AN APPLICATION OF GRAPHICAL MODELS TO THE INNOBAROMETER SURVEY: A MAP OF FIRMS' INNOVATIVE BEHAVIOUR

Cinzia Carota, Alessandra Durio
Department of Economics and Statistics "Cognetti de Martiis", University of Turin, Italy

Marco Guerzoni[1]
Department of Economics and Statistics "Cognetti de Martiis", University of Turin and CRIOS, Bocconi University, Milan, Italy

Abstract *Probabilistic graphical models successfully combine probability with graph theory and therefore provide applied statisticians with a powerful data mining engine. Graphical models are a good framework for formal analysis, allowing the researcher to obtain a quick overview of the structure of association among variables in a system. This paper is the first attempt to apply high-dimensional graphical models in innovation studies, since the increasing availability of data in the field and the complexity of the underlying processes are calling for new techniques which can handle not only a large amount of observations, but also rich datasets in terms of number and relations among variables. In this context, the process of variables and model selection became more arduous, influenced by biases of the scientist and, in the worst case scenario, subject to scientific malpractices such as the p-hacking behavior. On the contrary, high-dimensional graphical models allow for bottom-up, hypotheses free, data-driven, and see-through approach.*

Keywords: *Graphical Model, High-dimensional Data, Innovation Studies*

1. INTRODUCTION

"Probabilistic graphical models are an elegant framework which combines uncertanty (probabilities) and logical structure (independence constraints) to compactly represent complex, real-world phenomena. [...] In the last two decades graphical models have enjoyed a surge of interest due both to the flexibility and power of the representation and the increased ability to effectively learn and perform inference in large networks" (Koller et al., 2007, p.13). Fields of application range from bio-informatics to image segmentation, human pose estimation or language processing. A recent application in the field of financial risk is in Ahelegbey and Giudici (2014) We surmise that this approach could be productively applied to

[1]Corresponding author: marco.guerzoni@unito.it

innovation studies, in which the increasing availability of data is calling for new techniques which can handle not only a large amount of observations, but also rich datasets in terms of number and relations among variables. At the state of the art in the field, both variables and the model to be tested are based upon either existing theories or educated guesses. The process of variables selection might therefore be heavily influenced by biases of the scientist and, in the worst case scenario, subject to scientific malpractices such as the p-hacking behavior. Graphical models on the contrary allow for bottom-up, hypotheses free, data-driven, and see-through approach.

The aim of the paper is to explore the use of probabilistic graphical models in innovation studies, a cross-boarder field between economics and management studies. The novelty of the paper is twofold. First, we consider graphical models for high-dimensional data and we apply them to the realm of management and economics of innovation. Many graph-theoretic operations scale poorly, and the set of tools and strategies to deal with the computational problems in the high-dimensional case is very recent and differs from the traditional one. For instance, the pioneering application Giudici and Carota (1992), who made use of the conditional independence graph to study the innovation process in the software industry, considered 6 variables only and few given ex ante research hypotheses.

Secondly, this method allows us to approach debates in the field of innovation studies from a radically new perspectives since we are able to render a global view of the whole dependence structure of various phenomena which are usually and erroneously kept apart from each other. Thus, the contribution of the paper is not purely a methodological one, for we apply graphical models to all the variables in the Innobarometer dataset and address the many debated and interrelated issues in economics and management of innovation.

A key research area in the field concerns the determinants and the sources of innovation (Freeman, 1994): which firms are more likely to introduce new product and process? Do innovations origin from technological opportunities or rather from the recognition of emerging market needs? Which role do public research institutions play? Does the innovative performance simply depends by the innovative efforts of individual or rather by the framework conditions of a competitive system? Are public policies such as standard settings and R&D subsidies really effective?

Despite the wide acknowledgement in the community of innovation studies that the process of innovation is a complex one and that the aspects mentioned above are necessarily interrelated, scholars have been trying to answer these ex-

emplary questions with standard regressions techniques on a small subset of the variables at disposal. On the contrary, in this paper we map the overall structure of the variables in the dataset taking into account almost all variables and without any a priori on their conditional independence structure.

We show that graphical models applied to innovation studies can corroborate most of the previous results in the field such as the diverse nature of product and process innovations, the distinct effect of technology and market opportunities, the role of interactions with research institutes and users. More specifically, we will contribute to the technology-push vs.demand-pull debate which discusses whether human needs blindly follow technological advances or whether the process of technological development and scientific discoveries is pulled by priorities set up by the society and by potential consumers.

In the next section we briefly present the main debate in innovation studies. In section 3, we detail the statistical method we use. In section 4, we present the data and the results. Conclusions follow.

2. INNOVATION STUDIES: A QUICK OVERVIEW

Innovation can be broadly defined as the introduction of a new product, a new process or a new organizational form and it is recognized as a key driver for the economic performance both at the firm and the aggregate level (Nelson and Winter, 1982; Solow, 1957). A very much debated issue is whether the origin of innovation is technology-push or demand-pull, that is whether new ideas stem from the independent and serendipitous processes of scientific discoveries and technological improvements, or, conversely, whether the elicitation of human needs steers the route of science and technology.

Some scholars share a faith in the capability of mankind to continuously develop technology along any possible direction. This faith in technology dates back to the tremendous leap forward in the early second half of the last century in the realms of nuclear technology, computer science, aerospace, and telecommunication. Bush (1945) devised the expression "Science, the endless frontier" to describe the view of a world where technological opportunities are endless ones and the needs of individuals and firms define the priorities for the research. Therefore, in this social context of the postwar period the demand-pull hypothesis diffuses among scholars in economics of innovation.

In the late 1970s, the faith in the mankind and in its ability to fulfill any human needs with advancements in technology declined. In the literature, which in social science is often wired with historical developments of a society, a very dis-

ruptive critique to the demand-pull approach came from Mowery and Rosenberg (1979) and Dosi (1982). In their view, the process of innovation is reverted: technology ceases to be an endless frontier, since engineers cannot explore the whole space of technology opportunities but only a narrow contour of the state of the art. In other words, the logic of the discoveries of new products and processes is partly serendipitous and largely independent by any consideration of a possible final use. Conversely, the set of human needs is endless and, thus, unable to provide a clear route for the technological development. Within this approach, technical knowledge is developed either within or externally to the firm. The internal knowledge is the result of investment in research and development, in the hiring and training of scientific personnel and in the acquisition of new technologies embedded in machineries. Externally, a firm can tap into knowledge provided by other firms, but it can also intentionally exploit the knowledge produced in universities and research centers via collaborations (Arundel and Geuna, 2004).

This technology-based view persists as the most widely accepted in the literature. However, in the last twenty years, efforts have been also made to overcome the Dosi's critique and revitalize the demand-pull hypothesis. Most of the efforts have been put forward to empirically operationalize the conceptualization of demand as a blend of size of the market and information. In this approach, the role of demand is not generally linked with a ill-defined set of general needs, but the focus shifted either to the very precise case of the interaction between users and producers or to the size of market as an incentive device. Fontana and Guerzoni (2008) showed that the interaction with users impinges more on the likelihood of introducing product innovations, while the market size has a larger impact on process innovations. More recently also the role of governmental demand has been considered as a possible determinant for the innovative activity as it has been traditionally the case in sectors such as defense and aerospace (Guerzoni and Raiteri, 2015).

Although scholars agree that the dichotomy technology-push vs.demand-pull is a sterile one (Freeman, 1994), very few empirical works attempt to take into account both sides of the story. A reason for this disappointing state of the art could be that scholars in the field are not always equipped to deal with challenging empirical issues. Indeed, the complexity of the debate is mirrored by the intricate structure of the available data: it is arduous to measure innovation since it involves non codified knowledge, implicit know-how, and intangible assets. Moreover a successful innovation is the rare positive outcome of long efforts which alongside produced failures which do not leave a paper trial. As a result, some scholars

innovation studies opted long ago for a survey based approach, where firms can describe in detail their innovative activity.

Oddly enough, the recently increased availability of information did not raise any concern for the selection of the empirical model and for the choice of the variables. For instance, there are no controls for the existence of false positives: any statistical result should consider the possibility that the reality observed in data is a fortunate random construct of the sampling and not a property of the population (Nuzzo, 2014). Secondly, the p-value is often the only figure shown in the tables. The increasing size of the now available samples makes the simple p-value of a scarce information content since for large samples the p-value goes quickly to zero. Third, surveys have many variables which are characterized by an overlapping information content and there exist redundancies and correlations among them. The choice of the variables to include or to omit in the analysis is a crucial one, otherwise a perceived sloppiness in this step of the analysis would cast doubts of scientific malpractice such as the p-hacking behaviour, that is "trying multiple things until you get the desired result"(Simonsohn et al., 2014).

We propose the use of graphical models as a tool to help the researcher in understanding the structure of a dataset as a whole, without omitting any variables on the base of educated guesses or introducing simplifying assumptions on existing relationships. In this way we mitigate the risk of both relevant omissions and thus endogeneity in the empirical model, and we reduce biases in the estimation. As we explain in the next section, although we are forced to make statistical assumptions to handle computational problems, the approach remain bottom-up, free from economic hypotheses, and data-driven.

3. GRAPHICAL MODELS IN HIGH-DIMENSIONAL PROBLEMS

Essentially graphical models used in this paper are classes of multivariate distributions whose conditional independence properties are encoded by a graph in the following way. The random variables are represented as vertices (nodes), and two vertices are connected by an edge (line) when the corresponding variables are not conditionally independent given the other variables represented in the graph. An introduction to graphical models can be found in Lauritzen (1996), Wermuth and Lauritzen (1990), Whittaker (1990), and Pearl (1988), while high-dimensional graphical models are discussed for instance in Højsgaard et al. (2012, Ch. 7).

Aiming at a data analysis free of any economic hypotheses in order to rediscuss the empirical literature in innovation studies, a fundamental task of this paragraph is to stress the statistical assumptions underlying our analysis. In the

next paragraph we provide all formal definitions required to discuss them in details.

A *graphical model* is defined as an undirected graph $G = (\aleph, E)$, where $\aleph = \{X_1, ...X_p\}$ is the set of vertices and E is the set of edges, i.e. a subset of $\aleph \times \aleph$ (unordered pairs), where multiple edges and self-loops are not allowed. The number p of vertexes is assumed to be finite and the absence of an edge $e = (X_u, X_v) \in E$ connecting two vertices means

$$X_u \perp X_v | \aleph \setminus \{X_u, X_v\}$$

(pairwise Markov property). This is the key feature of a graphical model: conditional independence can be read in the graph.

A subset $A \subseteq \aleph$ is *complete* if every pair of vertices in A is connected by an edge. If a subset is maximally complete, that is complete and not contained in a larger complete subset, it is called a *clique* . In a graph G two vertices, X and Y, are said to be connected if there is a sequence $X = X_1, ..., X_k = Y$ of distinct vertices such that $(X_{i-1}, X_i) \in E$, $\forall i = 2, ..., k$. The sequence $X = X_1, ..., X_k = Y$ is a *path* of length $k-1$. A subset $C \subseteq \aleph$ *separates* two disjoint subsets of $\aleph$, A and B, if all paths from a vertex X in A to a vertex Y in B passes through C. A *cycle* is a path where the end vertices are the same $(X = Y)$. A cycle is chordless if X_u and X_v are only connected by an edge when $|u - v| = 1$.

A graph is called *triangulated* if it has no chordless cycles of length greater than three. A *perfect ordering* of the nodes, which is equivalent to a perfect sequence of the cliques, exist if and only if the graph is triangulated. Finally, a graph is *decomposable* if and only if it is triangulated (Lauritzen, 1996).

The variables in a graphical model can be discrete, continuous, or both (mixed). In the first case, in which each variable assumes a value in a set of levels, the models are based on the multinomial distribution. This is our case.

Labelling by $1, ...|X_v|$ the levels that the discrete random variable X_v may take, so that $|X_v|$ represents the number of its levels, we write a generic observation (or cell) as $\mathbf{x} = (x_1, ..., x_p)$, and the set of possible cells as χ. Given a dataset with n observations of the p discrete random variables $\mathbf{X} = (X_1, ..., X_p)$ with $X_v \in \aleph$, we assume that the observations are independent and are interested in modelling the probabilities $p(\mathbf{x}) = Pr(\mathbf{X} = \mathbf{x})$ for $\mathbf{x} \in \chi$.

This kind of problem is known as *structure learning* or structure estimation since we are interested in learning the structure of the probability function and not simply in quantitative learning, such as estimations or tests, of the unknown parameters of a given model. In other words, the graphical structure itself, that is

to say interactions *and* conditional independence relationships between variables have to be simultaneously estimated from the data.
In this sense, such problem emerging as a model selection problem in graphical models, is parallel to e.g. density estimation, or, alternatively, it could be interpreted as a data mining problem with focus on discovering relations between variables in a complex system.

Model selection in graphical models represents a very challenging issue when many structures have to be considered, particularly with sample sizes not much larger than the number of variables - the so called "small-n-large-p situations"-, or, more in general for the discrete case, with sparse tables resulting also from a few variables with a large number of levels. In this respect, in section 4 we highlight that the size of the Innobarometer dataset is significantly reduced by the high-dimensionality of the multivariate analysis, which requires all levels of all variables to be represented and missing values are not allowed. Therefore, although in principle model selection in graphical models consist of finding the, in a some sense, *best fitting* graph for the given dataset, it is often not possible to perform an exhaustive search and it becomes necessary the use of standard stepwise search methods. Forward search, i.e. edge addition starting from the model of complete independence or from the graph with no edges, is preferred to the backward search when the saturated model cannot be easily fitted. However, forward searches may terminate in local optima as it can always happen in greedy search, that is an algorithm that proceeds by repeated local optimizations.

Moreover, a massive efficiency gain in computations can be achieved if inference or model search are restricted to decomposable models, where the maximum likelihood estimates (MLE) of the expected cell counts, $m(\mathbf{x}) = n \times p(\mathbf{x})$, exist in explicit form and depend on the observed marginal tables on the biggest cliques:

$$\hat{m}(\mathbf{x}) = \frac{\prod_{C \in \mathscr{C}} n(\mathbf{x}_C)}{\prod_{S \in \mathscr{S}} n(\mathbf{x}_S)^{\nu(S)}},$$

where $\mathscr{C}$ is the set of cliques which form a perfect sequence, $\mathscr{S}$ is the set of separators for this sequence, and $\nu(S)$ is the multiplicity of S as separator in this sequence. This implies also that *good* inferences can be obtained if such marginal tables satisfy the conditions required to obtain the desired *good* inferences (for instance if table margins are positive when we desire MLEs) and this is clearly a remarkable simplification of the problem.

In our case, there are 61 variables (nodes) implying $2^{p(p-1)/2} = 2^{1830}$ possible undirected graphs. So we adopt a radical way of dealing with high-dimensional sparse data and restrict attention to forests and tree graphs. A forest is an undi-

rected graph with no cycles and it may be composed of several connected components called trees, i.e. a tree is a connected acyclic graph (Bondy and Murty, 2008). We adopt this approach as a preliminary step towards the understanding of the overall dependence structure of our high-dimensional discrete dataset. Trees and forests can be unrealistically simple models, but can provide useful insights about identification of distinct connected components which can be analysed separately (i.e. dimension reduction), identification of neighbourhoods for more detailed analyses, identification of hub nodes and other interesting features. Moreover, they can be useful as initial models for search algorithms with a larger search space, for example decomposable models. In other words they provide a tentative network approximating the joint distribution of our variables.

From a formal point of view trees (forests) are decomposable graphical model, with cliques of size two (or less) and such that any two non-adjacent nodes are separated by a set of (at most) size one. Under the assumption that the cell probabilities factorize according to an unknown tree τ, they can be written as

$$p(\mathbf{x}|\tau) = \frac{\prod_{u,v\in E} p(x_u,x_v)}{\prod_{v\in \aleph} p(x_v)^{deg(v)-1}} = \prod_{v\in \aleph} p(x_v) \prod_{u,v\in E} \frac{p(x_u,x_v)}{p(x_u)p(x_v)},$$

where for simplicity the nodes in the graph are denoted by their indices and the degree of v, $deg(v)$, is the number of edges incident to X_v.

An efficient algorithm to find the maximum weight spanning tree of a arbitrary undirected connected graph W with p nodes and positive edge weights has been proposed by Kruskal (1956). Starting with the null graph, the edge with the largest weight among the edges not yet chosen is added to the graph provided that it does not form a cycle with the ones already included. When $p-1$ edges have been added, the maximum weight spanning tree (MWST) of the graph W has been found. Building on this work, Chow and Liu (1968) showed that the log-likelihood maximized over p for a fixed τ yields the profile log-likelihood

$$\hat{l} = log(L(\tau,\hat{p})) = \sum_{(u,v)\in E} I_{u,v} + \sum_{(v)\in \aleph} I_v$$

where $I_{u,v}$ is the mutual information or empirical cross-entropy between endpoint variables of the edge $e = (X_u, X_v)$, i.e.

$$\sum \frac{n(x_u,x_v)}{n} log \frac{n(x_u,x_v)/n}{n(x_u)n(x_v)/n^2},$$

where the sum extends to all possible levels x_u and x_v of X_u and X_v and $n(x_u,x_v)$ is the number of observations with $X_u = x_u$ and $X_v = x_v$. Similarly, I_v denotes the

empirical entropy of X_v. Therefore, by using the $I_{u,v}$ as edge weights on the complete graph with vertex set $\aleph$, and applying a maximum spanning tree algorithm, one can obtain the maximum likelihood tree.
The mutual information is written emphasizing how the MLEs from the pairwise marginals enter the formula. Note also that $I_{u,v}$ is one half of the usual likelihood ratio test statistic for marginal independence of X_u and X_v, that is $G^2 = -2log\Lambda = 2I_{u,v}$ (Agresti, 2013, Ch.3). Under marginal independence G^2 has an asymptotic χ^2_k distribution, where $k = (|X_u| - 1)(|X_v| - 1)$ represents the number of additional free parameters required under the alternative hypothesis, compared with the null hypothesis.

A disadvantage of the previous approach, based on the complete graph W on $\aleph$ with edge weights given by $I_{(u,v)},\ u,v \in \aleph$, is that it always results in optimal trees, not forests. To take account of the number of model parameters in some fashion Edwards and Labouriau (2010) replace the maximun likelihood with other well-established information criteria, particularly AIC (the Akaike information criterion) (Akaike, 1974) and BIC (the Bayesian information criterion) (Schwarz, 1978). Both criteria correspond to penalized likelihoods. The former is defined as $-2log(L) + 2r$, where L is the maximized likelihood under the model and r is the number of parameters in the model, and the latter as $-2log(L) + log(n)r$. Accordingly, in order to find the forest F with the minimum AIC or BIC, Edwards and Labouriau (2010) add a penalty for each edge, proportional to the number of additional parameters $k_{u,v}$ of introducing the edge e. Then, banning all the edges whose weights are negative, they find the optimal forest F by carring out the Kruskal's algorithm on the rest. In this case the weight of $e = (u,v)$ is taken to be a penalized mutual information, $I_{u,v} - k_{u,v}$ or $I_{u,v} - \frac{1}{2}log(n)k_{u,v}$, respectively.

This approach is attractive with high-dimensional data, since if the selected forest does consist of multiple connected components these may then be analyzed separately allowing a dimension reduction. Interestingly, the connected components of the minimal AIC/BIC forest are also connected components of the minimal AIC/BIC decomposable model, providing further justification for this procedure. A further interesting property is that the global optimality of the selected minimal tree holds under marginalization when the subset of variables of interest in turn is a tree.

A completely different criterion leading to maximum a-posteriori (MAP) forests has been proposed by Panayidou (2011) in a Bayesian contest. We do not enter in details of this proposal.

4. AN APPLICATION TO THE INNOBAROMETER SURVEY

Innobarameter is a dataset of 81 variables describing firms' innovative behaviour and collected via telephone-aided interviews to a stratified sample of 5234 European firms in the manufacturing sector. The dataset has already been used for scientific analysis in innovation (Guerzoni and Raiteri, 2015) and policy reports by the European Commission. Moreover the structure of this survey mimics the Community Innovation Survey, another widely used source of data in innovative studies. The detailed description of the dataset, of the questions, including the distribution of the variables and a rich set of descriptive statistics can be found in (Vv.Aa, 2009).[2] Excluding 18 "filtered" or "accessories" variables (questions only asked to whom provided a specific answer to a previous question) and one irrelevant variable (d5), 62 variables are eligible for the multivariate analysis.

Only one variable (q8) belonging to the latter group has been removed because of a huge number of missing values. This happened because q8 presupposes a positive answered to q7, that is it depends *de facto* on such a filter question, even though this fact is not clearly established in the survey.

All of the remaining 61 variables have been codified as follows: the answers *Don't Know-Not Available* are taken to be NA, while the answer *Not Applicable* is taken to be *NO*. In addition, given the nature of questions q2, q3, q4, and q5, they have been codified according to a further rule: answers of subjects responding *No* or *Not Applicable* in q1_a-g and jointly *Don't Know-Not Available* in q2, q3, q4 and q5 have been re-codified into the same variables, respectevely q2, q3, q4 and q5, as *NO*. Such a decision originates from the fact that the answers *NO* or *Not Applicable* in q1_a-g implies the answer *NO* in the four subsequent questions (other variables). Also in this case the text of the survey was not well presented to the respondents.

In conclusion, 52 binary variables and 9 variables taking a number of levels from 4 to 6 enter our analysis. They are listed in Table 1, third column. As mentioned in the previous section, the joint analysis of the structure of their association by high-dimensional graphical models implies a reduction of the sample size from 5234 to 1531.

[2]The Innobarometer analytical report is also available at this web-page: `http://ec.europa.eu/public_opinion/flash/fl_267_en.pdf` or upon request to the authors.

4.1. VARIABLES

In Table 1, first and second columns, we group the variables in five sets according to their economic meaning. The first group consists of output variables, which grasp whether a firm introduced product, process, service innovations, new marketing strategies, and new form of organization. In this group, we also include patent and design applications, which have been widely used in the literature to proxy innovative activities.

The second group entails variables which describe firms' innovative strategies, id est routines, behaviours and actions put deliberately in place in order to improve the innovative performance. The most discussed in the literature are R&D investment strategies, which do not consist in formal R&D only, but also in training for scientist, acquisition of external R&D by hiring expert consultants and by acquiring of new machines. Firms can also collaborate with universities and research centres to tap into their knowledge.Concerning the demand side, firms can also improve their performance by actively engage in collaborations with users, since in many cases, and especially in B2B relationships, users can provide firm with relevant information about their needs to be fulfilled by the innovation (Guerzoni, 2010) or even suggest themselves new product or process improvements (Von Hippel, 1988). In the survey, firms are asked about the intensity of interaction with users or directly whether users are a source of knowledge for the introduction of new ideas. Finally, firms can also make efforts to improve the management of internal and external knowledge flows, for instance by creating ad hoc positions or by investing in training activities.

A further vital decision for firms is market positioning. The size of the market is measured by the potential number of buyers, the size of the local market, the countries in which a firm operate, and the level of internationalization. Firms have also been asked about their strategies for managing the internal and external flows of knowledge and information and their activity of training on the job.

The third group of variables describes the external conditions a firm is facing such as the institutional setting, the competitive pressure, the policy, and the existence of public procurement.

The fourth group consists of two variables only, which we decided to highlight since they grasp the techology-push vs. demand-pull debate. Firms have been asked about the factors having a positive influence upon the innovative activities and the two variables refer to the emergence of a new technology to be exploited (q16_b) and to new opportunities from the market (q16_d).

The last group of variables describe firms' characteristics. In the survey we

have information about the firm's size, age and turnover. Some scholars suggested that large firms should innovate more than others since they have more dedicated resources. Some others however puts forward some evidence for an inverted U-shape relation: small and large firms are the most innovative, while medium size ones lag behind (Acs and Audretsch, 1987). The role of age is less controversial. Young firms are very likely to fail and exit the market. However, if they manage to be successful they growth faster and they are more innovative than the market average (Audretsch, 1995).

Table 1: Variables list

Group	Content	Variable
Innovative output	Product Innovation	q6_a
	Process Innovation	q6_c
	Service Innovation	q6_b
	Organization Innovation	q6_e
	Marketing Innovation	q6_d
	Patent and Design Application	q1_f, q1_g
Firm Innovative Strategy	R&D Invesment	q1_a-e, q2, q3, q4, q5
	Interaction with the users	q12_a-c, q16_f , q18
	Collaboration with University	q12_e, q12_d, q16_a
	Market positioning	d1_a-c, d1_f, q_14, q10_a-e, q17
	Internal Knowledge Management	q9_a, q9_b, q9_c, q9_d
	External Knowledge Management	q11_a-d
	Training	q13_a-e
External Conditions	Institutional Settings	q15_a-e
	Public Procurement	q16_e, q7
	Competitive Pressure	q16_c
Demand vs. Technology	Market opportunities	q16_d
	Technological opportunities	q16_b
Firms Characteristics	Size	d2
	Turnover	d4
	Age	d3

4.2. RESULTS

Figure 1 maps the 61 variables listed in Tabel 1 and describes the conditional independence structure of the dataset through a minimum BIC forest, obtained by the R package gRapHD for efficient selection of high-dimensional undirected graphical models (Abreu and L., 2007). Figure 2 shows the same forest with labels for the most important variables or groups of variables relevant for the analysis we

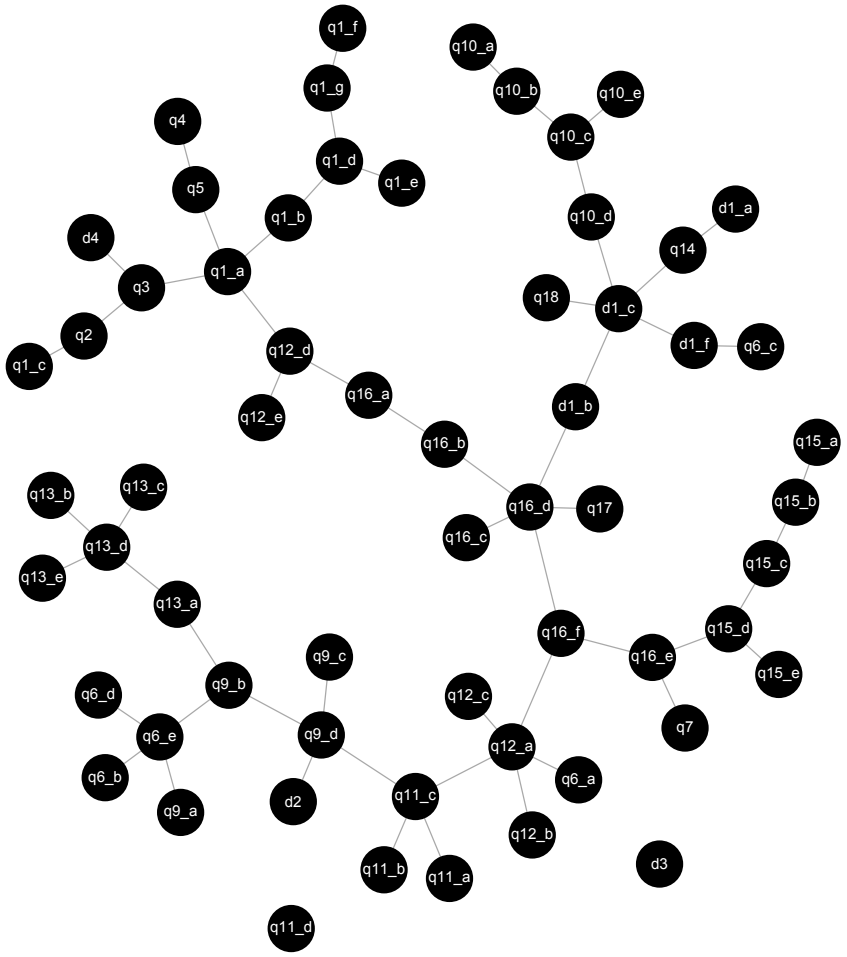

Figure 1: The minimum BIC forest

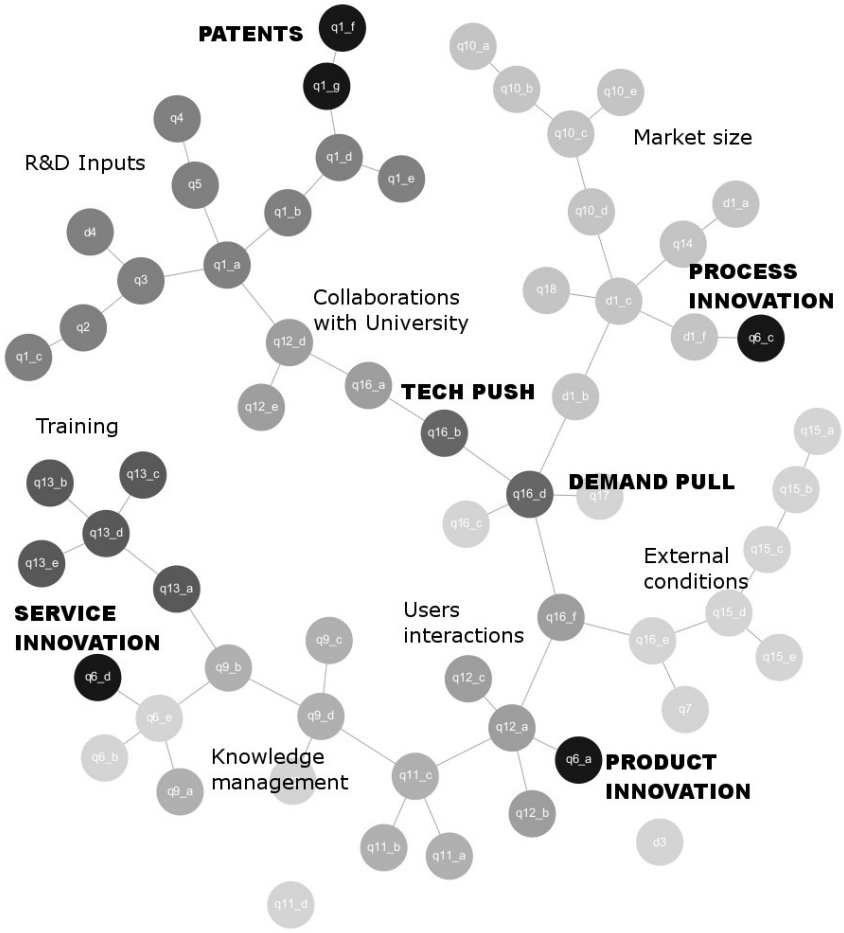

Figure 2: The minimum BIC forest with labels

present in the next paragraphs. For the sake of simplicity, in this paper we do not report the optimal decomposable graphical model corresponding to the minimum BIC forest since many relevant results can be derived from this preliminary analysis alone.

First of all in Figure 1 we observe that the forest consists of three trees or unconnected components: two variables stand alone and are independent from the rest of the structure, while all other variables are a connected in a large tree. One of the unconnected trees is the age of the firm (d3), the second the exchange of property rights (q11_d). While for the latter, there is not necessarily a surprising result because markets for knowledge are rather underdeveloped, for the former it probably depends on the structure of the question asking only whether the firms was founded before or after 2001, which do not really grasp the age of the firms. Before analyzing in details the large tree, we recall that in a tree, two variables can be directly linked by an edge or by a path via one or more nodes. Given that cycles are not allowed, two variables are connected by a unique path. When the connection of two variables is mediated by another node omitting the separating node/variable may induce dependence between the two variables. The longer is the path connecting two variables the weaker is their association since, under the constraint of no cycles, at each step the Kruskal algorithm compares their mutual information to the mutual information of all other pairs of variables included in the path. We first focus on the output variables, namely innovation (product, process, service and organizational) and invention (patent and design application). If they are the outcome of an underlying inventive and innovative process, it reasonable to expect them to be at the end of a chain of dependence. This is clearly the case for product and process innovations which are connected to tree by an edge only and therefore are labeled *leaves*, but also for groups of other output variables in Figure 2.

Both product and process innovation are closely linked to demand factors: product innovation depends on the group of variables capturing the interaction with the users. Similarly, all variables in the tree with the highest proximity to process innovation relate to the size of market. Service, organization, and marketing innovations are closely related with the nodes capturing firms behavior about internal and external knowledge management and training. Patents and design applications are dependent on R&D input variables and conditionally independent of any other output of inventive activity. Without additional information, it seems questionable to consider different types of innovation as homogeneous entities, since they are related to different variables and, thus, possibly generated by

different underlying processes.

At the core of the map, two nodes deserve a peculiar attention. These two nodes depict the variables "technological-push" and "demand-pull", which respectively describe whether a firm declared as important for the introduction of a innovation either a technological innovation or market opportunities. For this reason, the localization of these two variables in the tree, here interpreted as a map, can suggest how to empirically investigate the technology-push vs. demand-pull debate. With this purpose in mind and a slight abuse of terminology, we can observe that they have both a high "degree of centrality", that is they play an important role in the"cohesion" of tree. This is mostly true for the variable demand-pull which, "if removed", would have the large tree of the forest divided in three different and large unconnected components[3]. In other words, such a node separates three sub-trees in the large tree of the forest. It is also noticeable that the same variable has a high proximity with market size and user interaction, that is demand side variables, while the technology-push node is closer to the technology variables such as the interaction with university and the level of investment in R&D.

The level of investment in R&D has been usually considered as the main determinant of the innovative activity. Despite the overall consistency of the tree with this view, a researcher should be aware that the path from nodes describing the R&D investment activity to the innovative outcome consists of more than 10 steps. When modeling R&D investment as a determinant of product and process innovation a researcher should not omit all the variables/nodes in-between in that path. On the other hand, a different story concerns the relation between the level of investment in R&D and patent and design applications, which are directly linked. This layout of the tree suggests that a model which assumes or tests a dependence of patents application and R&D investment is a sound one. This is a very good news for innovative studies, since many scholars tested the parameters of a Cobb-Douglas production function where the outcome are patents as recorded in a patent office and the inputs are different investments in R&D (for a review see Griliches (1998)). However, the bad news is that the number of patent applications is not a good proxy for the number of product and process innovations: the forest suggests that this is strong assumption since the dependence relation between

[3]The terminology "centrality" is borrowed by social network theory (Wasserman and Faust, 1994). Borgatti and Everett (2006) show how the centrality of a node can be conceived as its contribution to the cohesion of a graph, measured using the distribution of the reachability of the nodes.

patents application and innovation emerges as mediated by many variables.[4]

Regarding the variables capturing the institutional conditions, it should be noticed that they are all grouped in the south-east part of the map and the award of a public procurement contract is the closest the product innovation. Given the increased attention of policy makers to the use of public procurement as an industrial policy, data suggest that in case of product innovation this is a very sound hypothesis, while for the case of other type of innovations or for R&D investment this hypothesis might be less reasonable.

Overall, we can send a clear message to scholars in innovation studies. If the aim of a research is to explain the determinants of innovation the following should be carefully taken into account:

- Product and process innovations should be considered as distinct outputs, resulting from partly different generative processes.
- There might be a causal link which flows from R&D investment to innovation. However, if it exist, our preliminary analysis shows that it is mediated by the role of other variables, namely the interaction with university, the existence of technological and market opportunities and the role of demand both as the size of the market and as the interaction with users. For this reason, any empirical study that tests or imposes a model where R&D investments directly influence the level of innovation might omit important endogenous variables.
- This evidence suggest that Freeman's intuition was the correct one: the seeds of future innovation rest in both technology and market opportunities. However, variables capturing demand conditions seems to be more related with the output of the innovative process. Anyway, the role of demand factors, even when not prominent, can not be omitted in an empirical analysis, which would otherwise results a biased one.
- Service, marketing and organizational innovations are very close and they directly depends by investment in human capital (knowledge management and training). In line with literature, they are not closely associated with R&D investment.
- In the map, institutional conditions and competitive pressure are clustered and they exert a direct impact on product innovation only.

[4]Such indication is substantially confirmed by the corresponding decomposable model, which we do not report in this paper for the sake of brevity.

- Patent applications are not a suitable proxy for product or process innovation.

5. CONCLUSIONS

Innovation is both a driver of the performance at the firm level and a determinant of long term productivity growth at the aggregate level. Innovation studies, a cross-boarder field between economics and management, deal with the complex process of innovation that generates and exploits new technological opportunities and compel them to satisfy human needs. The underlying process is uncertain and it involves various factors which affect the system in a non-linear way. The complexity of this phenomenon reflects in the lack of well-behaved data and in the use of survey data, where firms can describe in detailed the multifaceted aspects of their innovative activity.

The increased availability of survey data, the large sample sizes, and the high number of variables call for a cautious approach when performing empirical exercises. Specifically, major problems consist of the selection of the empirical model to test and the choice of the variables. In this paper, we show that high-dimensional graphical models, and specifically simple minimum BIC forests, can serve as a data mining tool and guide the researcher towards a sound empirical design.

The application to this method to a survey on firms' innovative behaviour led to remarkable results. Indeed we encompassed in one picture the most pressing debates in innovation studies. It is not the aim of this research to contribute to any of these single issues in depth. However, we showed that most of them in the past have been erroneously handled as independent ones, whereas the dependency structure of the variable suggests that they should considered together.

This work call for extension along different lines. First of all, Innobarometer is not the only survey, nor the most used. We therefore plan to repeat this exercise on the Community Innovation Survey (CIS) 2012 as soon as it is available. Moreover the CIS is available in different waves and this allows us analyzing the dynamic properties of graph over time. A second step consists of controlling whether models applied in past works are coherent with the graph present here and, if not, we aim at testing whether a different model, suggested by the graph, lead to different results. From a technical point of view, we can introduce many interesting variants in the analysis presented in this paper. They include different weights for edges in the tree such as Bayes factors for independence of X_u and X_v, instead of the mutual information, leading to maximum a posteriori forests.

We can deepen the local analysis on sub-graphs of a given forest, as for instance one including specific neighbours of a given node of interest. We can also remove the no-cycle restriction and search for decomposable graphical models. Finally, of course, the analysis can be restricted by resorting to directed graphical models for high-dimensional data when causal relationships among variables have to be explored.

References

Abreu, G. C. G., E. D. and L., R. (2007). High-dimensional graphical model search with the gRapHD R package. *Journal of Statistical Software*, 37(1):1–18.

Acs, Z. J. and Audretsch, D. B. (1987). Innovation, market structure, and firm size. *The review of Economics and Statistics*, pages 567–574.

Agresti, A. (2013). *Categorical Data Analysis, third edition*. John Wiley & Sons, Ltd, Hoboken, New Jersey.

Ahelegbey, D. F. and Giudici, P. (2014). Hierarchical graphical models, with application to systemic risk. *University CaFoscari of Venice, Dept. of Economics Research Paper Series No 63*, 1.

Akaike, H. (1974). A new look at the statistical identification problem. *IEEE Transaction Auto Contr*, 16(1):716–723.

Arundel, A. and Geuna, A. (2004). Proximity and the use of public science by innovative european firms. *Economics of Innovation and new Technology*, 13(6):559–580.

Audretsch, D. B. (1995). Innovation, growth and survival. *International journal of industrial organization*, 13(4):441–457.

Bondy, A. and Murty, U. (2008). *Graph Theory*, volume 244 of *Graduate Texts in Mathematics*. Springer.

Borgatti, S. P. and Everett, M. G. (2006). A graph-theoretic perspective on centrality. *Social networks*, 28(4):466–484.

Bush, V. (1945). Science: The endless frontier. *Transactions of the Kansas Academy of Science (1903)*, pages 231–264.

Chow, C. K. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *EEE Transactions on Information Theory*, 14:462–467.

Dosi, G. (1982). Technological paradigms and technological trajectories. *Research Policy*, 11(3):147–162.

Edwards, D.and de Abreu, G. and Labouriau, R. (2010). Selecting high-dimensional mixed graphical models using minimal aic or bic forests. *BMC Bioinformatics*, 11.

Fontana, R. and Guerzoni, M. (2008). Incentives and uncertainty: an empirical analysis of the impact of demand on innovation. *Cambridge Journal of Economics*, 32(6):927–946.

Freeman, C. (1994). The economics of technical change. *Cambridge Journal of Economics*, 18:463–514.

Giudici, P. and Carota, C. (1992). Symmetric interaction models to study innovation processes in the european software industry. In L. Fahrmeir, B. Francis, R. G. and Tutz, G., editors, *Advances in GLIM and Statistical Modelling*. Springer-Verlag, Berlin.

Griliches, Z. (1998). Patent statistics as economic indicators: a survey. In *R&D and productivity: the econometric evidence*, pages 287–343. University of Chicago Press.

Guerzoni, M. (2010). The impact of market size and users' sophistication on innovation: The patterns of demand. *Economics of Innovation and New Technology*, 19(1):113–126.

Guerzoni, M. and Raiteri, E. (2015). Demand side vs. supply side technology policies: Hidden treatment and new empirical evidence on the policy mix. *Research Policy, forthcoming*.

Højsgaard, S., Edwards, D., and Lauritzen, S. (2012). *Graphical models with R*. Springer.

Koller, D., Friedman, N., Getoor, L., and Taskar, B. (2007). Graphical models in a nutshell. In Getoor, L. and Taskar, B., editors, *Introduction to Statistical Relational Learning*. MIT Press.

Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7 (1):48–50.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford Statistical Science Series 17. Oxford.

Mowery, D. C. and Rosenberg, N. (1979). The influence of market demand upon innovation. a critical review of some recent empirical studies. *Research Policy*, 8(2):102–153.

Nelson, R. R. and Winter, S. G. (1982). *An evolutionary approach to economic change*. Belknap Press.

Nuzzo, R. (2014). Statistical errors. Nature International weekly journal of science.

Panayidou, K. (2011). Estimation of tree structure for variable selection. PhD thesis, Department of Statistics, University of Oxford.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(1):461–464.

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2):534.

Solow, R. M. (1957). Technical change and the aggregate production function. *The review of Economics and Statistics*, pages 312–320.

Von Hippel, E. (1988). *The source of innovation*. Oxford University Press, New York.

Vv.Aa (2009). Innobarometer analytica report. Technical report, Gallup Organization.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis*. Cambridge University Press.

Wermuth, N. and Lauritzen, S. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models (with discusssion). *ournal of the Royal Statistical Society B*, 52:21–50.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Jhon Wiley and Sons, NY.