Informatica e diritto, XXXIX annata, Vol. XXII, 2013, n. 1, pp. 243-261

Information Extraction and Social Network Analysis of Criminal Sentences. A Sociological and Computational Approach

DEBORAH DE FELICE, GIOVANNI GIUFFRIDA
GIUSEPPE GIURA, VILHELM VERENDEL, CALOGERO G. ZARBA*

1. INTRODUCTION

A main objective of our research is to obtain a description of the socio-economic environment characterizing a trial leading to a criminal sentence, as well as the differences in the conduct of the trial between different jurisdictional administrations.

In this paper we study the juridical response to organized crime activities in Sicily[1], Italy, by analyzing a *corpus* of criminal sentences using computational techniques of information extraction and social network analysis. In particular, the analyzed criminal sentences were pronounced in the four courthouses of Sicily from 2000 through 2006, and were declared irrevocable for at least one defendant. A first element of originality of this paper is in the choice of the source of sentences: criminal sentences are tools of social regulation (although as outcome of a trial), and are used as a "magnifying glass" to study mechanisms and contexts from which one can deduce the

* D. De Felice and G. Giuffrida are assistant professors at the Department of Political and social sciences of the University of Catania (Italy); G. Giura, PhD in Sociology and methods of social sciences, works at the Department of European studies and international integration of the University of Palermo (Italy); V. Verendel has a PhD in the Complex systems group at the Chalmers University of Technology in Gothenburg (Sweden); C. Zarba, PhD, works for Neodata Intelligence S.r.l.

[1] C. PENNISI, G. GIURA, *Un'analisi empirica della giurisprudenza sulla criminalità organizzata e di tipo mafioso*, in "Antigone", 2009, n. 2-3, pp. 125-162.

real balances of power and value hierarchies established by the institutional response to organized crime and mafia-related crimes. A second element of originality is the synergy, pursued since the phase of data retrieval, between computer scientists and social scientists when it comes to methodology and the approach of the research team.

By using direct (expert-based) knowledge of the semantic importance of the source of sentences, we made a preliminary coding and performed statistical exploration of a *corpus* of sentences. To the same *corpus*, we designed and applied automatic textual extraction algorithms trained using computational machine learning techniques. A result of using this approach, especially compared to the source of judiciary statistics, has the following advantages: a larger variety of the information material contained in the sentences, greater reliability of the classification of the crime categories dealt by the sentences, greater versatility of the information to the cognitive requirements of the researcher rather than that of the administrative monitoring requirements of the judiciary machine.

We use all of these advantages and report here on our results of information extraction and social network analysis between actors (pairs of actors that repeatedly co-occur in the text). An outcome of our result is the study of a social network with several familiar properties, which suggests that the combination of information extraction and social network analysis may provide novel insights on crime as a social phenomenon.

In this paper we describe the first example in Italy of a large digital archive of criminal sentences and demonstrate how a computational structural analysis can be performed on the *corpus* of data.

1.1. Criminal Sentences from a Sociological Perspective

In the analysis of the jurisprudence a sentence is the result of a reconstruction process of reality that depends on the purpose for which such reconstruction is made (that is to judge), and on the (juridic) technique used in order to arrive to the qualification of the facts.

The judge must evaluate using the current law, and therefore his action is limited to the possibilities contemplated by the law. This inevitably implies, on the one hand, that the qualificatory process influences the way that the reality is reconstructed in the sentence and, on the other hand, that the outcome of the sentence will depend upon what the judge will deem significant for the application of the relevant norm to the concrete case. Moreover,

difficulties stem because, on the one hand, juridical constructs live on infrequency, that is, they are not ascribable to dogmatic constructs in which one can make out methodologies and objectives, and on the other hand, they are the result of a juridical culture that uses proximity models followed by the judge.

Until now, the reflections on the jurisprudence as a topic of research, with an explanatory focus of the interplay between juridical and social dimensions in the judiciary institutions are limited, especially in Italy, both in the juridic doctrine and in the philosophic and sociologic analysis. Little is known, since these analyses suffer from a lack of empirical studies. Considered as a meeting point of different philosophies coming from different interests coming from different interests in the social system, juridical decisions can represent a significant magnifying glass of mechanisms and phenomena that, although still considered within the "area" of the process, are a precious source of information for those which are motivated by heuristic needs: the juridic decision, "because of its attitude to become an immediate tool of social regulation, can form an indicator [...] of the real balances of power [...] existing in a society"[2]. The construction of the juridical reality as expression of the complexity of a social phenomenon that comprises, or at least, interacts with the juridical phenomenon is far from being ascribable to a paradigm oriented for the understanding of basic mechanisms; it is rather mainly oriented to the construction of meta-theorization which tends to overlook the observation of the phenomena. In this context, however, the debate exhibits a consciousness raising of the unilateral perspective employed until now in the formulation of possible theories on the juridical phenomenon in its entirety (be it that of the judge, of the legislator, or of the citizen), thus considering the different roles and the different perspectives of the analysis.

From this point of view, a perspective of socio-juridical analysis may be fertile; pointing out that the scientific character of sociology of law, today, is the employment of methodological tools that allow it to understand and describe in a scientifically-founded way, on one hand, the proper function of the modern jurisprudence, and on the other hand, the empirically observable reality of the behavior of jurisprudence operators or subjects interacting with the juridical system.

[2] O. Abbamonte, *Le ragioni del decidere. Per un possibile studio della giurisprudenza e della mentalità del giudice*, in "Sociologia del diritto", Vol. 28, 2001, n. 2, pp. 5-44.

The incremental development of the analysis of sentences has mainly considered the areas of civil jurisprudence and the case law of the Court of Cassation (more than in the penal area), favoring an analytic-classificatory approach. The techniques of analysis are the classic ones of the traditional dogmatic investigation (given a sentence, analyze citations as well as massive basic comments from jurisdictional authorities to see what differs from the regularity of many previous pronounced sentences) applied to the study of the juridic material. Also, the common methodology of investigation is suited to the formation and cultural interests of the jurists that use it. Moreover, the examination of the sentences and of the tendencies of the jurisprudence is almost always inside the source, that is analyzed in its structural and argumentative components, with notably lucid results for knowing what and how the judge has pronounced and how much reliable and reasonable his conclusions are, on the basis of the logic coherence of the arguments contained in the decision. In this sense, the decisions, which fully enter in the "material sources" of the positive jurisprudence, would constitute the written document to interpret using as a key reading the linguistic signs that compose it. According to Abbamonte[3], the limit of this kind of methodological orientation is that a significant part of the cognitive contribution that the analysis of the juridical material could supply remains unexplored. He argues over the importance of the historicization, of the juridical decisions with respect to the connections with the political, economic, social, and ideological contexts that could have caused them, in order to "have access" to the qualifying moments of the juridical source (namely that define the characteristics of the jurisprudence source), and to search in the depths of the politics that the jurisprudence pursues. In other words, the question arises as to reconsider our understanding of the law, since this vision implies a process in which the so-called "internal legal culture" moves, communicates, both to its members and to the so-called "external legal culture", a perspective of reality that it is substantially theoretical in nature. This means that this reality, so represented, is disconnected both from the social and cultural contexts in which it originated and from the facts in which it is produced, because it is a result of logical constructs that can be good for each "historical time". The problem of rethinking the way of understanding the jurisprudence arises in many ways in the debate faced by the sociology of law. With a growing attention to the trial, to the mechanisms, and to

[3] *Ibidem*.

the relation between the roles as a measure to observe the real balances of power and relational hierarchies the sentence becomes a natural ground of the possible explanatory and observing these phenomena.

1.2. Outline of This Work

Although the *corpus* of criminal sentences is limited in both time and space, the results of the analysis are significant for three main reasons: there is not yet in the literature a comparative analysis of criminal sentences pronounced at the four Sicilian courthouses or in the rest of Italy: the present one is hence at the best of our knowledge a seminal work; there is not yet a digital database collecting data on the institutional response to the phenomenon of organized crime activities in Sicily: the present work may hence be used as a starting experience toward the creation of such knowledge base; in the case of organized crime activities, the Sicilian jurisprudence *de facto* orients the Italian jurisprudence and it is hence relevant to better investigate the internal working of this activity.

Our research is organized into three main stages[4]. In the first stage, we collected the criminal sentences from the courthouses of Sicily. Since there is not yet a digital archive of criminal sentences in Sicily, all sentences had to be collected in their paper format. The paper sentences have been scanned into digital format, and then converted into plain text files by means of computer-based OCR technology. Furthermore, we have constructed a codebook, which is basically a collection of well-thought variables to be devised from the text of each criminal sentence. This is described in Section 2. and Section 3.

In the second stage, the text files were analyzed using information extraction technology, in order to extract from the text of the sentences the actors involved in the facts and the relationships between them. In particular we extracted information on the judge, the members of the court, the prosecutor, the defendants, the lawyers, and the other people involved in the sentence facts. Relationships between actors were also extracted. The information extraction has been performed by implementing opportune *finite state transducers*, automatons capable to recognize specific patterns in an input string. This is described in Section 4.

[4] D. De Felice, G. Giuffrida, G. Giura, C. Zarba, *La descrizione dei reati di criminalità organizzata e di tipo mafioso nel testo delle sentenze*, in "Quaderni di Sociologia", Vol. 54, 2010, n. 54, pp. 57-80.

In the third stage, we constructed a social network using the information extracted so far. The social network consists of the nodes that are actors and edges denote repeated co-occurrence in *corpus* texts. The resulting network was inspected in order to detect central nodes as well as communities. These nodes should be relative to pivotal character of the trials. This is described in Section 5.

We also discuss limitations as well as possible avenues of our research and conclude our paper. This is done in Section 6.

2. Data Collection

We have restricted the analysis of criminal sentences according to the following criteria:

- the sentences describe only mafia crimes: these are those encompassed by the Italian criminal procedure code, Article 51.3-bis;
- all sentences were declared final and irrevocable for at least one defendant;
- all sentences have been pronounced by Sicilian judicial authorities between January 1, 2000 and December 31, 2006.

Since in Italy there is not yet a unified digital database of criminal sentences, it was necessary to perform a complex and time-consuming data collection activity in order to gather the criminal sentences required for our analysis. The entire process took more than two man-years work. Specifically a preliminary interrogation to the Italian computerized archive RE.GE. – the general register of criminal proceedings – has been performed. This interrogation has been parametrized with the criteria above and produced a list of about 1,200 criminal sentences satisfying our query. Then a formal request to the applicable Sicilian courthouses has been made in order to gain physical access to the paper printed criminal sentences. After the authorizations were granted, we went physically to the various Sicilian courthouses in order to obtain copies of the files. Eventually we collected 1,147 sentences; we performed an extensive manual data quality verification process. Due to misleading classification in the RE.GE. archive, only 728 criminal sentences really satisfied all our criteria. For instance, about 90% of the non-pertinent criminal sentences were initially recorded in the RE.GE. as concerning illicit drug smuggling. These sentences were later assessed as concerning the less grave crime of Art. 73 of the same DPR (*Decreto del Presidente della Repubblica*, a statute law), but without this correction being made in the RE.GE.

The entire remaining set of criminal sentences is made of about 55,000 pages, and the sentence length varies from a minimum of 2 pages to a maximum of 3,268 pages. Every page has been scanned producing digital (pdf) files, and processed with computerized (OCR) technology in order to produce plain text files suitable to automatic computerized analysis.

The collected sentences can be classified according to authority, degree of judgement and proceedings format, as shown in the following tables, where the rows indicate the authority entitled, the degrees (first three for the first degree and last two for the second degree) of judgements in the Italian judicial system, while the columns indicate the proceedings format (Tab. 1).

Authority	Standard	Abbreviated	Plea	Total
GIP/GUP	0	147	117	264
Tribunale	115	2	6	123
Corte d'Assise	11	2	0	13
Corte d'Appello	89	10	62	161
Corte d'Assise d'Appello	90	4	73	167
Total	305	165	258	728

Tab. 1 – The collection of 728 sentences on which the rest of the analysis is based

3. The Codebook

While analyzing the collected criminal sentences, we specified a codebook, whose purpose is to organize the results of the analysis. More precisely, the codebook consists of a table that contains one row for each criminal sentence. The columns denote classificatory variables that describe the features of the criminal sentences that are important for the analysis[5].

The process leading to the specification of the codebook started with an initial *a priori* definition of the variables. Then, each sentence has been read by human experts and analyzed in order to fill the codebook. While the criminal sentences were analyzed, the specification of the variables of the codebook has been gradually refined. Many sentences had to be re-analyzed several times. This manual approach is satisfactory to meet the designed goals, but it definitively limits additional analysis based on different variables. The lesson learned at this stage has been, hence, that an automatic

[5] A. Bruschi, *Metodologia delle scienze sociali*, Milano, Bruno Mondadori, 1999; A. Bruschi, *Metodologia della ricerca sociale*, Roma-Bari, Laterza, 2005.

process to harvest variables from free text is imperative in order to avoid long repetitive manual processes.

At the end of the analysis, the final codebook is made of 44 variables belonging to four categories: temporal, procedural, social, and environmental.

The *temporal* dimension includes variables describing the durations of each phase of the trial process, starting from the registration to the RE.GE., including the pronouncements of the verdicts at each degree of judgement, and terminating with the final declaration of irrevocability of the sentence. Therefore, the total duration of the trial process can be ascertained. The *procedural* dimension includes variables describing the legal events occurring during the temporal dimension. These events include custody measures, proceedings formats, contested crimes, modifications and integration of contested crimes, recognition of extenuating circumstances, and the final verdict. The *social* dimension includes variables describing the occupation or profession of the defendants, as well as their social and economic conditions. The *environmental* dimension includes variables describing the geographic, political, and institutional, aspects of the events discussed in the sentence. They also identify the economic sector that is harmed by the contested crimes, and report the official quantification of the economic cost suffered because of the contested crimes.

Making observations efficiently depends on the data being relatively structured, and the variables outlined above indicates the information which is possible to collect. However, making observations from the text in general would be a cumbersome manual process due to the size of the material. We proceed to describe an attempt to automatically extract structural information from the *corpus* of 728 sentences.

4. Information Extraction

Information extraction is generally the process of extracting structured data from "unstructured" (or explicitly less structured) observation. In this research we consider the extraction of relational data from natural language documents[6]. Typically, given a document written in natural language, there are four kinds of information that can be extracted: entities, attributes, relations, and events. Entities can be individuals, things, dates, or measurements. Attributes are features associated to entities. For instance, an indi-

[6] S. Sarawagi, *Information Extraction*, in "Foundations and Trends in Databases", Vol. 3, 2007, n. 1, pp. 261-377.

vidual has attributes like birthdate, birthplace, profession, education, title, telephone number, email address. Relations are associations between entities. Events are relations where time is of primary importance.

There are two main approaches to information extraction: deep and shallow. Deep information extraction is based on natural language processing. Information is extracted from the document by lexical analysis, semantic analysis, and the interpretation of the discourse[7]. Deep information extraction can be quite effective, but in some cases computationally too slow. Furthermore, the (manual) construction of the model necessary to carry out the interpretation of the discourse is complex and laborious. Shallow information extraction does not aim at a human-like comprehension of the document, but aims only at filling the relational tables. This is done using a pipeline consisting of a finite number of finite state transducers (FSTs). A finite state transducer takes a sequential input and, if some conditions are verified, returns an output that depends on the input and on the internal state of the transducer[8]. Essentially, a finite state transducer performs a simple linguistic task. The idea is that a finite number of simple linguistic tasks is sufficient in order to fill the relational tables.

4.1. Information Extraction from Criminal Sentences

We have implemented an automated analyzer that performs information extraction from our *corpus* of criminal sentences. Given a criminal sentence, our analyzer extracts the following entities representing individuals: judges, members of the court, defendants, lawyers, prosecutors, and other people involved. Our analyzer also extracts the crimes mentioned in the criminal sentence. Furthermore, it extracts, for each defendant, the lawyers(s) that represent them, and whether the defendant is convicted or acquitted. Finally, our analyzer extracts associations between entities representing individuals, by detecting when two distinct individuals co-occur in the same phrase of a criminal sentence. The extraction is performed by means of a pipeline of finite state transducers, and exploits the fact that Italian crimi-

[7] K. HUMPHREYS, R. GAIZAUSKAS, S. AZZAM, *Event Coreference for Information Extraction*, in "Anaresolution '97, Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts", Stroudsburg, Association for Computational Linguistics, 1997, pp. 75-81.

[8] S. BRIN, *Extracting Patterns and Relations from the World Wide Web*, in Atzeni P., Mendelzon A., Mecca G. (eds.), "The World Wide Web and Databases", Berlin-Heidelberg, Springer, 1999, pp. 172-183.

nal sentences are written following a relatively standard structure. We next describe this standard structure, and afterwards we describe the finite state transducers implemented.

4.2. Standard Structure of an Italian Criminal Sentence

An Italian criminal sentence always start with the denomination of the legal authority, and the wording "Repubblica Italiana", followed by "In nome del popolo italiano". The names of the members of the court follow. The first name mentioned is always that of the judge. Then, the other members of the court. Then another section starts, where the defendants are listed. Each defendant has to be properly identified by his/her biographical data such as birthplace and birthday. In the criminal sentence, the defendant name is typically followed by the wording "nato a" (i.e., born in). The name of each defendant is followed by the name(s) of the defending lawyer(s). The name of each lawyer is preceded by the title "avv.". The name of the prosecutor is preceded by the acronym "PM". The first name that is not preceded by "avv." or "PM", and is not followed by a "nato a" statement indicates the end of the defendant list, and this name is an involved part in the events discussed by the sentence (for instance, it could be an injured party or a witness). The verdict of the sentence is always preceded by the acronym "PQM" or "PTM". For first-degree sentences (GIP/GUP), each defendant is either convicted or acquitted. In second-degree sentences, before the acronym PQM/PTM and after the defendant list, the first-degree verdict is described. Then, after the acronym PQM/PTM, it is explained how the first-degree verdict is modified.

4.3. The Finite State Transducers

We now list and describe the finite state transducers implemented in order to perform information extraction on our *corpus* of criminal sentences.

People-FST. This transducers uses a dictionary of Italian first names and family names in order to recognize individuals. If necessary, the user can extend the dictionary. An individual is considered as a sequence of at least two names, or as a capital letter followed by a point, a space, and a name.

Defendants-FST. Each defendant is always accompanied by its birthplace and birthday. If an individual is followed by the wording like "nato a", then we assume that he/she is a defendant.

Lawyers-FST. Lawyers are individuals preceded by their title "avv.".

Judge-FST. If at this point the first individual appearing in the text of the sentence is not a defendant, then it must be the judge. If the first individual is a defendant, then the information about the judge is unavailable.

Court-FST. If the name of the judge has been extracted, then all individuals comprised between the judge and the first defendant must be members of the court.

Prosecutor-FST. The prosecutor is an individual preceded by the abbreviation "PM".

Other-FST. At this point, all individuals that are not the judge, members of the court, assistants, defendants, or lawyers, are categorized as "other people involved".

Defendants-lawyers-FST. This transducer associates each defendant to the list of lawyers that represent him/her.

Crimes-FST. This transducer recognizes the crimes disputed in the trial using regular expressions.

Verdict-FST. This transducer attempts to deduce if a defendant has been condemned or absolved. This is done by analyzing the text of the sentence following the acronym "PQM" or "PTM", and looking for words such as "condanna" or "assolve" written before the name of the defendant.

Associations-FST. This transducer detects when two individuals co-occur in the same phrase anywhere in a criminal sentence. When this happens, we say that there is an association between the two individuals. These associations are then used in order to construct a graph, which will be then analyzed using techniques of social network analysis, as described in the next section.

5. Social Network Analysis

We now turn to reporting the social network analysis we performed on the *corpus* of criminal sentences. Social network analysis uses the mathematical theory of graphs, in order to define and quantify various measures of network structure. This leads to the possibility to explain and describe sociological concepts such as social capital by describing the role of individuals and groups in social networks. The various quantitative concepts measure both properties on the network, taken as a whole, as well as properties of

the individual and her position in the network. For an introduction to the field, see e.g. Scott's book[9].

More precisely, we have analyzed a network graph generated using the information extracted using our computer-based tools. The nodes are the individuals and their relationship, their co-occurrence in sentences, extracted with the set of finite-state transducers (as described in Section 4.3.).

5.1. A First Look: Pre-processing the Data

Presented with a *corpus* of many associations, with the possibility of two names co-occurring only a few times, one must make a distinction between noise and interesting data. In order to reduce the possibility to observe occurrences by chance, we performed pre-processing of data before the analysis in three main steps. First, we eliminate self-cycles, i.e. a name occurring with itself in phrases. It is straightforward that a person co-occurring with itself is uninteresting for studying the social relations between actors. Second, we need to distinguish between a casual co-occurrence of names and an unambiguous indication in the text of an actual relationship. At this point, we chose to discard co-occurrences between pairs of nodes that co-occur less than 5 times in the *corpus*. Out of 67,586 extracted edges (co-occurring associations between two distinct individuals) from the *corpus*, discarding weights less than 5 led us to discard 49,498 edges (73%) of the available associations. Third, we chose to discard nodes representing names that had no co-occurrences to other names. Out of 5,280 names extracted out of the *corpus*, this led us to discard 2,990 (57%) nodes from the network.

5.2. Overview of the Network

After the pre-processing of the data, what remains is a network of 2,290 persons with 18,088 associations (occurring at least 5 times in the same phrase). In mathematical terms, we are thus studying a graph with weighted and undirected edges, with weights being at least 5 and relationships going in both directions between nodes. Furthermore, all nodes lack self-cycles and are connected to at least another node, thus having a degree of at least 1. The components of this network are visualized in Fig. 1, from which one can see that the network consists of many small but only one "giant" component (1,497 persons).

[9] J. SCOTT, *Social Network Analysis: A Handbook*, London, SAGE Publications, 2000.

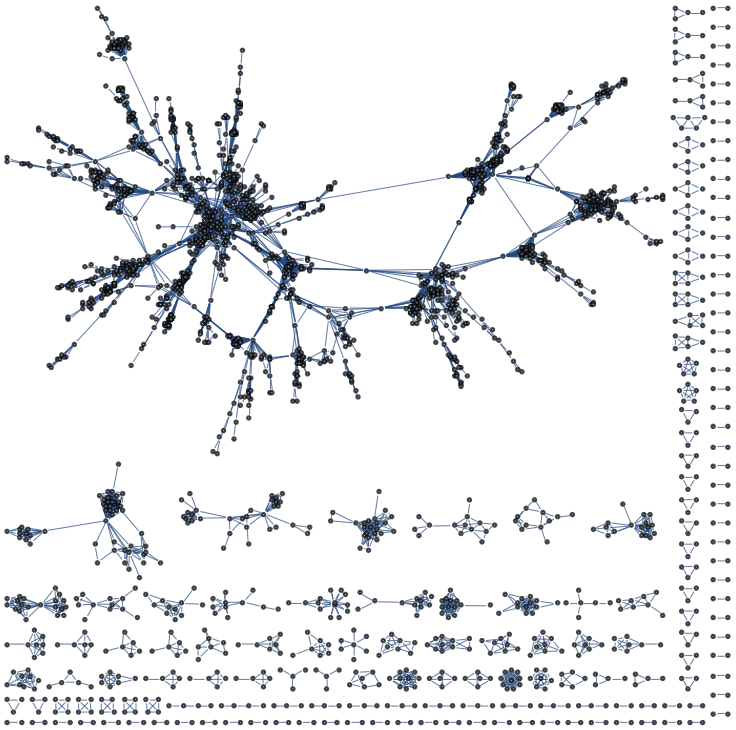

Fig. 1 – The various components in 2,290 out of 5,280 persons in the data set (the remaining 2990 nodes are isolated). The giant component contains 1,497 *persons, followed by sizes of* 52, 32, 23 *and* 20 *(the* 5 *largest components, in ascending order)*

The quantitative properties of these components, as well as the giant component separately, are summarized in Tab. 2 and Tab. 3, as well as the degree distributions in Fig. 2. We note, first, that a qualitative inspection of the network directly suggests that smaller components are possible to analyze by hand. Second, we also note that there appears to be a clear community structure with sub-groups of nodes where there are dense connections among nodes compared to that to the rest of the nodes in the graph. Third, individual nodes appear to play a significant role in linking up groups into the connectivity of the giant component. These observations lead us to analyze the network both from the perspective of groups (communities) as well as individuals.

All components (N=2290)	
Average degree	7.89
Density	0.003
Diameter	∞

Tab. 2 – Basic statistics for the full network

Giant component (N=1497)	
Average Degree	9.88
Density	0.006
Diameter	16

Tab. 3 – Basic statistics for the single giant component

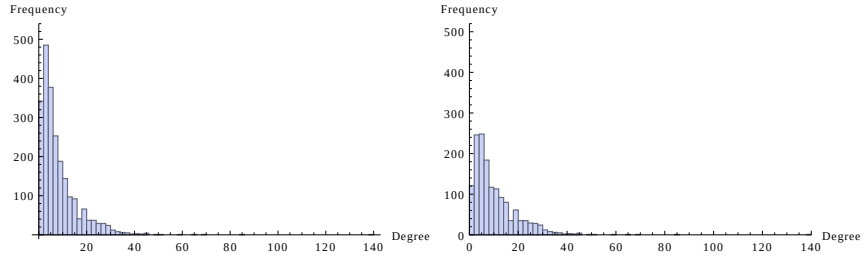

Fig. 2 – Degree distributions of the full network (N=2,290 persons, to the left) and the giant component (N=1,497, to the right). An upward shift of the average degree can be expected, as the smaller components are degree-bounded by available number of nodes

For the rest of the analysis, we will focus on the single giant component.

5.3. A Central Core in the Giant Component

Is there a dense core in the network, where everyone is associated in some way to everyone? To study this question, but with the more reasonable criterion that "many know many" in such a group, we apply the *k*-core algorithm[10]. This method iteratively eliminates nodes with a low degree, similar to removing "outliers". An iteration leads to the removal of nodes with a

[10] M. NEWMAN, *Networks: An Introduction*, Oxford, Oxford University Press, 2009.

degree less than k, and the method is repeatedly applied to the graph until no more elimination of nodes (and their associated edges) can be done. The result is presented in Fig. 3.

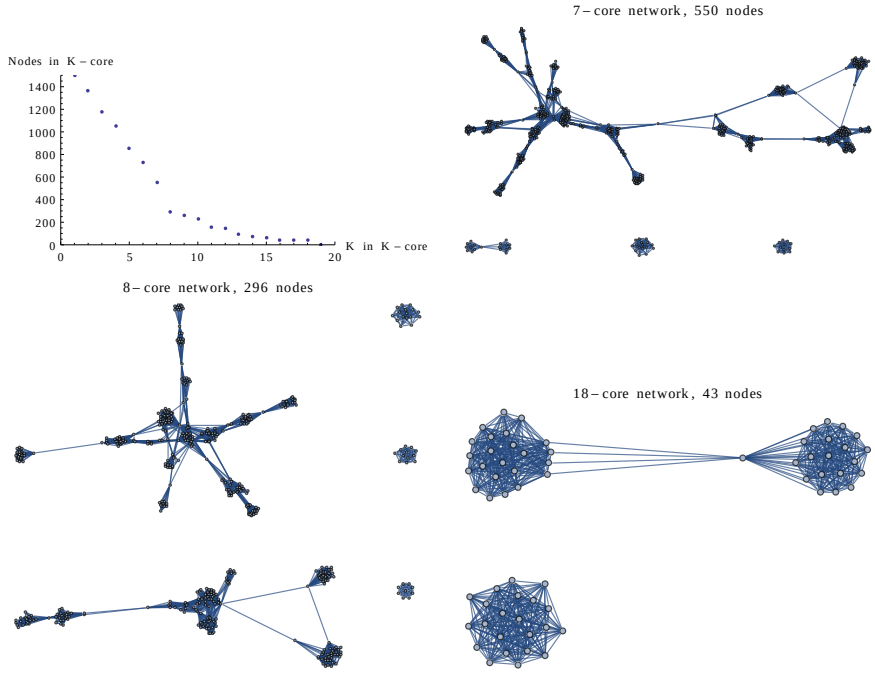

Fig. 3 – Application of the K-core algorithm to the giant component, varying k. The most dense core has at least 18 *in-group associations, and is a set of 43 nodes*

5.4. Community Structure

We proceed to perform a community structure analysis on the giant component with $N = 1,497$. Community detection[11] consists of dividing the nodes of a network into distinct, or possibly overlapping, groups of nodes that are relatively dense to the rest of the network. As argued above, smaller components also indicate a natural idea of community structure (that of be-

[11] A. CLAUSET, M.E.J. NEWMAN, C. MOORE, *Finding Community Structure in Very Large Networks*, in "Physical Review E", Vol. 70, 2004, n. 6.

ing densely connected as well as relatively isolated), but on the level of the network they are less ambiguous when it comes to visual inspection.

Focusing on the main component, we use the *modularity* metric[12], a quantitative description aiming to capture the intuition of communities as being both dense (nodes in a community have most neighbors within that community) as well as relatively isolated (there are few edges leaving the community). The modularity of a graph takes its value between -1 and 1. We execute a fast greedy algorithm that attempts to maximize modularity[13] by splitting the network into communities. In principle, modularity can be maximized in various ways[14], where we use a direct way by iteratively joining single nodes into bigger groups and evaluate what maximizes modularity change. This results in 33 different communities and a modularity value of 0.884, which is relatively high (as compared to other networks). While getting high modularity scores can be consistent with several different divisions into communities[15], with different solutions possibly giving similar modularity values, a high modularity demonstrates that the network structure is relatively easily split into communities. The resulting network contains 7,028 out of 7,400 edges *inside* the communities, i.e. only 5% of the relations are directed outside the community to which an individual is classified. The resulting communities are displayed in Fig. 4, and their size in Fig. 5.

5.5. Important Nodes in the Network

The previous results on community structure suggests the possibility to study nodes that appear important for the overall network connectivity. Due to the high modularity value, with most edges inside the communities, we can expect a relatively small number of nodes to act as "bridges" between different communities in the network. In order to assess this quan-

[12] M. NEWMAN, *Modularity and Community Structure in Networks*, in "Proceedings of the National Academy of Sciences", Vol. 103, 2006, pp. 8577-8582.

[13] U. BRANDES, D. DELLING, M. GAERTLER, R. GÖRKE, M. HOEFER, Z. NIKOLOSKI, D. WAGNER, *On Finding Graph Clusterings with Maximum Modularity*, in Brandstädt A., Kratsch D., Müller H. (eds.), "Graph-Theoretic Concepts in Computer Science", Berlin-Heidelberg, Springer, 2007, pp. 121-132.

[14] M. NEWMAN, *Networks: An Introduction*, cit.

[15] B.H. GOOD, Y.-A. MONTJOYE, A. CLAUSET, *Performance of Modularity Maximization in Practical Contexts*, in "Physical Review E", Vol. 81, 2010, n. 4, p. 19.

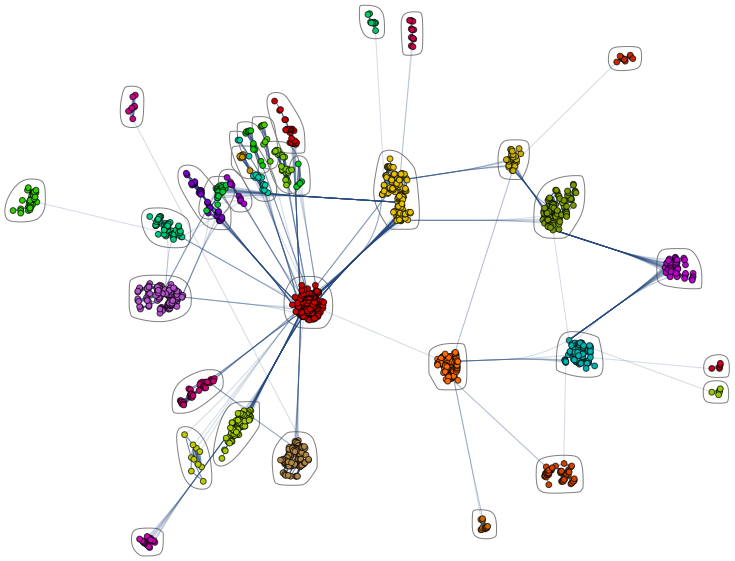

Fig. 4 – Community detection applied to the giant component. Using modularity maximization, the method suggests 33 *distinct communities with a high modularity value of* 0.884. 95% *of the edges stay within the classified communities*

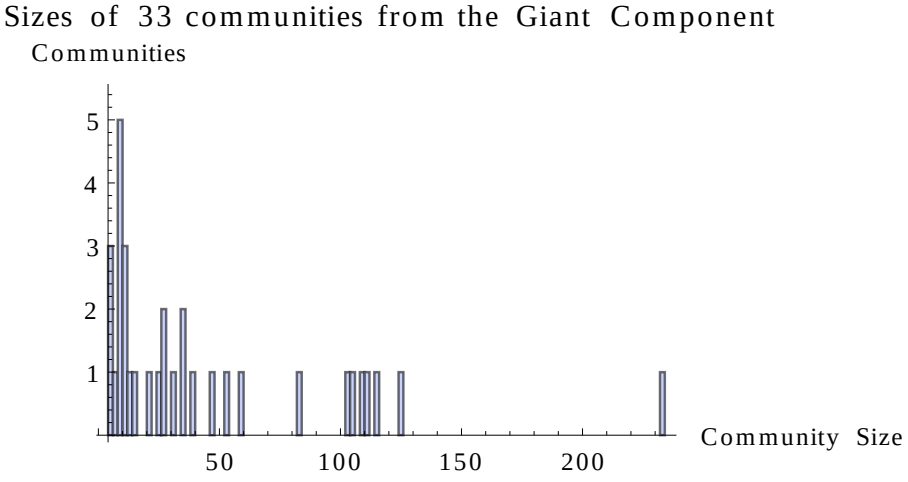

Fig. 5 – Distribution of the 33 community sizes of the giant component

titatively, we calculate the *betweenness centrality*[16] for the nodes in the giant component. The result can be found in Fig. 6. We find that this measure varies widely, which is consistent with the findings above.

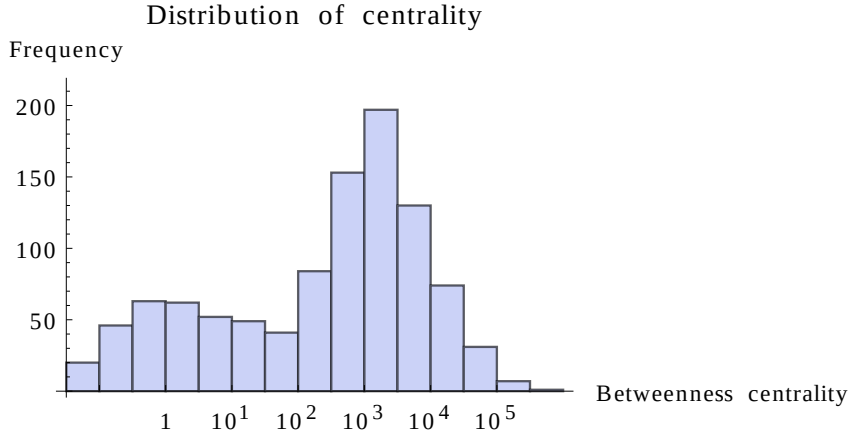

Fig. 6 – Distribution of betweenness centrality measure in the giant component. The values take a wide range, showing how nodes vary to a big extent in the connectivity of the giant component

6. Conclusion

In this paper we have described the first example in Italy of a large digital archive of criminal sentences accompanied by a structural analysis of response to organized crime. We set out to describe the socio-economic environment of criminal sentences by using automatic tools of information extraction. We extract information for social network analysis of the different actors in the jurisdictional system (judges, members of the court, defendants, lawyers, prosecutors, etc.). Our results are very reasonable in terms of what is typically found in a social network: some highly connected nodes, a giant component, community structure as well as varying dependence on the connectivity among individual nodes[17].

It is also possible to keep important assumptions and limitations in mind. The structure of this network is gathered from a specific representation of

[16] Betweenness centrality is a standard metric, attempting to quantify the intuitive notion that there can be "bottlenecks" or "bridges" in social networks that connect separate subgraphs.

[17] M. Newman, *Networks: An Introduction*, cit.

reality, created by the jurisdictional system in terms of written sentences. Specifically, the concept of relating actors by repeated co-occurrence in text depends on a number of decisions by a number of people in the jurisdictional system. In this research we have worked on a particular representation when it comes to understanding a network. The problem of the representation of the structured data is particularly important and is not completely defined, since it is influenced by the kind of analysis that one may decide to apply to the *corpus* of data. In any case, whichever will be its final form, the current activity of entity and relation extraction is necessary to avoid too large manual efforts. The sociological approach and the computational approach both have virtues and limitations, as well as the risks originating from the potentially increasing level of abstraction that one may achieve using more sophisticated computer models. The application of computer science gets its power from the large-scale capability of symbolic manipulation of the information, upon which it exercises its capabilities using imposed models, in need of simplified assumptions rather than real correspondence between what is real and what is represented. Obviously, while providing the power of simplicity, only studying text co-occurrence cancels out many other possible relations that could be read out manually from the text. The task of the social scientist is often exactly the opposite than finding restrictive assumptions: to highlight the differences, i.e. of the relations, because only through understanding these crucial differences, it is possible to, if not explain, then at least fully describe the phenomena that are the subjects of study. This task is easier when the information is structured, and the structure, when a large *corpus* of data is present, can also use the contribution of applied computer science. This fact is an additional reason that makes the project of interest to both the scientific communities.

The main challenges of additional research will be to overcome the difficulty to perform information extraction on the *corpus* of available sentences and classify in a sociologically correct way the different nodes we have found out using social network analysis. At another level, challenges also exist for the OCR software, with e.g. some of the sentences containing not only typed text, but also some handwriting, and the OCR is unable to satisfactorily process the handwriting. Future progress of this research will therefore need to use more sophisticated OCR tools, specifically tailored to the kind of criminal sentences that need to be analyzed. In general, there remain a number potentially fruitful ways to apply methods from network analysis and improve our knowledge about this system.