Communities in criminal networks: A case study

Francesco Calderoni[1]

Università Cattolica del Sacro Cuore and Transcrime, Milan, Italy

Domenico Brunetto

MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy

Carlo Piccardi[1]

DEIB, Politecnico di Milano, Milan, Italy

Abstract

Criminal organizations tend to be clustered to reduce risks of detection and information leaks. Yet, the literature has so far neglected to explore the relevance of subgroups for their internal structure. The paper applies methods of community analysis to explore the structure of a criminal network representing the individuals' co-participation in meetings. It draws from a case study on a large law enforcement operation ("Operazione Infinito") tackling the 'Ndrangheta, a mafia organization from Calabria, a southern Italian region. The results show that the network is significantly clustered and that communities are partially associated with the internal organization of the 'Ndrangheta into different "locali" (similar to mafia families). Furthermore, community analysis methods can effectively predict the leadership roles (above 90% precision in classifying nodes as either bosses or non-bosses) and the locale membership of the criminals (up to two thirds of any random sample of nodes). The implications of these findings on the interpretation of the structure and functioning of the criminal network are discussed.

[1] Correspondence and requests for materials should be addressed to F.C. (email: francesco.calderoni@unicatt.it) or C.P. (email: carlo.piccardi@polimi.it).

Keywords: Criminal networks; Community analysis; Centrality measures; Leadership identification; Membership identification

1. Introduction

Academics and law enforcement agencies are increasingly applying network analysis to organized crime networks. While the current applications mainly focus on the identification of the key criminals through centrality measures (Varese, 2006b; Morselli, 2009; Calderoni, 2014) and other individual attributes (Carley et al., 2002; Morselli and Roy, 2008; Malm and Bichler, 2011; Bright et al., 2015), the analysis of the subgroups and their influence on the criminal activities received very limited attention so far.

Subgroups are a natural occurrence in criminal networks. Criminal organizations may structure themselves in functional, ethnic, or hierarchical units. Furthermore, the constraints of illegality limit information sharing to prevent leaks and detection, as criminal groups face a specific efficiency vs. security trade-off (Morselli et al., 2007). This makes criminal organizations globally sparse but locally clustered networks, often showing both scale-free and small-world properties (Malm and Bichler, 2011). Also, the larger the criminal organization, the most likely and relevant is the presence of subgroups. These considerations suggest that the analysis of subgroups in criminal networks may provide insight on both the internal structure of large organized crime groups and on the best preventing and repressive strategies against them.

The mafias are a clear example of large organized crime groups, often comprising several families or clans with a specific hierarchy and a strong cohesion. These units may show different interactions among them, ranging from open conflict to pacific cooperation. Each mafia family is a subgroup within a larger criminal network, and inter-family dynamics are determinant for the activities of the mafias. Nevertheless, possibly due to the difficulties in gathering reliable data, the literature has so far neglected the role of the family in the structure and the activities of the mafias.

In the literature of network analysis (e.g., Boccaletti et al., 2006; Barrat et al., 2008; Newman, 2010), one of the most challenging areas of investigation in recent years is *community analysis*, which is aimed at revealing possible subnetworks (i.e., groups of nodes called communities, or clusters, or modules) characterized by comparatively large internal connectivity, namely whose nodes tend to connect much more with the other nodes of the group than with the rest of the network. A large number of contributions have explored the theoretical aspects of community analysis and proposed a broad set of algorithms for community detection (Fortunato, 2010). Most notably, community analysis has revealed to be a powerful tool for deeply understanding the properties of a number of real-world complex systems in virtually any field of science, including biology (Jonsson et al., 2006), ecology (Krause et al., 2003), economics (Piccardi et al., 2010), information (Flake et al., 2002; Fortuna et al., 2011) and social sciences (Girvan and Newman, 2002; Arenas et al., 2004).

This paper aims to apply the methods of community analysis to criminal networks analyzing the co-participation in the meetings of a large mafia organization. The exercise aims to explore the relevance of subgroups in criminal networks, with a specific focus on the characterization of mafia clans and families and the identification of bosses. The case study draws data from a large law enforcement operation in Italy ("Operazione Infinito"), which arrested more than 150 people and concerned the establishment of several 'Ndrangheta (a mafia from Calabria, a southern Italian region) groups in the area around Milan, the capital city of the Lombardy region and Italy's "economic capital" and second largest city. The exploration has a double relevance. First, it improves the understanding of the internal functioning of criminal organizations, demonstrating that the Infinito network is clustered in subgroups, and showing that the subgroups identified by community analysis overlap with the internal organization of the 'Ndrangheta. Second, it may contribute in the development of law enforcement intelligence capacities, providing tools for early identification of the internal structure of a criminal group.

The internal organization of the 'Ndrangheta provides an interesting opportunity to explore the relevance of subgroups in criminal networks. Indeed, this mafia revolves around the blood family (Paoli, 2003; Varese, 2006a). One or several 'Ndrangheta families, frequently connected by marriages, godfathering and similar social ties, form a "'ndrina". The "'ndrine" from the same area may form a "locale", which controls a specific territory (Paoli, 2007). The "locale" is the main structural unit of the 'Ndrangheta. Each "locale" has a number of formal charges, tasked with specific functions: the boss of the "locale" is the "capobastone" or "capolocale", the "contabile" (accountant) is responsible for the common fund of the locale, the "crimine" (crime) oversees violent actions, and the "mastro di giornata" (literally "master of the day") takes care of the communication flows within the "locale".

Since the organization in "locali" plays such an important role in the structure of the 'Ndrangheta, our investigation is specifically oriented to assess their significance in the sense of community analysis. Therefore, after illustrating some details on the network data (Sec. 2), we first quantify the cohesiveness of each "locale" in the Infinito network, discovering a quite diversified picture where very cohesive "locali" coexist with others apparently not so significant. The results of community analysis (Sec. 3) show that the Infinito network is significantly clustered, suggesting that subgroups play an important role in its internal organization. If we try and match the clusters obtained by community analysis with the "locali" composition, we interestingly discover that in most cases clusters correspond either to "locali" or to unions of them. In the last two sections, we use the results from community analysis to identify the bosses and the locale membership of each network participant. Section 4 shows that community analysis can effectively identify the bosses in the Infinito network, yielding up to approximately 93% of correct predictions (nearly 60% for bosses only). Section 5 demonstrates the utility of community analysis in identifying the "locale" membership of the nodes. The best method correctly attributes the "locale" of up to 65% of a random sample of nodes.

2. The Infinito network

"Operazione Infinito" was aimed at disentangling the organizational structure of the 'Ndrangheta in Lombardy, with a special care in charting the hierarchical structure and the different "locali" existing in the region. The documentation[2] provides information on a large number of meetings among members. Indeed, most of the investigation focused on meetings occurring in private (e.g., houses, cars) or public places (e.g. bars, restaurants or parks). The two sets, namely meetings and participants, define a standard bipartite (two-mode) network. The projection of the bipartite network onto the set of 256 participants leads to a (one-mode) weighted, undirected network, whose largest connected component – which we will denote hereafter as the *Infinito network* – has $N = 254$ nodes and $L = 2132$ links (the density is $\rho = 2L/(N(N-1)) = 0.066$). The weight w_{ij} is the number of meetings co-participation between nodes i and j, and it ranges from 1 to 115. However, the mean value of the (nonzero) weights is $\langle w_{ij} \rangle = 1.88$ and about 70% of them is 1, denoting that only very few pairs of individuals co-attended a large number of meetings. Similarly, the distributions of the nodes degree k_i and strength $s_i = \sum_j w_{ij}$ display a quite strong heterogeneity: indeed, their average values are, respectively, $\langle k_i \rangle = 16.8$ and $\langle s_i \rangle = 31.5$, but the most represented individual in the sample has both degree and strength equal to 1.

The affiliation of an individual to the "locale", namely the group controlling the criminal activities in a specific territory, is formal and follows strict traditional rules. Each "locale" has a boss who is responsible of all the activities in front of the higher hierarchical levels (see Calderoni (2014) for further details). The investigation activity of "Operazione Infinito" was able to associate 177 individuals (out of 254) to one of the 17 "locali" identified in Milan area, the region under investigation. Of the remaining ones, 35 were known to belong to

[2] Pretrial detention order issued by the preliminary investigation judge upon request by the prosecution (Tribunale di Milano, 2011).

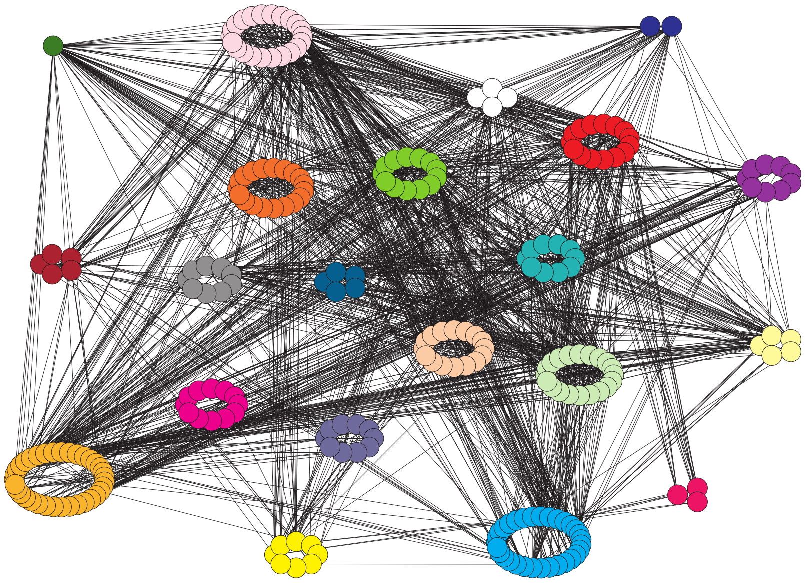

Figure 1: The Infinito network: nodes are grouped and colored according to the "locali" partition (Table 1).

"locali" based in Calabria (the region of Southern Italy where the 'Ndrangheta had origin and still has its headquarters), 3 came from a Lombardy "locale" not in the area of investigation (Brescia), and 8 were known to be non affiliated to 'Ndrangheta, whereas the correct classification of the remaining 31 individuals remained undefined. The Infinito network is displayed in Fig. 1. The figure[3] puts in evidence the 17 "locali" and the other groups above described.

As a first analysis, we assess whether the partition defined by the "locale" membership is significant in the sense of community analysis, namely whether the intensity of intra-"locale" meetings is significantly larger than that of the contacts among members of different "locali". If so, this would confirm, on one hand, the actual modular structure of the crime organization; on the other hand, it would provide a tool for investigations, as the composition of the "locali" could

[3] All network figures in the paper were produced with Pajek (Batagelj and Mrvar, 2004).

endogenously be derived by mining meetings data.

We denote by C_k the subgraph induced by the nodes belonging to "locale" k. We quantify the cohesiveness of C_k by the *persistence probability* α_k, namely the probability that a random walker, which is in one of the nodes of C_k, remains in C_k at the next step. This quantity, which proved to be an effective tool for mesoscale network analysis (Piccardi, 2011; Della Rossa et al., 2013), reduces in an undirected network to:

$$\alpha_k = \frac{\sum_{i \in C_k} \sum_{j \in C_k} w_{ij}}{\sum_{i \in C_k} \sum_{j \in \{1,2,\ldots,N\}} w_{ij}}, \qquad (1)$$

namely to the fraction of the strength of the nodes of C_k that remains within C_k (the same quantity is referred to as *embeddedness* by some authors (e.g., Hric et al., 2014)). Radicchi et al. (2004) defined *community* a subnetwork which has $\alpha_k > 0.5$. Obviously, the larger α_k, the larger is the internal cohesiveness of C_k. Notice that, since α_k tends to grow with the size N_k of C_k (trivially, $\alpha_k = 1$ for the entire network), large α_k values must be checked for their statistical significance. We derive the empirical distribution of the persistence probabilities $\bar{\alpha}_k$ of the connected subgraphs of size N_k (we do that by randomly extracting 1000 samples), and we quantify the significance of α_k by the *z-score*:

$$z_k = \frac{\alpha_k - \mu(\bar{\alpha}_k)}{\sigma(\bar{\alpha}_k)}. \qquad (2)$$

A large value of α_k (i.e., $\alpha_k > 0.5$) reveals the strong cohesiveness of the subgraph C_k, while a large value of z_k (i.e., $z_k > 3$) denotes that such a cohesiveness is not trivially due to the size of the subgraph, but it is anomalously large with respect to the subgraphs of the same size.

Table 1 summarizes the values of α_k and z_k computed on the subgraphs corresponding to the "locali" (see Fig. 1). Notice that L2 to L18 actually refer to the 17 "locali" under investigation, all based in Milan area (Milan itself plus 16 small-medium towns); L19 collects the individuals, participating in some of the meetings, belonging to any of the Calabria "locali", and L20 contains those affiliated to Brescia, not subject to investigation and whose members participated in the meetings only occasionally; L0 are the individuals with non

	“locale”	N_k	α_k	z_k
L0	*not specified*	*31*	*0.08*	*-3.15*
L1	*not affiliated*	*8*	*0.03*	*-0.84*
L2	Bollate	13	0.25	1.31
L3	Bresso	15	0.39	2.72
L4	Canzo	2	0.10	0.47
L5	Cormano	22	0.41	3.96
L6	Corsico	4	0.12	0.21
L7	**Desio**	**19**	**0.63**	**6.40**
L8	Erba	9	0.37	2.44
L9	**Giussano**	**10**	**0.63**	**5.26**
L10	Legnano	10	0.20	0.77
L11	Limbiate	1	0	-
L12	Mariano Comense	9	0.27	1.40
L13	**Milano**	**16**	**0.62**	**5.78**
L14	Pavia	5	0.13	0.25
L15	Pioltello	20	0.43	3.83
L16	Rho	5	0.18	0.78
L17	**Seregno**	**12**	**0.93**	**8.73**
L18	Solaro	5	0.06	-0.42
L19	*Calabria locali*	*35*	*0.19*	*-0.97*
L20	*Brescia*	*3*	*0.17*	*0.98*

Table 1: Testing the “locali” partition. In bold, the four “locali” with significant cohesiveness ($\alpha_k > 0.5$).

specified affiliation, L1 those who are not affiliated. Overall, only 4 “locali” out of 17 reveal strong – and statistically significant – cohesiveness, proving to actually behave as communities in the sense of network analysis. Most of the other ones, however, display very mild cohesiveness. It cannot be claimed, therefore, that the “locali” partition as a whole is significant in functional terms. In the next section, we analyze whether the network is actually organized around a different clusterization.

3. Community analysis

Given a partition $C_1, C_2, \ldots, C_K$ of the nodes of a weighted, undirected network into K subgraphs, the *modularity* Q (Newman, 2006; Arenas et al., 2007) is given by

$$Q = \frac{1}{2s} \sum_{k=1,2,\ldots,K} \sum_{i,j \in C_k} \left(w_{ij} - \frac{s_i s_j}{2s} \right), \qquad (3)$$

	N_k	α_k	z_k
C1	12	0.93	9.07
C2	18	0.72	7.79
C3	25	0.66	9.85
C4	25	0.63	9.11
C5	45	0.68	8.20
C6	62	0.78	8.30
C7	67	0.67	5.72

Table 2: Results of max-modularity community analysis

where $s = \sum_i s_i/2$ is the total link weight of the network. Modularity Q is the (normalized) difference between the total weight of links internal to the subgraphs C_k, and the expected value of such a total weight in a randomized "null network model" suitably defined (Newman, 2006). Community analysis seeks the partition with the largest Q: large values ($Q \to 1$) typically reveal a high network clusterization. Although the exact max-Q solution cannot be obtained because computationally unfeasible even for small-size networks (Fortunato, 2010), many reliable sub-optimal algorithms are available: here we use the so-called "Louvain method" (Blondel et al., 2008).

The result is a partition with 7 clusters ($Q = 0.48$), whose data are reported in Table 2. All clusters are strongly cohesive (α_k much larger than 0.5, with large z_k). Overall, the Infinito network is therefore strongly clusterized, with community size from small (12) to medium-large (67, about 26% of the network size).

The max-modularity partition of the Infinito network is displayed in Fig. 2. The patterns of node colors – which refer to the "locali", see Fig. 1 – denote a non trivial relationship between the "locali" partition and the max-modularity partition. To disentangle this aspect, we pairwise compare the "locali" $L0, L1, \ldots, L20$ (Table 1) and the communities $C1, C2, \ldots, C7$ obtained by max-modularity (Table 2), quantifying similarities by *precision* and *recall* (e.g., Baeza-Yates and Ribeiro-Neto, 1999). Let m_{hk} be the number of nodes classified both in L_h and in C_k. Then the precision $p_{hk} = m_{hk}/|C_k|$ is the fraction of the nodes of C_k that belongs to L_h whereas, dually, the recall $r_{hk} = m_{hk}/|L_h|$ is the fraction of the nodes of L_h that belongs to C_k. If we interpret L_h as the

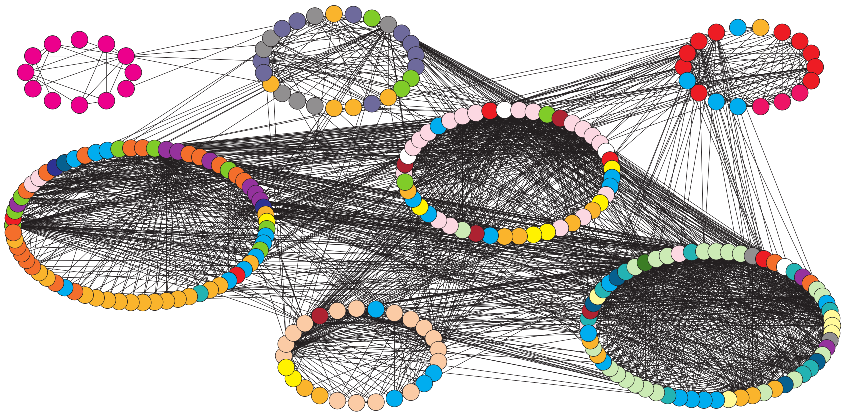

Figure 2: The Infinito network: nodes are grouped according to the max-modularity partition (Table 2) and colored according to the "locali" partition (Table 1).

"true" set and C_k as its "prediction", then the precision quantifies how many of the predicted nodes are true, and the recall how many of the true nodes are predicted. Then $p_{hk} = r_{hk} = 1$ if and only if the sets L_h and C_k coincide, while $p_{hk} \to 1$ if most of the nodes of C_k belong to L_h, and $r_{hk} \to 1$ if most of the nodes of L_h are included in C_k.

Figure 3 (upper panels) summarizes the results of this analysis by a graphical representation of the precision and recall matrices. We firstly note that "locale" L17 perfectly matches community C1 (it is the community in the upper-left corner of Fig. 2). Moreover, "locale" L13 can be approximately identified with C3, whereas C2 corresponds to a large extent to the union of L3 and L20, and C4 to the union of L9 and L12. But also the last three columns of the recall matrix clearly put in evidence that C5, C6 and C7 actually behave, to a large extent, as unions of "locali". This clearly emerges from the lower panels of Fig. 3, where the precision/recall analysis is performed again but after "locali" have been partially aggregated in 7 supersets: the diagonal dominance of the matrices p_{hk}, r_{hk} highlights that, overall, the Infinito network is quite strongly compartmentalized (see again Table 2), and the compartments coincide to a

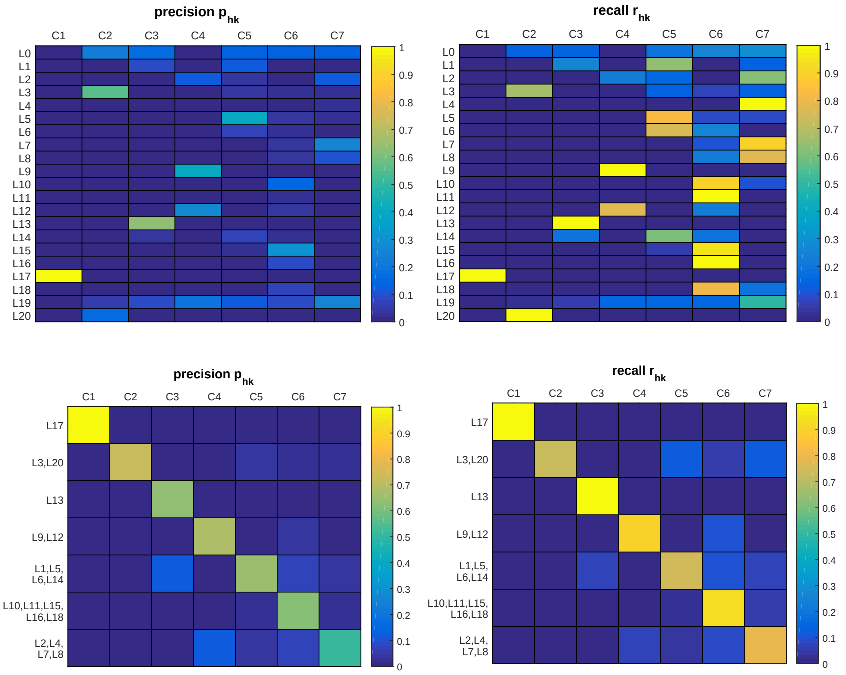

Figure 3: Precision/recall matrices of the comparison between the "locali" and the max-modularity communities. Above: the "locali" $L0, L1, \ldots, L20$ are compared with the communities $C1, C2, \ldots, C7$. Below: after "locali" have been partially aggregated, the diagonal dominance of the precision/recall matrices evidences that communities coincide to a large extent with unions of "locali".

large extent with single "locali" or unions of them.

These findings support the intuition that subgroups are important elements in the internal organizations of the mafias. The clusterization of unions of "locali" may suggest that clans or families may have closer connections with a few others. Several investigations showed that "locali" may raise and decline, compete or collaborate, merge or separate. Based on meeting co-participation patterns, community analysis methods can effectively reveal a clusterization closely connected with the formal structure of the mafia. The next two sections will explore whether community analysis techniques can further contribute to identifying the bosses and the "locale" membership.

4. Identifying bosses

In this section we focus on the relation between the hierarchical role of individuals within the 'Ndrangheta organization, and the pattern of their meeting attendance, as modeled by the Infinito network. The aim is to explore whether the results from community analysis can provide tools to identify individuals with leading roles, who will be referred to as *bosses* from now on. As already pointed out in Sec. 1, the 'Ndrangheta relies on a formal hierarchy with multiple ranks and offices. In particular, each locale normally appoints a few major officers: the capobastone or capolocale is the head of the locale; the contabile is the accountant who manages the common fund of the group; the crimine (crime) oversees violent actions; the mastro di giornata (master of the day) ensures the flow of information within the locale (Calderoni, 2014). Information on the actual number and roles of the offices in the 'Ndrangheta is incomplete. Yet, in some investigations the suspects discuss about the different offices: these conversations are sometimes tapped by the police, as in the Infinito case.

The judicial documentation classifies 34 of the 254 nodes of the Infinito network as bosses. Calderoni (2014), working on the unweighed network, investigated the correlation between a set of node centrality measures (including degree, strength, betweenness, closeness, and eigenvector centrality) and the boss role of the node, finding that betweenness is by far the most effective predictor. Indeed, the average betweenness of bosses turns out to be about 15 times larger than that of non-bosses, testifying a brokering role of bosses within the criminal network.

Here we want to further improve the predictive performance by exploiting the information provided by community analysis. As a matter of fact, the partition induced by max-modularity has the effect of placing each node in a specific position in terms of intra-/inter-community connectivity, an information that can potentially be useful in assessing its functional role.

4.1. z-P analysis

We follow the *z-P analysis* approach proposed by Guimera and Amaral (2005) (see also Guimera et al. (2005)) where, after community analysis has identified a partition into K modules, the intra- vs inter-community role of each node i is quantified by a pair of indexes (z_i, P_i). We denote by $c(i) \in \{1, 2, \ldots, K\}$ the community node i belongs to, and by $s_i^{c(i)} = \sum_{j \in c(i)} w_{ij}$ the *internal strength* of i, i.e., the strength directed towards nodes of $c(i)$. By straightforwardly extending the definitions of Guimera and Amaral (2005) to the case of weighted networks, we define the *within-community strength* as

$$z_i = \frac{s_i^{c(i)} - \mu(s_i^{c(i)})}{\sigma(s_i^{c(i)})}, \tag{4}$$

where $\mu(s_i^{c(i)})$ and $\sigma(s_i^{c(i)})$ are the mean and standard deviation of $s_i^{c(i)}$ over all nodes $i \in c(i)$, and the *participation coefficient* as

$$P_i = 1 - \sum_{c=1}^{K} \left(\frac{s_i^c}{s_i} \right)^2 , \tag{5}$$

where $s_i^c = \sum_{j \in c} w_{ij}$ is the strength of node i directed towards nodes of community c. The normalized internal strength z_i measures how strongly a node is connected within its own community. On the other hand, P_i quantifies to what extent a node tends to be uniformly connected to all communities ($P_i \to 1$) rather than only to its own community ($P_i \to 0$).

Figure 4 shows the results of the z-P analysis of the Infinito network (notice that we normalize z_i to take values in the $[0, 1]$ interval, i.e., $z_i \to (z_i - \min z_i)/(\max z_i - \min z_i)$). The figure highlights that bosses tend to concentrate on the upper-right part of the plot, namely they have both within-module strength z_i and participation coefficient P_i larger than average. As a matter of fact, the ratio between the values of the two indicators for bosses and non-bosses is 2.51 for z_i, and 2.30 for P_i. It seems, therefore, that leading individuals have a twofold characterization, namely a connectivity larger than average within their own community, and at the same time the capability of connecting to a large number of the other communities. In order to get the most effective prediction,

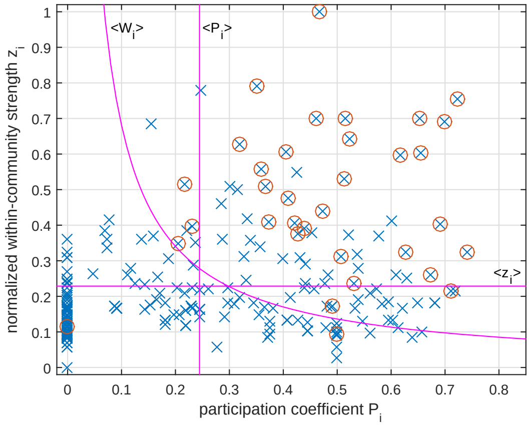

Figure 4: z-P analysis of the Infinito network. Each node is identified by a cross corresponding to the (z_i, P_i) coordinates: bosses are highlighted by red circles. The magenta lines correspond to the average value of z_i, P_i, and W_i, over all nodes.

we can combine the role of z_i and P_i in a unique indicator defined as the product $W_i = z_i P_i$. The ratio between the W_i value for bosses and non-bosses is 5.46: as evidenced in Fig. 4, only 2 bosses out of 34 have W_i lower than average.

We now want to explicitly quantify the predictive ability of the z-P analysis in identifying the leading roles within the criminal network, and compare it with a non community-based indicator such as the betweenness b_i. For that, first notice that all the indicators b_i, z_i, P_i, and W_i induce a ranking in the set of 254 nodes. Table 3 summarizes the performance of the above indicators in terms of their predictive precision, assuming to know the exact number of bosses to be guessed (i.e., 34). In other words, we count how many of the top-34 nodes in the relevant indicator's ranking are actually bosses. While the P-score alone seems unable to effectively capture the leading nodes, the z-score and the betweenness both identify 23 bosses (although the two sets are slightly different), but the zP-score outperforms all the methods identifying 25 bosses over 34.

One may wonder to what extent the above performances are influenced by the assumption of knowing exactly the number of bosses, an information not available in reality. For these reasons, we refine our analysis and compute the

method	precision
zP-score W_i	0.735
z-score z_i	0.677
betweenness b_i	0.677
P-score P_i	0.294

Table 3: Identifying bosses: for each method, the precision is computed as the fraction of true bosses among the top 34 nodes ranked by the related indicator.

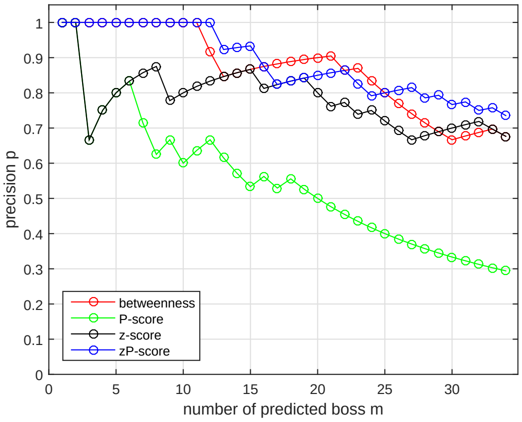

Figure 5: Identifying bosses: for a given number of predicted boss m, the precision is computed as the fraction of true bosses among the top m nodes ranked by the related indicator.

precision p for all methods as a function of the number $m = 1, 2, \dots, 34$ of guessed bosses, i.e., we take the top-m nodes for each index and we compute how many of them actually correspond to bosses:

$$p = \frac{\text{\# of nodes correctly guessed among } m \text{ nodes}}{m}. \tag{6}$$

290 The precision p as a function of m is depicted, for all methods, in Fig. 5. Overall, the zP-score has the best performance, with 100% precision up to $m = 12$ and a good performance even for the largest m values. Betweenness is a valid alternative, displaying comparable performances except for large m.

4.2. Integrating network-based measures

295 We complement the previous analysis through a set of multiple logistic regressions estimating the influence of different factors on the probability of being

	variable	min	max	mean	st.dev.		2	3	4	5	6
1	boss	0	1	0.134	0.341						
2	betweenness	0	100	4.23	11.7	2	-	.77	.73	.76	.83
3	strength	1	361	31.6	48.2	3		-	.81	.84	.88
4	z-score	0	1	0.228	0.158	4			-	.82	.67
5	zP-score	0	100	12.5	17.2	5				-	.70
6	n. of meetings	1	179	7.29	16.7	6					-
7	mafia charge	0	1	0.5	0.5						

Table 4: Descriptive statistics (left) and Pearson's correlation coefficients (right) of the variables used in the regression (all correlations are statistically significant at $p < 0.001$ level). To improve the readability of the results, betweenness and zP-score have been normalized to the $[0, 100]$ range.

a boss. This integrates and expands the analyses of Calderoni (2014, 2015), which were restricted to the individual centrality measures on the subset of meetings with more than 3 participants (215 nodes).

The dependent dichotomous variable is derived from the judicial documents (1 for bosses, and 0 for non-bosses). Independent variables include two of the network centrality measures retained in Calderoni (2014), namely the betweenness and the strength, and the z-score and zP-score from the previous subsection. The models also include two control variables: the first is the number of meetings attended by each individual, the second (mafia charge) is a dummy one describing whether an individual was charged with the offence of mafia-type association in the court order, a possible bias in the network (Table 4).

Given the low number of bosses in the sample (34 out of 254), in the logistic regressions we adopt the penalized maximum likelihood estimation proposed by Firth (1993). This method compensates for low numbers in one of the categories of the dependent variable, making it a good approach for the Infinito network. As for the standard logistic regressions, it models a dichotomous dependent variable y (in this case, the boss attribute) as a linear combination of independent variables x_i ($y = a + b_1 x_2 + b_2 x_2 + \ldots$). The outcomes can be expressed as odds ratio (OR), where $OR = \exp(b_i)$. In the present application, OR expresses the change in the probability that a node is a boss per unitary increase in any independent variable, all other variables equal. For $OR = 1$ the probability is the same, for $OR > 1$ it increases, and for $OR < 1$ it decreases. For example, $OR = 1.1$ means that a unitary increase in the independent variable implies a +10% increase in the probability of a node being a boss. Since the logistic regression predicts the value of the dependent variable based on the values of the independent variables, comparison between predicted and observed values enables to assess its predictive power (percentage of correct predictions) (Hosmer et al., 2013).

The results are summarized in Table 5. Model I replicates the best model from Calderoni (2014) on a wider sample, yielding very similar results. A unit increase in betweenness centrality provides +11% increase in the probability of being a boss, all other variables equal. The strength contributes with a +3.5% increase in probability. The model correctly classifies 94.1% of the population and 61.8% of bosses (compare with a random probability of 13.3%). Model II relies only on the control variables, mafia charge and number of meetings. Both are significant and positive. Yet the overall capacity of the model is lower than the first one (90.9%), with a remarkable decrease in the identification of bosses (41.2%). Model III includes both individual centrality measures and the controls. Both strength and betweenness maintain their significant and positive effect, whereas the controls are non-significant. The prediction success are similar to model I, especially for bosses. Models IV to VI test the community measures identified in the previous section. Model IV shows that z-score has no significant impact on the probability of being a boss, once tested along with the control variables. Conversely, in Model V the zP-score has a statistically significant and positive influence (+9.8% per unitary increase of zP-score) despite the presence of the controls. The last model (VI) includes the controls and both betweenness and zP-score. The latter results as the only significant variable with an impact of +8.6% on the probability of being a boss, all other variables equal. Overall, the share of correct predictions is slightly lower than models I and III, with the best results in model VI (92.9% and 58.8% for total correct predictions and correct boss predictions, respectively).

The regressions corroborate the results of the previous section. Network

	I	II	III	IV	V	VI
strength	1.035***	-	1.032**	-	-	-
betweenness	1.111**	-	1.108*	-	-	1.035
mafia charge	-	12.87*	4.501	8.400*	5.444	5.172
n. of meetings	-	1.167***	0.982	1.116**	1.057	1.046
z-score	-	-	-	33.34	-	-
zP-score	-	-	-	-	1.098***	1.086**
true non-bosses	218	217	217	217	216	216
false non-bosses	13	20	13	16	15	14
true bosses	21	14	21	18	19	20
false bosses	2	3	3	3	4	4
precision (total) [%]	94.1	90.9	93.7	92.5	92.5	92.9
precision (bosses) [%]	61.8	41.2	61.8	52.9	55.9	58.8

Table 5: Results of Firth's logistic regressions on bosses. The upper part of the table reports the odds ratio with the statistical significance (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$), the bottom part summarizes the predictive capabilities (percentage of correct predictions) of Models I-VI described in the text.

analysis measures can effectively predict the leadership roles of individuals in a criminal network. All network measures perform better than naturally observable variables such as the two controls. Centrality measures are effective and yield the highest share of correct predictions. Among community measures, zP-score has a significant capacity to predict bosses. In a model with centrality measures and controls, zP-score is the only statistically significant variable, indicating a strong capacity to capture the behavior of leaders in criminal networks.

These findings expand the literature on leadership in criminal networks, as previous studies mainly relied on centrality measures only, often finding that betweenness centrality identified leadership roles within crime groups (Morselli, 2009; Calderoni, 2014). Whereas the previous studies pointed out the role of brokering positions, they neglected the analysis of subgroups and its implications for leadership. The application of community analysis measures shows that criminal leaders not only have a notable brokering capacity, but also manage to balance the connection within and outside their group. These results advocate for expanding the concept of brokerage beyond individuals measures. In fact, bosses not only meet unconnected individuals, but also have a crucial function in bridging their group with other groups.

5. Identifying the "locale" membership

In this section we consider the problem of identifying the "locale" membership of those individuals for which such an information is unknown. In the Infinito network (254 nodes), this problem arises for 31 nodes (see Table 1, row L0).

The problem can be set in the general framework of *label prediction* (Zhang et al., 2010): we are given a set of network nodes $X = \{x_1, x_2, \ldots, x_{254}\}$ and a set of labels $L = \{L_1, L_2, \ldots, L_{20}\}$ which, in our case, code the "locali" of the criminal organization (Table 1). The majority of the nodes have a label: L_h is assigned to node x_i (and we write $L(x_i) = L_h$) if x_i is affiliated to "locale" L_h. The correspondence nodes/labels is, however, partially unknown, since there are 31 nodes of X whose labeling is unknown and must be predicted based on the network structure and on the known labels.

A very general approach to the above problem relies on the notion of *node similarity*, based on the assumption that the more two nodes are similar (in a sense to be defined – see below), the more likely their label is the same. Therefore, once defined a similarity score s_{ij} between nodes (x_i, x_j), the probability that the unlabeled node x_i has label L_h is assumed equal to

$$p(L(x_i) = L_h) = \frac{\sum_{\{x_j | j \neq i, L(x_j) = L_h\}} s_{ij}}{\sum_{\{x_j | j \neq i, L(x_j) \in L\}} s_{ij}}, \quad h = 1, 2, \ldots, 20. \tag{7}$$

In words, $p(L(x_i) = L_h)$ counts the relative abundance of nodes labeled L_h in the network, and weights each of these nodes by its similarity to x_i. The label predicted for node x_i is the one attaining the largest $p(L(x_i) = L_h)$.

5.1. Node similarities

We consider and test four definitions of the similarity score s_{ij}: (i) and (ii) are very popular and find many applications in social network analysis (e.g., Lü and Zhou (2011)), (iii) and (iv) exploit the partition found by max-modularity community analysis (Sec. 3).

(i) *Common Neighbors (CN):* denoting by $\Gamma(x_i)$ the set of nodes neighbors to x_i, we let

$$s_{ij} = |\Gamma(x_i) \cap \Gamma(x_j)|, \tag{8}$$

where $|Q|$ denotes the number of elements of the set Q.

(ii) *Weighted Common Neighbors (wCN):* it generalizes the above definition by exploiting the information on link weights (Lü and Zhou, 2010):

$$s_{ij} = \sum_{k \in \{\Gamma(x_i) \cap \Gamma(x_j)\}} \frac{w_{ik} + w_{kj}}{2}. \tag{9}$$

(iii) *Common Community (CC):* a binary indicator, stating that similarity is equivalent to the membership to the same community:

$$s_{ij} = \begin{cases} 1, & \text{if } c(i) = c(j), \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

where $c(i)$ denotes the community node i belongs to.

(iv) *Weighted Common Neighbors - Common Community (wCN-CC):* it combines (ii) and (iii). It is equal to the Weighted Common Neighbors similarity, but it is nonzero only when (x_i, x_j) are in the same community:

$$s_{ij} = \begin{cases} \sum_{k \in \{\Gamma(x_i) \cap \Gamma(x_j)\}} \frac{w_{ik} + w_{kj}}{2}, & \text{if } c(i) = c(j), \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

5.2. Results

The label identification procedure, with the different node similarities above defined, has been tested on the Infinito network. Unfortunately, the specificity of the case does not allow one to validate the method on the 31 nodes which are actually unlabeled – their "locale" is unknown by definition. Thus the procedure has been applied to the 177 nodes with known label $L2, L3, \ldots, L18$ (the "locali" in Milan area, the region under investigation – see Table 1), assuming their label is unknown and trying to recover it.

In order to mimic the real situation, in which an entire pool of labels have to be simultaneously identified, in our experiments we assume that the labels of m nodes have to be reconstructed at the same time, and we test the effectiveness of the procedure by letting m increasing from 1 to 30. For each m, we randomly extract 5×10^3 samples of m nodes in "locali" $L2, L3, \ldots, L18$, and predict simultaneously their labels via equation (7). For each sample, we compute the precision as the fraction of correct guesses. More in detail, for each node under test we increment a success counter s by 1 if the label which maximizes the probability (7) is the correct node label, while if the probability of $r > 1$ labels is equally maximal in (7) we increment the counter by $1/r$ if the correct node label is one of them. For the m-node sample, the precision of the reconstruction is eventually given by s/m.

Figure 6 summarizes the results, in terms of mean and standard deviation of the precision over the samples, for all $m = 1, 2, \ldots, 30$ and for the four similarity measures above defined. In principle, we expect that the larger m, the more difficult the prediction task, since the latter is based on a smaller set of known labels. In this respect, the results are rather counterintuitive. Firstly, the average precision is largely insensitive to m, and ranges from about 45% to 65% according to the similarity measure adopted. Notably, the best performing method (wCN-CC) exploits the analysis of the community structure of the network. Secondly, the variability of the precision rate displays a clear decreasing trend as m increases. This behaviour is due to a sort of "large numbers" effect: when very few labels are to be guessed, the success depends very much on the specific nodes under scrutiny. When a large pool of nodes are instead investigated, successes and failures tend to balance in a proportion which mildly depends on the specific set of nodes. Overall, this latter analysis confirms that, on the Infinito network, the precision of the label reconstruction procedure can reach a proportion of about two thirds, even for sets of the same order of magnitude of the real unlabeled set $L0$.

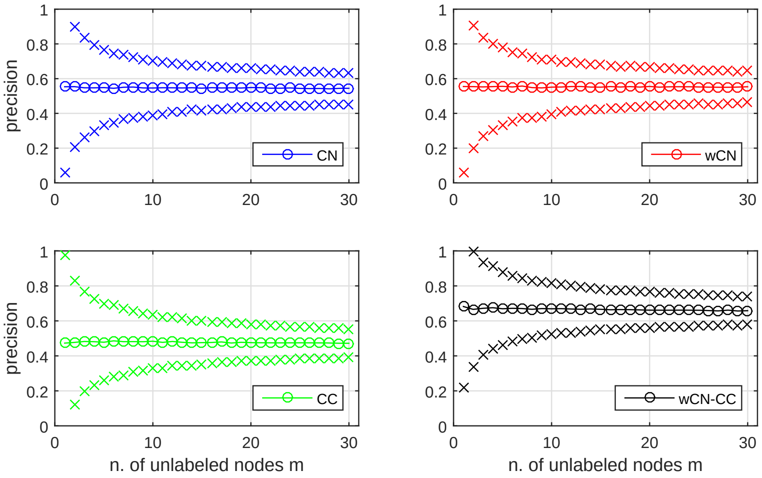

Figure 6: Precision of the label identification methods with respect to the number m of unlabeled nodes. The curves represent the average precision (circles) plus/minus standard deviation (crosses) over 5×10^3 random samples of m nodes (CN: Common Neighbors; wCN: Weighted Common Neighbors; CC: Common Community; wCN-CC: Weighted Common Neighbors - Common Community).

6. Concluding remarks

This paper applied community analysis methods to investigate the structure of a mafia organization. Focusing on meeting participation as a proxy for the relationships among criminals, community analysis assessed the clusterized structure of the mafia and showed that it often mirrors the internal subdivision of the mafia among several clans or “locali”, or unions of them. This supports the intuition that subgroups matter in this type of organizations.

In the light of these findings, the study tested the capacity of community analysis techniques to identify relevant characteristics of the criminal organization, namely leadership roles and “locali” membership. The results show that the zP-score, which captures the interplay between a node connectivity within its community and to the other communities, can effectively single out the bosses of the mafia. Furthermore, the most effective method for identifying the “locale” membership of the nodes focuses again on the connectivity within the same community.

Overall, these findings reinforce the idea that the tools of network analysis can be fruitfully adopted to enhance the understanding of the structure and function of organized crime, albeit their use as a support for law enforcement intelligence still needs further exploration.

The research can be extended in many directions. First of all, a deeper structural analysis on a pool of criminal networks would be needed, aimed at assessing whether peculiar structural attributes turn out to be recurrent in such networks. Then, coming back to the problem of community detection, other methods might prove to be more effective – including those specifically devoted to bipartite networks, as it is our data structure before projection (see Sec. 2). Finally, once the structure has been thoroughly understood, the challenge is clearly that of linking it with the function of the network, namely to fully understand how structural properties relate to criminal activities.

Acknowledgements

The authors would like to thank Giulia Berlusconi, Vera Ferluga, Nicola Parolini, Samuele Poy, and Marco Verani for many useful discussions.

References

Arenas, A., Danon, L., Diaz-Guilera, A., Gleiser, P., Guimera, R., 2004. Community analysis in social networks. European Physical Journal B 38, 373–380. doi:10.1140/epjb/e2004-00130-1.

Arenas, A., Duch, J., Fernandez, A., Gomez, S., 2007. Size reduction of complex networks preserving modularity. New Journal of Physics 9, 176. doi:10.1088/1367-2630/9/6/176.

Baeza-Yates, R., Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addison Wesley.

Barrat, A., Barthélemy, M., Vespignani, A., 2008. Dynamical Processes on Complex Networks. Cambridge University Press.

Batagelj, V., Mrvar, A., 2004. Pajek - Analysis and visualization of large networks, in: Jünger, M., and Mutzel, P (Ed.), Graph Drawing Software. Springer-Verlag Berlin. Mathematics and Visualization, pp. 77–103.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. Journal of Statistical Mechanics - Theory and Experiment , P10008. doi:10.1088/1742-5468/2008/10/P10008.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.H., 2006. Complex networks: Structure and dynamics. Physics Reports 424, 175–308. doi:10.1016/j.physrep.2005.10.009.

Bright, D.A., Greenhill, C., Reynolds, M., Ritter, A., Morselli, C., 2015. The use of actor-level attributes and centrality measures to identify key actors: A case study of an Australian drug trafficking network. Journal of Contemporary Criminal Justice 31, 262–278. doi:{10.1177/1043986214553378}.

Calderoni, F., 2014. Identifying mafia bosses from meeting attendance, in: Masys, A (Ed.), Networks and Network Analysis for Defence and Security. Springer. Lecture Notes in Social Networks, pp. 27–48.

Calderoni, F., 2015. Predicting organized crime leaders, in: Bichler, G and Malm, Aili E. (Ed.), Disrupting Criminal Networks: Network Analysis in Crime Prevention. Lynne Rienner Publishers, Boulder, CO. volume 28 of *Crime Prevention Studies*, pp. 89–110.

Carley, K.M., Krackhardt, D., Lee, J.S., 2002. Destabilizing networks. Connections 24, 79–92.

Della Rossa, F., Dercole, F., Piccardi, C., 2013. Profiling core-periphery network structure by random walkers. Scientific Reports 3, 1467. doi:10.1038/srep01467.

Firth, D., 1993. Bias reduction of maximum-likelihood-estimates. Biometrika 80, 27–38. doi:{10.1093/biomet/80.1.27}.

Flake, G., Lawrence, S., Giles, C., Coetzee, F., 2002. Self-organization and identification of web communities. Computer 35, 66–71.

Fortuna, M.A., Bonachela, J.A., Levin, S.A., 2011. Evolution of a modular software network. Proceedings of the National Academy of Sciences of the United States of America 108, 19985–19989. doi:10.1073/pnas.1115960108.

Fortunato, S., 2010. Community detection in graphs. Physics Reports 486, 75–174. doi:10.1016/j.physrep.2009.11.002.

Girvan, M., Newman, M., 2002. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America 99, 7821–7826. doi:10.1073/pnas.122653799.

Guimera, R., Amaral, L., 2005. Cartography of complex networks: modules and universal roles. Journal of Statistical Mechanics - Theory and Experiment , P02001. doi:10.1088/1742-5468/2005/02/P02001.

Guimera, R., Mossa, S., Turtschi, A., Amaral, L., 2005. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. Proceedings of the National Academy of Sciences of the United States of America 102, 7794–7799. doi:10.1073/pnas.0407994102.

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied Logistic Regression, 3rd ed. John Wiley & Sons, Hoboken, NJ.

Hric, D., Darst, R.K., Fortunato, S., 2014. Community detection in networks: Structural communities versus ground truth. Physical Review E 90, 062805. doi:10.1103/PhysRevE.90.062805.

Jonsson, P., Cavanna, T., Zicha, D., Bates, P., 2006. Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. BMC Bioinformatics 7. doi:10.1186/1471-2105-7-2.

Krause, A.E., Frank, K.A., Mason, D.M., Ulanowicz, R.E., Taylor, W.W., 2003. Compartments revealed in food-web structure. Nature 426, 282–285. doi:10.1038/nature02115.

Lü, L., Zhou, T., 2010. Link prediction in weighted networks: The role of weak ties. EPL 89. doi:10.1209/0295-5075/89/18001.

Lü, L., Zhou, T., 2011. Link prediction in complex networks: A survey. Physica A - Statistical Mechanics and its Applications 390, 1150–1170. doi:10.1016/j.physa.2010.11.027.

Malm, A., Bichler, G., 2011. Networks of collaborating criminals: Assessing the structural vulnerability of drug markets. Journal of Research in Crime and Delinquency 48, 271–297. doi:10.1177/0022427810391535.

Morselli, C., 2009. Hells angels in springtime. Trends in Organized Crime 12, 145–158. doi:10.1007/s12117-009-9065-1.

Morselli, C., Giguere, C., Petit, K., 2007. The efficiency/security trade-off in criminal networks. Social Networks 29, 143–153. doi:10.1016/j.socnet.2006.05.001.

Morselli, C., Roy, J., 2008. Brokerage qualifications in ringing operations. Criminology 46, 71–98. doi:{10.1111/j.1745-9125.2008.00103.x}.

Newman, M.E.J., 2006. Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America 103, 8577–8582. doi:10.1073/pnas.0601602103.

Newman, M.E.J., 2010. Networks: An Introduction. Oxford University Press.

Paoli, L., 2003. Mafia brotherhoods: organized crime, Italian style. Oxford University Press, New York, NY.

Paoli, L., 2007. Mafia and organised crime in Italy: The unacknowledged successes of law enforcement. West European Politics 30, 854–880. doi:10.1080/01402380701500330.

Piccardi, C., 2011. Finding and testing network communities by lumped Markov chains. PLoS One 6, e27028. doi:10.1371/journal.pone.0027028.

Piccardi, C., Calatroni, L., Bertoni, F., 2010. Communities in italian corporate networks. Physica A - Statistical Mechanics and its Applications 389, 5247–5258. doi:10.1016/j.physa.2010.06.038.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D., 2004. Defining and identifying communities in networks. Proceedings of the National Academy of Sciences of the United States of America 101, 2658–2663. doi:10.1073/pnas.0400054101.

Tribunale di Milano, 2011. Ordinanza di applicazione di misura coercitiva con mandato di cattura - art. 292 c.p.p. (Operazione Infinito). Ufficio del giudice per le indagini preliminari (in Italian).

Varese, F., 2006a. How mafias migrate: The case of the Ndrangheta in northern Italy. Law & Society Review 40, 411–444. doi:10.1111/j.1540-5893.2006.00260.x.

Varese, F., 2006b. The structure of a criminal network examined: The Russian mafia in Rome. Oxford Legal Studies Research Paper No. 21/2006 .

Zhang, Q.M., Shang, M.S., Lue, L., 2010. Similarity-based classification in partially labeled networks. International Journal of Modern Physics C 21, 813–824. doi:10.1142/S012918311001549X.