press.

[19] Kauffman S. (1993): *The Origins of Order*, New York, Oxford University Press.

[20] Litterer J. A. (ed.)(1969): *Organizations: Systems, Control and Adaptation*, 2, New York, J. Wiley.

[21] Muff S., Rao F. and Caflisch A. (2005): "Local Modularity Measure of Network Clusterization", *Physical Review E*, 72, 056107.

[22] Mokyr J. (1990): *The Lever of Riches. Technological Creativity and Economic Progress*, New York, Oxford University Press.

[23] Newman M. E. J. (2004): " Analysis of Weighted Networks", *Physical Review E*, 70, 056131.

[24] Newman M. E. J. (2006): "Modularity and Community Structure in Networks", *Proceedings of the National Academy of Sciences*, 103, pp. 8577-8582.

[25] Newman M. E. J. and Girivan M. (2004): "Finding and Evaluating Community Structure in Networks", *Physical Review E*, 69, 026113.

[26] Olson G. M., Malone T. W., and Smith J. B. (eds.) (2002): *Coordination Theory and Collaboration Technology*, Mahwah, NJ, LEA Publishers

[27] Page E. S. (1998): "Uncertainty, Difficulty, and Complexity", Mimeo.

[28] Patee H. (ed.): *Hierarchy Theory*, New York, Braziller

[29] Pavitt K. (1998): "Technologies, Products and Organization in the Innovating Firm: What Adam Smith Tells Us and Schumpeter Doesn't", *Industrial and Corporate Change*, 7, 433-452.

[30] Prencipe A, Davies A. and Hobday M. (2003): *The Business of System Integration*, New York, Oxford University Press.

[31] Reiter S. (1992): Knowledge, Discovery and Growth. Discussion Paper #1011, Northwestern University. Revised version in [26]

[32] Rosenberg N. (1998): "Chemical Engineering as a General Purpose Technology" in Helpman E. (ed.): *General Purpose Technologies and Economic Growth*, Cambridge, MA, MIT Press, 167-192.

[33] Samuelson L. (2005): "Economic Theory and Experimental Economics", *Journal of Economic Literature*, XLIII, 65-107.

[34] Schmookler J. (1966): *Invention and Economic Growth*, Cambridge, MA, Harvard University Press.

[35] Simon H. A. (1962): "The Architecture of Complexity", *Proceedings of the American Philosophical Society*, 106, 467-482; reprint in [20], pp. 98-114.

[36] Simon H. A. (1973): "The Organization of Complex Systems", in [28], 1-27."

[37] Simon H. A. and Ando A. (1961): "Aggregation of Variables in Dynamic Systems", Econometrica, 29, pp. 111-138.

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

[39] Weitzman M. L. (1998): "Recombinant Growth", *Quarterly Journal of Economics*, 113, 331-360.

[40] Watson R. A. (2006): *Compositional Evolution. The Impact of Sex, Symbiosis, and Modularity on the Gradualist Framework of Evolution*, Cambridge, MA, MIT Press.

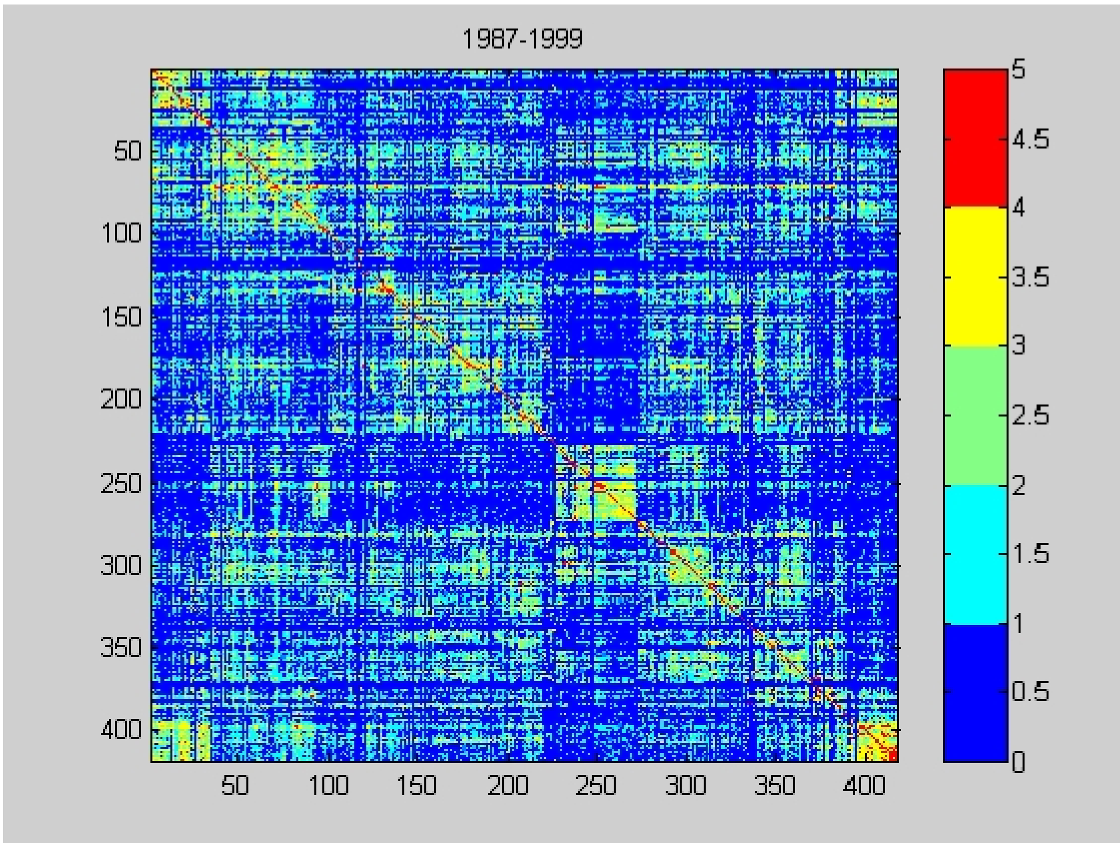



Figure 13: Average citation flows of 1 patent issued in a column technology class, by patents issued in the row technology-class: 1987-1999. Representation based the permutation of $C(W2)$ generated by our adaptation of the algorithm proposed in [24]. Colours identify link intensity as in Fig. 1 .

1: 1 340 345 365 370 379 380 395 700 702 704 705 706 707 708 709 710 711 712 713 714

2: 84 178 186 194 235 273 341 348 358 360 368 382 386 434 463

3: 323 326 327 330 331 333 334 377

4: 329 332 342 375 381 455

5: 33 73 147 250 324 351 356 367 374 376 378 600 607

6: 136 257 313 315 359 361 362 372 385 438 445

7: 270 271 346 347 355 399 400

8: 349 352 353 369 396 453

9: 37 56 104 105 172 212 213 278 298 414 460

10: 42 114 124 188 199 242 244 248 258 276 294 343 446 482

11: 60 91 102 280 440

12: 74 123 180 192 291 303 475 476 477 701

13: 70 116 185 191 292

14: 200 218 231 335 336 337 338 439

15: 177 187 236 246

16: 290 307 310 318 320 322 363 388

17: 2 12 26 28 36 38 54 57 66 87 128 156 168 245 425 450 602

18: 40 47 53 141 150 190 206 215 217 220 221 222 223 229 281 283 383 401 402 412 452 462 493

19: 4 5 16 24 43 89 108 119 211 224 232 267 296 297 312 410 441 473 601

20: 14 27 49 52 109 135 160 169 182 238 249 256 293 403 404 405 449 454 472

21: 7 30 51 59 63 69 76 79 81 82 86 125 140 142 144 157 163 164 173 174 193
225 227 228 234 254 269 279 289 295 300 301 305 407 408 409 411 413 429 470 483 492

22: 19 29 72 83 100 101 112 198 216 219 226 314 373 392 419 451 606

23: 15 111 139 152 171 175 299 420 474

24: 92 137 138 165 181 184 239 251 277 285 384 406 415 416 417 418 464

25: 8 106 162 264 427 428 430 433 442 501 503 505 522 604 623

26: 149 166 252 260 502 507 508 520 521 523 524 525 526 527 528 556 585

27: 44 424 504 514 530 534 536 540 544 546 548 549 552 554 558 560 562 564 568 570

28: 71 127 132 426 435 436 510 512 516 800

29: 23 48 159 196 201 202 203 208 423 518 588

30: 55 62 68 95 96 134 209 210 241 261 366 422 494

31: 34 99 110 122 126 131 237 431 432

32: 65 75 117 118 148 204 205 266

Figure 14: Partition of the set of 418 3-digits classes into 32 structurally 'similar' groups performed by the algorithm CONCOR on the connection matrix $C(W2)$. Colours identify NBER 1-digit Categories, as in 4

1: 246, 307, 320, 323, 324, 327, 329, 330, 331, 332, 333, 334, 340, 342, 343, 363, 367, 370, 375, 377, 379, 381, 455

3: 186, 235, 359, 360, 369, 386

4: 23, 29, 51, 65, 75, 79, 117, 118, 125, 134, 136, 148, 156, 165, 174, 200, 204, 205, 218, 219, 225, 228, 252, 257, 310, 313, 315, 322, 335, 336, 337, 338, 361, 373, 419, 420, 427, 428, 429, 438, 439, 445, 451, 501, 505

5: 250, 349, 353, 355, 356, 362, 372, 374, 385

6: 216, 347, 430, 503, 522

7: 73, 132, 424, 435, 436, 530

9: 2 , 4, 5, 7, 12, 30, 36, 54, 66, 72, 87, 112, 128, 140, 147, 163, 227, 251, 278, 289, 351, 378, 392, 417, 433, 442, 450, 482, 600, 601, 602, 604, 606, 607, 623

10: 16, 49, 70, 114, 160, 180, 185, 188, 213, 244, 254, 258, 267, 280, 292, 293, 294, 295, 296, 297, 301, 305, 403, 440

13: 60, 74, 91, 92, 123, 137, 181, 184, 192, 236, 237, 239, 277, 290, 291, 303, 415, 416, 418, 475, 477

14: 24, 42, 63, 84, 89, 108, 116, 124, 144, 150, 190, 211, 223, 224,248, 273, 312, 410, 446, 472, 473

21: 8, 44, 71, 106, 127, 159, 166, 196, 201, 203, 208, 260, 502, 504, 507, 508, 510, 512, 514, 516, 518, 520, 521, 523, 524, 525, 526, 527, 528, 534, 536, 540, 544, 546, 548, 549, 552, 554, 556, 558, 560, 562, 564, 568, 570, 585, 800

22: 19, 28, 57, 68, 69, 152, 164, 264, 425

25: 48, 55, 95, 96, 110, 122, 126, 169, 202, 210, 261, 285, 376, 405, 422, 423, 431, 454, 588

27: 99, 138, 206, 217, 220, 221, 383, 413, 426, 452

39: 26, 34, 100, 141, 162, 209, 222, 232, 241, 266, 298, 366, 406, 414, 432

40: 38, 83, 101, 102, 149, 199, 226, 234, 270, 271, 283, 314, 346, 399, 400, 412, 462, 492

42: 318, 388

44: 1, 178, 187, 276, 326, 341, 345, 348, 358, 365, 380, 382, 395, 434, 700, 701, 702, 704, 705, 706, 707, 708, 709, 710, 711, 712, 713, 714

Figure 15: Class composition of the 'main' groups (groups with 9 or more classes, or sharing classes with the leading ACS of $W2_A$) in the partition of the set of 3-digits classes into 44 $Q$ maximizing groups performed by our adaptation of the algorithm in [24]. Colours identify NBER 1-digit Categories, as in 4. Interestingly, there is a meaningful overlapping between the large ITC group 31 of the Newman partition for period $W1$ and the union of groups 44 and 1 in this partition. groups 44 and 1 give the highest and second highest weighted contribution to modularity, respectively.

| CONCOR | | | | NEWMAN | | | |
|---|---|---|---|---|---|---|---|
| i | n(i) | Q(i) | 100Q(i)/n(i) | n(i) | Q(i) | 100Q(i)/n(i) | i |
| 1 | 21 | 0,067554 | 0,1976741 | 23 | 0,042229 | 0.1836 | 1 |
| 2 | 15 | 0,028377 | 0,1162501 | 5 | 0,0056993 | 0.1140 | 2 |
| 3 | 8 | 0,012253 | 0,0941175 | 6 | 0,011658 | 0.1943 | 3 |
| 4 | 6 | 0,010877 | 0,1113976 | 45 | 0,067952 | 0.1510 | 4 |
| 5 | 13 | 0,021183 | 0,1001295 | 9 | 0,016011 | 0.1779 | 5 |
| 6 | 11 | 0,022052 | 0,1231894 | 5 | 0,0083192 | 0.1664 | 6 |
| 7 | 7 | 0,014192 | 0,1245844 | 6 | 0,0089004 | 0.1483 | 7 |
| 8 | 6 | 0,009475 | 0,0970338 | 2 | 0,002659 | 0.1330 | 8 |
| 9 | 11 | 0,012394 | 0,0692368 | 35 | 0,052306 | 0.1494 | 9 |
| 10 | 14 | 0,019002 | 0,0834045 | 24 | 0,031749 | 0.1323 | 10 |
| 11 | 5 | 0,006616 | 0,0813038 | 8 | 0,010468 | 0.1308 | 11 |
| 12 | 10 | 0,020869 | 0,1282389 | 7 | 0,0075338 | 0.1076 | 12 |
| 13 | 5 | 0,0039 | 0,0479293 | 21 | 0,030472 | 0.1451 | 13 |
| 14 | 8 | 0,009211 | 0,0707498 | 21 | 0,028067 | 0.1337 | 14 |
| 15 | 4 | 0,004089 | 0,0628167 | 1 | 0,0011228 | 0.1123 | 15 |
| 16 | 8 | 0,01374 | 0,1055394 | 2 | 0,0026095 | 0.1305 | 16 |
| 17 | 17 | 0,020218 | 0,0730815 | 1 | 0,0010576 | 0.1058 | 17 |
| 18 | 23 | 0,033912 | 0,0906032 | 2 | 0,0019853 | 0.0993 | 18 |
| 19 | 19 | 0,027984 | 0,0905054 | 2 | 0,0018687 | 0.0934 | 19 |
| 20 | 19 | 0,024231 | 0,0783675 | 1 | 0,00096758 | 0.0968 | 20 |
| 21 | 42 | 0,038334 | 0,0560858 | 47 | 0,063144 | 0.1343 | 21 |
| 22 | 17 | 0,01974 | 0,0713537 | 9 | 0,010899 | 0.1211 | 22 |
| 23 | 9 | 0,010216 | 0,069752 | 1 | 0,0013517 | 0.1352 | 23 |
| 24 | 17 | 0,022459 | 0,081182 | 6 | 0,0078091 | 0.1302 | 24 |
| 25 | 15 | 0,026703 | 0,1093923 | 19 | 0,028877 | 0.1520 | 25 |
| 26 | 17 | 0,029153 | 0,1053786 | 1 | 0,0017649 | 0.1765 | 26 |
| 27 | 20 | 0,020944 | 0,0643499 | 10 | 0,01361 | 0.1361 | 27 |
| 28 | 10 | 0,012898 | 0,0792575 | 4 | 0,0054 | 0.1350 | 28 |
| 29 | 11 | 0,010113 | 0,0564944 | 7 | 0,0092415 | 0.1320 | 29 |
| 30 | 13 | 0,020976 | 0,0991511 | 1 | 0,00056132 | 0.0561 | 30 |
| 31 | 9 | 0,011306 | 0,0771942 | 1 | 0,0013457 | 0.1346 | 31 |
| 32 | 8 | 0,009525 | 0,0731625 | 1 | 0,0005597 | 0.0560 | 32 |
| | | | | 1 | 0,0010779 | 0.1078 | 33 |
| | | | | 1 | 0,0012627 | 0.1263 | 34 |
| | | | | 1 | 0,00048704 | 0.0487 | 35 |
| | | | | 1 | 0,00062563 | 0.0626 | 36 |
| | | | | 7 | 0,0085514 | 0.1222 | 37 |
| | | | | 6 | 0,0073054 | 0.1218 | 38 |
| | | | | 15 | 0,015592 | 0.1039 | 39 |
| | | | | 18 | 0,025304 | 0.1406 | 40 |
| | | | | 1 | 0,0011443 | 0.1144 | 41 |
| | | | | 2 | 0,0032124 | 0.1606 | 42 |
| | | | | 4 | 0,0047314 | 0.1183 | 43 |
| | | | | 28 | 0,080356 | 0.2870 | 44 |
| **C(W2)** | **418** | **0,614495** | | **418** | **0,62785** | | **C(W2)** |

Figure 16: Group contributions to $Q$ modularity, weighted and unweighted by number of classes per group. Partitions generated by CONCOR and Newman algorithms on $C(W2)$. Blue entries refer to the 'Core Groups'.

**Modularity: class partition by subcategory**

| | 1975-1986 | 1987-1999 |
|---|---|---|
| subcat. | Q(i) | Q(i) |
| 11 | 0,012064 | 0,0083618 |
| 12 | 0,011066 | 0,0098523 |
| 13 | 0,020171 | 0,016662 |
| 14 | 0,011017 | 0,0068588 |
| 15 | 0,023214 | 0,018136 |
| 19 | 0,020158 | 0,016958 |
| 21 | 0,020131 | 0,022677 |
| 22 | 0,027234 | 0,031 |
| 23 | 0,026664 | 0,023277 |
| 24 | 0,020847 | 0,023604 |
| 31 | 0,012689 | 0,012595 |
| 32 | 0,030077 | 0,03077 |
| 33 | 0,016459 | 0,015226 |
| 39 | 0,020641 | 0,016632 |
| 41 | 0,014486 | 0,01609 |
| 42 | 0,015028 | 0,017048 |
| 43 | 0,015384 | 0,014143 |
| 44 | 0,01641 | 0,014861 |
| 45 | 0,018402 | 0,017237 |
| 46 | 0,028925 | 0,030943 |
| 49 | 0,015925 | 0,015967 |
| 51 | 0,015154 | 0,013566 |
| 52 | 0,013972 | 0,011623 |
| 53 | 0,020369 | 0,017744 |
| 54 | 0,02023 | 0,018484 |
| 55 | 0,01539 | 0,015657 |
| 59 | 0,012985 | 0,013803 |
| 61 | 0,01739 | 0,015989 |
| 62 | 0,016927 | 0,017526 |
| 63 | 0,015216 | 0,014982 |
| 64 | 0,023013 | 0,018805 |
| 65 | 0,014843 | 0,015793 |
| 66 | 0,019905 | 0,013316 |
| 67 | 0,014378 | 0,012687 |
| 68 | 0,016992 | 0,018203 |
| 69 | 0,014118 | 0,013657 |
| Total | 0,647874 | 0,6107339 |

Figure 17: Group contributions to modularity according to the NBER partition of 418 3-digit classses into 36 2-digits subcategories. Colours identify NBER categories as in Fig. 4

**H* on unweighted 32x32 matrix**

| block | input distr. | output distr. |
|---|---|---|
| 1 | 0,55179 | 0,62112 |
| 2 | 0,49952 | 0,45803 |
| 3 | 0,49585 | 0,38539 |
| 4 | 0,40006 | 0,42929 |
| 5 | 0,47073 | 0,44328 |
| 6 | 0,50928 | 0,52245 |
| 7 | 0,52574 | 0,54867 |
| 8 | 0,44905 | 0,4655 |
| 9 | 0,4388 | 0,44783 |
| 10 | 0,41861 | 0,43937 |
| 11 | 0,44019 | 0,43809 |
| 12 | 0,54338 | 0,55387 |
| 13 | 0,49401 | 0,48727 |
| 14 | 0,54077 | 0,44595 |
| 15 | 0,30394 | 0,28059 |
| 16 | 0,37925 | 0,40569 |
| 17 | 0,31749 | 0,34243 |
| 18 | 0,50404 | 0,48005 |
| 19 | 0,50726 | 0,5056 |
| 20 | 0,44991 | 0,47037 |
| 21 | 0,40343 | 0,35045 |
| 22 | 0,32057 | 0,34701 |
| 23 | 0,42794 | 0,36886 |
| 24 | 0,4514 | 0,3756 |
| 25 | 0,36602 | 0,39246 |
| 26 | 0,47451 | 0,43047 |
| 27 | 0,5949 | 0,60781 |
| 28 | 0,38655 | 0,42523 |
| 29 | 0,34918 | 0,30634 |
| 30 | 0,39227 | 0,43734 |
| 31 | 0,4445 | 0,42645 |
| 32 | 0,31926 | 0,2845 |

**H* on 36x36 unweighted matrix**

| subcategory | input distr. | output distr. |
|---|---|---|
| 11 | 0,17672 | 0,17323 |
| 12 | 0,17233 | 0,16366 |
| 13 | 0,37932 | 0,32954 |
| 14 | 0,33494 | 0,25759 |
| 15 | 0,41147 | 0,40042 |
| 19 | 0,42455 | 0,41261 |
| 21 | 0,50328 | 0,45919 |
| 22 | 0,38226 | 0,55381 |
| 23 | 0,29912 | 0,38995 |
| 24 | 0,50524 | 0,4704 |
| 31 | 0,40314 | 0,47732 |
| 32 | 0,6095 | 0,61185 |
| 33 | 0,32431 | 0,47084 |
| 39 | 0,43435 | 0,39762 |
| 41 | 0,43571 | 0,36686 |
| 42 | 0,50844 | 0,45073 |
| 43 | 0,35293 | 0,32019 |
| 44 | 0,33316 | 0,31895 |
| 45 | 0,39287 | 0,38283 |
| 46 | 0,4924 | 0,6228 |
| 49 | 0,33032 | 0,28123 |
| 51 | 0,37365 | 0,33923 |
| 52 | 0,28814 | 0,30873 |
| 53 | 0,51518 | 0,52843 |
| 54 | 0,43445 | 0,47727 |
| 55 | 0,49357 | 0,46358 |
| 59 | 0,38952 | 0,32853 |
| 61 | 0,50637 | 0,52189 |
| 62 | 0,60722 | 0,58737 |
| 63 | 0,53646 | 0,50044 |
| 64 | 0,49248 | 0,58134 |
| 65 | 0,46961 | 0,46472 |
| 66 | 0,4208 | 0,38367 |
| 67 | 0,34455 | 0,25277 |
| 68 | 0,46893 | 0,40769 |
| 69 | 0,31904 | 0,32433 |

Figure 18: Normalized Herfindal concentration indexes concerning the distributions of patent citations made (input) and received (output) by each group in the 32 and 36 partitions for the period 1987-1999.