This the author's accepted version of the following article Calderoni, F., Catanese, S., De Meo, P., Ficara, A., Fiumara, G. (2020). Robust link prediction in criminal networks: A case study of the Sicilian Mafia. Expert Systems with Applications, 161, 113666. https://doi.org/10.1016/j.eswa.2020.113666

Robust Link Prediction in Criminal Networks: a case study of the Sicilian Mafia

Francesco Calderoni[a], Salvatore Catanese[b], Pasquale De Meo[c,*], Annamaria Ficara[b], Giacomo Fiumara[b]

[a]*Transcrime, Università Cattolica del Sacro Cuore, Largo Gemelli 1, 20123, Milan, Italy*
[b]*MIFT Department, University of Messina, V.le F. Stagno D'Alcontres, 31, 98166, Messina, Italy*
[c]*Department of Modern and Ancient Civilizations, Polo Universitario Annunziata, 98122, Messina, Italy*

Abstract

Link prediction exercises may prove particularly challenging with noisy and incomplete networks, such as criminal networks. Also, the link prediction effectiveness may vary across different relations within a social group. We address these issues by assessing the performance of different link prediction algorithms on a mafia organization. The analysis relies on an original dataset manually extracted from the judicial documents of operation "Montagna", conducted by the Italian law enforcement agencies against individuals affiliated with the Sicilian Mafia. To run our analysis, we extracted two networks: one including meetings and one recording telephone calls among suspects, respectively. We conducted two experiments on these networks. First, we applied several link prediction algorithms and observed that link prediction algorithms leveraging the full graph topology (such as the Katz score) provide very accurate results even on very sparse networks. Second, we carried out extensive simulations to investigate how the noisy and incomplete nature of criminal networks may affect the accuracy of link prediction algorithms. The experimental findings suggest the soundness of link predictions is relatively high provided that only a limited amount of knowledge about connections is hidden or missing, and the unobserved edges follow some kind of generative law. The different results on the meeting and telephone call networks indicate that the specific features of a network should be taken into careful consideration.

Keywords: Criminal Networks, Social Network Analysis, Network Science, Link Prediction in Uncertain Graphs.

1. Introduction

Methods from Social Network Analysis (in short, SNA) (Sparrow, 1991; Klerks, 2001; Xu & Chen, 2005; Van der Hulst, 2009; Agreste et al., 2016; Berlusconi et al., 2016) greatly contributed to intelligence and criminal investigations: for example, SNA allows to identify, within a criminal network, the most central members in terms of connections or information flow (Calderoni & Superchi, 2019), the presence of different communities (Catanese et al., 2014; Calderoni et al., 2017), and the most efficient strategies for dismantling the network (Agreste et al., 2016).

A crucial application of SNA methods to intelligence is the so-called *link prediction problem* (Liben-Nowell & Kleinberg, 2003; Pandey et al., 2019): given a graph G which describes interactions between pairs of criminals, we wish to predict which edges are more likely to appear in G in the near future. Algorithms to solve the link prediction problem may hugely impact police activities: in fact, if we would be able to accurately predict the formation of new links, we would be able to discover pairs of criminals who are likely to collaborate and, thus, we could early detect and prevent crimes.

*Corresponding author
Email addresses: francesco.calderoni@unicatt.it (Francesco Calderoni), salvocatanese@gmail.com (Salvatore Catanese), pdemeo@unime.it (Pasquale De Meo), aficara@unime.it (Annamaria Ficara), gfiumara@unime.it (Giacomo Fiumara)

Many link prediction methods have been designed and implemented in a broad range of domains (see the excellent reviews by Liben-Nowell & Kleinberg (2003) and Pandey et al. (2019)) and, more recently, Berlusconi et al. (2016) applied link prediction algorithms on a dataset derived from an Italian criminal case against a Mafia group.

Almost all of the existing approaches to link prediction focus on maximizing the accuracy and they overlook fundamental aspects such as the *robustness* of predictions, namely the extent to which the incompleteness of information about relations may affect the quality of predictions. By construction, in fact, datasets associated with criminal networks are noisy and incomplete: on one hand, investigations often encounter individuals unrelated to the criminal organization (e.g. friends, relatives, and other frequent contacts) and, on the other hand, some members of the organization actively attempt to avoid detection, e.g. by refraining from the use of telephone, using intermediaries, and coding messages. As a consequence, imprecise and incomplete information is a critical impediment to understand network boundaries and topology and, ultimately, it constitutes a main challenge for law enforcement agencies (hereafter LEAs) which plan to get reliable results from the application of link prediction algorithms.

In this paper we tackle the problem of estimating the robustness of link prediction algorithm in criminal networks. To do so, we analysed judicial sources on operation "Montagna", a long investigation on a large criminal organization belonging to Cosa Nostra (i.e., the Sicilian Mafia) (Paoli, 2004, 2008) active in the north of Sicily. We extracted two graphs (Ficara et al., 2020): the former (called *Meeting Graph* G_M) maps meetings between person under investigations and the latter (called *Phone Call Graph* G_P) is built on the monitoring of phone communications (also known as *wiretapping*). Our dataset is unique and we believe it might represent a valuable resource for better understanding complex criminal phenomena from a quantitative standpoint.

We applied many classical methods of link prediction (Liben-Nowell & Kleinberg, 2003) on both G_M and G_P such as the Common Neighbors (CN), the Jaccard Coefficient (JC), the Adamic Adar (AA) coefficient, the Preferential Attachment (PA), the Katz score, and more recent methods such as the *Node2Vec* (Grover & Leskovec, 2016) graph embedding algorithm and the Personalized PageRank (PPR) similarity score (Avrachenkov et al., 2019). We used the *Area Under the Receiving Operating Curve* (AUROC) (Fawcett, 2006) to assess the accuracy of a prediction algorithm. We recall that the AUROC ranges from 0 to 1 and the larger the AUROC, the more accurate a link prediction algorithm.

Subsequently, we assumed that both G_M and G_P graphs are not completely known and introduced some generative models to study the incomplete information about connections among criminals that we define as *uncertainty*. Specifically, we considered G_M (resp., G_P) as a sample of a true graph G'_M (resp., G'_P) and we introduced a scoring function (called *likelihood*) to decide whether an edge non observed in G_M (resp., G_P) actually exists in G'_M (resp., G'_P). A core assumption of our method is that we know all nodes of the "real" networks, but we have an incomplete knowledge of the edges of the observed networks, namely of the interconnections among criminals. Our assumptions are backed on previous research results showing that LEAs may rarely miss important individuals in a well-built criminal investigation (Campana & Varese, 2012; Berlusconi, 2013) as well as on the length and relevance of the investigation pursued in "Montagna" operation. Conversely, LEAs often need to identify the relevant ties among thousands of communications and meetings and this process may be biased by lack of resources or by criminals' strategies to prevent detection. We resort to simulation to create graphs G'_M and G'_P and, in our experimental analysis, we considered multiple likelihood functions. Our simulation method allowed also to specify the fraction p of non observed edges in G'_M (resp., G'_P) which are actually placed in G'_M (resp., G'_P).

The main findings of our analysis are as follows:

1. The Katz and PPR scores prove as the most successful method to predict missing edges and they achieve an AUROC larger than 0.95, thus signalling a high degree of accuracy. It is worth observing that the highest AUROC is achieved when only short paths are taken into account, which is consistent with the short-range structure of criminal networks.

2. Graph topology significantly affects the accuracy of a link prediction algorithm: specifically, algorithms which are very accurate on G_M performs badly on G_P and vice versa. In detail, if a graph is poorly connected (i.e., it displays a low edge density and a small clustering coefficient), then *local methods* (i.e., link prediction algorithms which rely only on the local knowledge of graph topology) are to be preferred to *global* ones (i.e., link prediction algorithms leveraging the knowledge of the full graph topology). Vice versa, global methods such as the Katz score achieve their best AUROC on graphs with greater levels of connectivity (i.e. in graph with higher edge density and clustering coefficients).

3. The knowledge LEAs have acquired about the Meeting Network G_M is quite complete whereas the Phone Call Network G_P is more susceptible to uncertainty, encoded in our study through the parameter p. We can generalize such a result: if the amount of uncertainty is relatively small and non-observed edges derive from a specified generative model we can hope for robust edge prediction. In the light of our studies, we recommend LEAs not only to build a detailed map of connections between criminals but also to investigate how such a map evolves over time: in this way we would be able to design sophisticated likelihood functions which fits fairly well experimental observations and help LEAs to early detect and prevent crimes.

The rest of this paper is organized as follows: in Section 2 we present the related literature, whereas in Section 3 we outline our research questions. In Section 4 we introduce some basic definitions from graph theory. In Section 5 we introduce the operation "Montagna" and provide some details about the Meeting and Phone Call Graphs. In Sections 6 we compare the accuracy of some link prediction methods on our graphs. We then present our experiments aimed at modelling missing edges in Section 7. We draw our conclusions and highlight some possible future work in Section 8.

2. Related Work

Social Network Analysis (SNA) is increasingly used by law enforcement agencies (LEAs) to analyze criminal networks as well as to investigate on the relations among criminals based on calls, meetings and other events derived from investigations (Sparrow, 1991; Xu & Chen, 2005; Van der Hulst, 2009; Strang, 2014).

Given the social embeddedness of organized crime and, in particular, of Mafia-like organizations, the analysis of the social structure of Sicilian Mafia syndicates generated a great scientific interest (Kleemans & Bunt, 1999; Kleemans & De Poot, 2008). For instance, Morselli (2003) studied the connections within a New York-based family (the Gambino family). The study focused on the career of one of its members, Saul Gravano. One of the main findings is that Gravano's ability of building and extending over time his personal network of contacts was a key factor to climbing the Gambino's family organization. Calderoni (2012) showed that high status Mafia members were able to indirectly manage illicit drug traffics leaving in more central and visible position middle-level criminals.

SNA is not only a tool to describe the structure and functioning of a criminal organizations but it has been largely employed in the construction of crime prevention systems (Chen et al., 2004). For instance, Xu & Chen (2005) jointly applied SNA with hierarchical clustering algorithms. The proposed approach worked in two stages: first, a criminal network was partitioned into subgroups by means of a clustering algorithm. Then, block modelling techniques have been used to extract interaction patterns between these subgroups. Agreste et al. (2016) applied percolation theory to efficiently dismantle Mafia syndicates. Calderoni & Superchi (2019) showed that the node's betweenness centrality in a meeting network is evidence of Mafia leadership, suggesting that this variable could be exploited by LEAs in selecting the most suitable targets for additional investigations and disruption. Duxbury & Haynie (2019) used an agent-based model to evaluate how criminal networks recover from disruption and identified which disruption strategies are most effective at damaging various criminal networks. Grassi et al. (2019) explored different betweenness centrality including the classic betweenness by Freeman (1979) and three inspired by the dual projection approach recently suggested by Everett & Borgatti (2013), which proved to be more successful than classic approaches in identifying the criminal leaders. Bouchard (2020) used a network approach to specify and model collaboration among people involved in organized crime. His approach provides methodological guidelines for clarifying boundaries and helps solve four puzzles: social boundaries, boundaries of group membership, ethnic boundaries and recruitment. Overall, while the studies above provided insight into the social organization and possible countermeasures against criminal groups, the application of SNA to them nearly inevitably faces problems of *noisy or incomplete information*.

Information on a criminal network is often likely to be missing or hidden, due to the covert and stealthy nature of criminal actions (Krebs, 2002; Xu & Chen, 2005). Consequently, the derived networks are incomplete, incorrect, and inconsistent, either due to deliberate deception on the part of criminals, or to limited resources or unintentional errors by LEAs (Calderoni, 2010; Campana & Varese, 2012; Catanese et al., 2014; Ferrara et al., 2014; Agreste et al., 2016). These limitations may bias the analysis and they cause problems of uncertain information, potentially jeopardizing the effectiveness of the investigations (Strang, 2014).

In the analysis of criminal networks, missing data can refer to missing *nodes* and/or missing *edges* (Calderoni, 2010). The problem of missing nodes has already received attention (Kim & Leskovec, 2011; Hric et al., 2016) but,

from our perspective, missing nodes in criminal networks are not particularly relevant, in that it is quite improbable that LEAs may disregard central criminals during such prolonged investigations. On the other hand, while it is possible to predict some missing edges among already detected criminals, it is impossible to detect missing criminals relying only on pre-trial detention orders. Missing edges refer to the lack of information on the relations between two known criminals. LEAs, in fact, may miss a lot of criminal activities such as meetings or phone calls, and therefore relevant plans of the criminal organization (Campana & Varese, 2012; Ferrara et al., 2014; Catanese et al., 2014; Agreste et al., 2016). For instance, criminals may use different telephone lines, according to the nature of the conversation and the interlocutor, and investigators may be able to identify only some of them. The frequent change of mobile phones and SIM cards and the use of particular lines to communicate with high-ranking affiliates may also prevent LEAs from identifying all conversations among suspects.

Several recent contributions have addressed the problem of missing links with particular attention to criminal networks. Fan et al. (2017) proposed a combined link prediction index considering both the nodes' types effects and nodes' structural similarities, and demonstrated that it is remarkably superior to all the 25 existing similarity-based methods both in predicting missing links and identifying spurious links in a real military network data. This method is also suitable to many other social organizations, such as criminal networks. Marciani et al. (2017) proposed three new similarity social network metrics, specifically tailored for criminal link detection and prediction and evaluated them through a flexible data stream processing application observing that the new metrics could reach up to 83% accuracy in detection and 82% accuracy in prediction and be competitive with the state of the art metrics. Bahulkar et al. (2018) described a framework which predicts the missing links in the social network data and then algorithms are applied to the augmented data to detect the communities of a transnational criminal organization (TCO). Parisi et al. (2018) proposed an entropy-based method to predict a given percentage of missing links from an observed network structure, by identifying them with the most probable non-observed ones. Diviák (2019) tried and followed a systematic and general solution to deal with the problem of missing data. He proposed three potentially synergistic and combinable techniques for data collection for each stage of data collection – biographies for data extraction, graph databases for data storage, and checklists for data reporting. Lim et al. (2019) explored the application of deep reinforcement learning (DRL) in developing a criminal network hidden links prediction model from the reconstruction of a corrupted criminal network dataset. De Moor et al. (2020) also considered the problem of missing data in criminal networks. They compared the statistics on a reduced or incomplete network with those from a known network, integrating police data on known offenders with DNA data on unknown offenders and showing how networks with both known and unknown offenders are bigger but also have a different structure to networks with only known offenders.

Despite the growing scholarly attention to missing links in criminal networks and the important consequences of missing relations, there is no previous research work aiming at modelling uncertainty in criminal networks and how such an uncertainty affects the analysis of a criminal network. Our work aims at filling this gap and, to the best of our knowledge, it represents the first step toward an objective assessment of the robustness of a link prediction algorithm in a criminal network.

3. Research Questions

We assume that *multiple interactions* can be observed among the members of a Mafia group: typical interactions are phone calls, meetings, interactions on Social Media platforms, financial transactions and so on. For each interaction, we represent a Mafia group as a graph in which a node identifies an individual who belongs (or is close) to the group and an edge indicates that an interaction occurred between two individuals represented by the nodes tied by that edge.

If we assume that K type of interactions exist, we can understand a Mafia group as a collection of graphs $\mathcal{G} = \{G_A, G_B, \ldots, G_K\}$. We call each graph $G_D \in \mathcal{G}$ as *interaction graph*. The topology of each interaction graph G_D has a remarkable value for investigation purposes: for instance, we could use the topology of G_D to: *(i)* infer collaboration between criminals, *(ii)* to identify individuals who intercept most of the information flow in the criminal organization and *(iii)* to design police operations to dismantle the underlying criminal organization.

In this paper we will concentrate on task *(i)*, and, more specifically, we seek an answer to the following question:

> **Q**$_1$ *Given an interaction graph $G_D \in \mathcal{G}$, can we design algorithms to accurately predict edges between pairs of criminals*?

Our research question is strictly linked to a popular problem in Network Science, namely the *link prediction problem in networks* (Liben-Nowell & Kleinberg, 2003). The link prediction problem has been extensively studied in a number of domains, such as e-commerce (Chen et al., 2005), homeland security (Hasan et al., 2006) and bioinformatics (Menon & Elkan, 2011). In the criminology the problem of predicting links is well known but there are few studies approaching it (Rhodes & Jones, 2009; Berlusconi et al., 2016).

Given a link prediction algorithm $\mathcal{A}$, one could ask if the accuracy of $\mathcal{A}$ depends on the topological features of the interaction graph, say G_D, on which we decide to run $\mathcal{A}$. More formally, we are interested in answering the following question:

> **$\mathbf{Q}_2$** *Given a Mafia network $\mathcal{G} = \{G_A, G_B, \ldots, G_K\}$ and a link prediction algorithm $\mathcal{A}$, how does the topological features of a dimension graph $G_D \in \mathcal{G}$ impact on the accuracy of $\mathcal{A}$?*

Finally, a fundamental issue of criminal networks is that they are incomplete and noisy and, thus, a key scientific challenge is about the *robustness* of the results that the algorithm $\mathcal{A}$ produces. In other words, we are interested in estimating how uncertainties in the topology of G_D impact of the accuracy of $\mathcal{A}$ and such a reasoning leads us to formulate the following question:

> **$\mathbf{Q}_3$** *How sensitive is the link prediction algorithm $\mathcal{A}$ to uncertainty in G_D? Can police forces get reliable outcomes when they apply the algorithm $\mathcal{A}$ on a specified dimension graph G_D?*

In the next subsections we illustrate the outcomes of our study to answer **$\mathbf{Q_1}$-$\mathbf{Q_3}$**; our study builds upon a case study draw from "Montagna", a law enforcement operation tackling Mafia gangs in the North of Sicily.

4. Fundamentals of Graph Theory

In this section we introduce some basic definitions from graph theory which will be largely used throughout the paper.

Definition 1 (Graphs). *A graph $G = \langle N, E\rangle$ is a pair in which N is the set of nodes and $E \subseteq N \times N$ is the set of edges. A graph is* undirected *if $\langle i, j\rangle \in E$ implies that $\langle j, i\rangle \in E$ for each pair of nodes i and j, directed otherwise. The* non-edge set *$T \subseteq N \times N$ is the complement of E, i.e., $T = \{\langle i, j\rangle : i \in N, j \in N \wedge \langle i, j\rangle \notin E\}$.*

Given a graph $G = \langle N, E\rangle$, we say that a graph $G' = \langle N', E'\rangle$ is a *subgraph* of G if $N' \subseteq N$ and $E' \subseteq E$.

In this paper we consider only undirected graphs. A graph G is associated with an adjacency matrix $\mathbf{A}$, whose entries are defined as follows: $\mathbf{A}_{ij} = 1$ if and only if $\langle i, j\rangle \in E$, 0 otherwise. The *order* of a graph G is the number $n = |N|$ of its nodes and the *size* of G is defined as the number $m = |E|$ of its edges. A graph with n nodes may contain at most $\binom{n}{2}$ edges and it is said *complete*. The ratio $\delta = \frac{m}{\binom{n}{2}}$ is known as *graph density*: if δ is $O(n^{-1})$ we say that the graph is *sparse*, *dense* otherwise. We define the *neighbour-set* $N(i)$ of i as the set of nodes connected to i and the *degree* d_i of i as $d_i = |N(i)|$. The average degree $\overline{d}$ of a graph G is defined as $\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$. We also define the *local clustering coefficient* (Watts & Strogatz, 1998) of G as follows:

Definition 2 (Clustering Coefficient). *Let G be a graph with non-edge set T. Let $i \in N$ be a node in G with neighbour-set $N(i)$. Let us define the set $S(i) = \{\langle j, k\rangle : j \in N(i), k \in N(i), \langle j, k\rangle \in E\}$, i.e., $S(i)$ contains pairs of nodes j and k which are both connected to i and which are connected through an edge.*
The local clustering coefficient *$lc(i)$ of i is defined as follows:*

$$lc(i) = \frac{|S(i)|}{\binom{d_i}{2}} \tag{1}$$

The average clustering coefficient *ac is defined as:*

$$ac = \frac{1}{|N|}\sum_{i \in N} lc(i) \tag{2}$$

Parameter	G_M	G_P
Number of Nodes $\|N\|$	101	100
Number of Edges $\|E\|$	256	124
Average Degree $\overline{d}$	5.07	2.48
Density δ	0.051	0.025
Average Clustering Coefficient *ac*	0.656	0.105

Table 1: Statistics of G_M and G_P graphs.

The clustering coefficient $lc(i)$ quantifies to which extent the neighbours of the node i tend to form a tie. It is of crucial importance in the study of criminal networks because, due to previous criminological studies (Berlusconi, 2013; Agreste et al., 2016), we expect that criminals form dense clusters, thus implying large values of clustering coefficients.

For our research purposes we are also interested in more complex structures, such as *walks* and *paths*, which are specified below:

Definition 3 (Walks and Paths). *Let G be a graph. A* walk of length $r-1$ *in G is a sequence $i_0, i_1, \ldots, i_r$ of nodes such that $\langle i_x, i_{x+1}\rangle \in E$ for each $x \in \{0, 1, \ldots, r\}$. A walk with* no-repeated *nodes is called* path.

5. The Montagna Police Operation

Our case study concerns the anti-mafia operation called "Montagna" concluded in 2007 by the Public Prosecutor's Office of Messina (Sicily) and conducted by the R.O.S. (*Reparto Operativo Speciale*, or Special Operations Group, a specialized anti-mafia police unit of the Italian Carabinieri). The investigation, one of the most important of the period, focused on the Cosa Nostra groups known as "*Mistretta*" family (hereafter, clan A) and the "*clan Batanesi*" (hereafter, clan B).

From 2003 to 2007, these families had infiltrated several economic activities including the public works in the area, through a cartel of entrepreneurs close to Cosa Nostra. The groups engaged in extortion racketeering and provided illegal protection to achieve illegal profits from the public construction works, with dynamics similar to those described by Gambetta (1993) and Gambetta & Reuter (1995). Furthermore, the investigation showed that the "*Mistretta*" family had taken on the role of mediator between the Cosa Nostra families of Palermo and Catania and the other criminal organizations around Messina. Indeed both the Mistretta and Batanesi families had close connections with other Cosa Nostra families located in the province of Messina, namely the "*Barcellona*" family (hereafter clan C), and the "*Caltagirone*" family (hereafter clan D). The charges were upheld by several trials and the majority of the individuals have been sentenced to long prison terms.

The main data source is the pre-trial detention order by the Court of Messina's preliminary investigation judge issued on March 14, 2007 towards the end of the investigation. The order concerned a total of 52 suspects, all charged with the crime of participation in a Mafia clan (Article 416 bis of the Italian Criminal Code) as well as other crimes (e.g., theft, extortion, damaging followed by arson). According to the Italian Criminal code, the affiliation to a Mafia clan carries a penalty of between ten and fifteen years of imprisonment. The Court ordered the pre-trial detention for 38 individuals and provided detailed motivations for the decision in a document of more than two hundred pages with an important amount of information about the suspects' crimes, activities, meetings, and calls.

Most of the information from judicial documents were about clan A and clan B. From the analysis of legal documents we built two graphs: *a) the Meeting Graph G_M*, in which nodes are uniquely associated with suspected criminals and edges specify meetings among individuals *b) the Phone Call Graph G_P*, in which nodes are uniquely associated with suspected criminals and edges records phone calls between pairs of individuals (Ficara et al., 2020).

The G_M graph had 101 nodes and 256 edges while G_P had 100 nodes and it contained only 124 edges. There were 47 individuals who jointly belonged to G_M and G_P; some statistics about G_M and G_P are displayed in Table 1.

Nodes in G_M and G_P can take active roles in the criminal organization: for instance, some nodes correspond to individuals who can be classified as "boss" (i.e., leaders of the criminal organization) while others are classified as "picciotti" (i.e., soldiers of the organization). Of course, some individuals can have regular contacts with members of

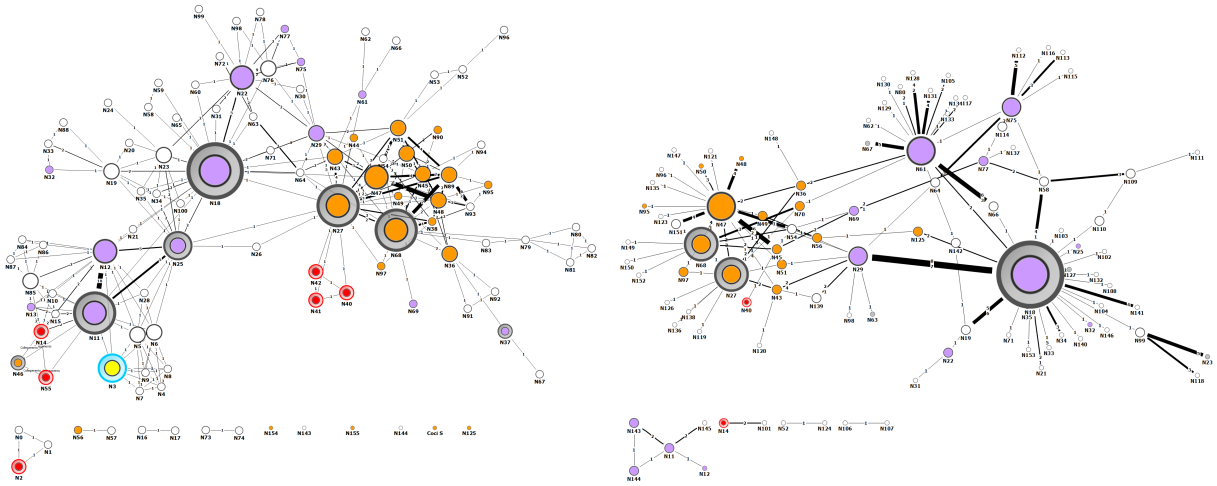

Figure 1: **Left Panel**. A graphical representation of the G_M graph. In G_M, nodes which represent the members of the "Mistretta" family and the "Batanesi" family are highlighted in violet and orange, respectively. Circled nodes correspond to the subjects investigated for having promoted, organized and directed the Mafia association (leaders). The red and yellow circled nodes refer to bosses of Mafia families of other districts. The white knots represent the other subjects considered to be: i) close to the association and ii) not classifiable in any of the previous categories, but nevertheless useful for the purposes of the Mafia-type association and the realization of its plans. **Right Panel**. A graphical representation of the G_M graph. The color of nodes has the same meaning as in the G_M. In both G_M and G_P, the width of the edges is proportional to the number of meetings (or phone calls) and the size of the nodes to their degree.

a criminal organization (due, for instance, to kinship relations) but they are not involved in any criminal activity. To guarantee anonymity, we used the symbol Nx (being x an integer) to identify an individual in both G_P and G_M.

In Figure 1 we graphically report the G_M graph (left panel) and the G_P graph (right panel): here the width of a node is proportional to its degree, while the width of an edge is proportional to the total number of meetings (or telephone calls) recorded between the nodes that edge connects. Members of the "Mistretta" and "Batanesi" families are colored in orange and purple, respectively.

6. Link Prediction in Montagna

In this section we consider the problem of predicting links in the G_M and G_P graphs.

The link prediction problem (Liben-Nowell & Kleinberg, 2003) is defined as follows:

Definition 4. *Let $G = \langle N, E\rangle$ be an undirected graph and let $G' = \langle N, E'\rangle$ be a subgraph of G which contains all nodes in G and a subset $E' \subseteq E$ of its edges. The link prediction problem consists of printing a list of non-edges in G' which are edges in G.*

We will call the set E' as the *training set* and the set $E - E'$ as the *test set*.

In practice, algorithms to solve the link prediction problem build a matrix $\mathbf{\Omega}$ in which the entry $\mathbf{\Omega}_{ij} = \sigma_{ij}$ specifies the *degree of similarity* between the nodes i and j; all pair of non edges $\langle i, j\rangle$ in G are thus ranked in decreasing order of similarity and non-edges with the largest similarity scores are the most likely to exist (Liben-Nowell & Kleinberg, 2003).

We can define many similarity scores to compute the similarity degree of two nodes in G. In what follows we first illustrate some of these similarity metrics and, thus, we analyse their accuracy in predicting edges in G_M and G_P. Methods to compute node similarity can be classified into *local* and *global* methods

6.1. Local Methods to calculate node similarity

A first class of methods to calculate node similarity in graphs is known as *local methods* (Liben-Nowell & Kleinberg, 2003; Leicht et al., 2006). because they only require the knowledge of the neighbours of two nodes i and j. Some of the most popular local methods are as follows:

1. *Jaccard Coefficient* (JC) (Jaccard, 1912; Liben-Nowell & Kleinberg, 2003):

$$JC(i, j) = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|} \tag{3}$$

2. *Common Neighbors* (CN) (Newman, 2001; Liben-Nowell & Kleinberg, 2003):

$$CN(i, j) = |N(i) \cap N(j)| \tag{4}$$

3. *Preferential Attachment* (PA) (Newman, 2001; Liben-Nowell & Kleinberg, 2003):

$$PA(i, j) = d_i \times d_j \tag{5}$$

4. *Adamic-Adar coefficient* (AA) (Adamic & Adar, 2003; Liben-Nowell & Kleinberg, 2003):

$$AA(i, j) = \sum_{x \in N(i) \cap N(j)} \frac{1}{\log |N(x)|} \tag{6}$$

6.2. Global Methods to calculate node similarity

Observe that both the G_M and G_P graphs are highly sparse and, thus, we expect that the task of predicting edges is hard if we would rely only on local information.

However, both G_M and G_P display a very high clustering coefficient (see Table 1), which is much higher than that we observe in other type of real-life social networks of roughly equal size. A large clustering coefficient implies that if two nodes i and j share at least one neighbour, then there is a high chance that i and j will be linked by an edge too. Therefore, methods to calculate node similarity which leverage higher order structures (e.g., as walks or paths) or, more in general, the *full knowledge* of the graph topology, might be more accurate than local methods in predicting edges.

We will call these methods as *global methods* and one of the most popular global methods is the so called *Katz score* (Katz, 1953).

The Katz score $\kappa(i, j)$ associated with a pair of nodes i and j considers the whole ensemble of walks connecting i and j and it assumes that each walk provides a contribution to determine the degree of similarity between i and j. A core assumption in the calculation of $\kappa(i, j)$ is that long walks are to be penalized with respect to short ones, which implies that two nodes are highly similar if they are connected by many short walks in G. To formally encode such a principle, we introduce a discount factor α and we denote $w_k(i, j)$ as the number of walks of length $k = 0, 1, \ldots$, from i to j. The Katz coefficient score is then computed as follows:

$$\kappa(i, j) = w_0(i, j) + \alpha w_1(i, j) + \alpha^2 w_2(i, j) + \ldots + \alpha^k w_k(i, j) + \ldots = \sum_{k=0}^{\infty} \alpha^k w_k(i, j) \tag{7}$$

Observe that $w_0(i, j) = 1$ if and only if nodes i and j coincide, 0 otherwise. If we let $\mathbf{A}$ be the adjacency matrix of G and suppose that α is less than $\frac{1}{\lambda_1}$, λ_1 being the largest eigenvalue of $\mathbf{A}$[1], then the Katz score between any pair of nodes in G can be seen as a matrix $\mathcal{K}_\alpha$, which can be computed as follows:

$$\mathcal{K}_\alpha = (\mathbf{I} - \alpha\mathbf{A})^{-1} - \mathbf{I} \tag{8}$$

Here $\mathbf{I}$ is the identity matrix.

In our analysis we consider also *Node2Vec* (Grover & Leskovec, 2016), a recent but promising approach for embedding graphs onto vectors. More specifically, given a graph $G = \langle N, E \rangle$, Node2Vec seeks at finding out a function $f : N \to \mathbb{R}^k$ where k is a fixed constant and $\mathbb{R}^k$ is the set of k-th dimensional arrays of real numbers. The main requirement we impose on f is that if two nodes i and j are "close"in G, then their representations $f(i)$ and

[1]The parameter λ_1 is also known as the *spectral radius* of $\mathbf{A}$.

$f(j)$ should be close in $\mathbb{R}^k$ too. To detect pairs of close nodes, Node2Vec simulates a random walk on G which can be thought as an interpolation of two popular procedures to explore a graph, namely the Breadth First Search (BFS) and the Depth First Search (DFS). More specifically, such a random walk is regulated by two parameters, namely the *return parameter* p (which specifies the likelihood the random walk will immediately revisiting a node) and the *in-out parameter* q: if $q > 1$, the random walk acts as a BFS because it tends to visit nodes which are close to the currently visited node; vice versa, if $q < 1$, the walk tends to move to nodes that are farther away from the current, thus simulating a DFS.

After applying the Node2Vec algorithm, each node i is associated with a vector $\mathbf{v}_i$ and the similarity of two nodes i and j is defined as the cosine similarity of vectors $\mathbf{v}_i$ and $\mathbf{v}_j$.

A further method to compute node similarity is the *Personalized PageRank similarity score* (PPR) (Avrachenkov et al., 2019), which, in matrix form, is defined as follow:

$$PPR_\alpha = (\mathbf{I} - \alpha\mathbf{P})^{-1} \tag{9}$$

The matrix $\mathbf{P}$ is a row-stochastic matrix defined as $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$: here $\mathbf{D}$ is a diagonal matrix storing the degrees of nodes in G and, $\mathbf{A}$ is the adjacency matrix of G. Therefore, the sum of the elements within each row of $\mathbf{P}$ is 1 and we can interpret $\mathbf{P}$ as the transition probability of a random walk over G in which the random walker, at any step, chooses uniformly at random one of its neighbors.

6.3. Simulation setup

As a first step of our analysis, we compare local and global methods. Specifically, let $\sigma_{\text{met}}(i, j)$ be the similarity score between a pair of nodes i and j calculated by means the method met, where met is one of the methods previously introduced. Similarity scores generated by each method were normalized to range from 0 to 1. We claim that i and j are connected if and only if $\sigma_{\text{met}}(i, j)$ is bigger than a threshold θ and we negate the existence of that edge if $\sigma_{\text{met}}(i, j) < \theta$. In this way, we were able to map continuous similarity scores onto discrete labels (i.e., 0 and 1 to claim/negate the existence of an edge).

We used two metrics to evaluate the level of association between a particular measure of similarity and the existence of an edge: *(i)* the *True Positive Rate (TPR)* and *(ii)* the *True Negative Rate (TNR)*. The TPR measures the proportion between the number of edges that a similarity measure claims exist and the real number of edges. The TNR is the proportion between the number of node pairs that according to a particular similarity measure are not connected and the actual number of pairs of nodes not connected. Space limitations preclude us from reporting the TPR and TNR of local methods for a broad range of values of θ. However, it is instructive to comment the configuration $\theta = 0.5$: here we observe that the TPR of all local methods was around 0 and their TNR were close to 1. Such a result implies that local methods almost always negate the existence of an edge and, thus, due to the sparsity of G_M and G_P, their guesses are almost always exact. Of course, local methods fail to identify edges actually existing. The Katz and PPR scores, instead, work much better than local methods. Because of space limitations, in Figure 2 we plot only the TPR and TNR for Katz score as function of α; similar results hold true for PPR as function of α.

The main conclusions we drew from our analysis are as follows: *(i)* An increase of α yields a decrease in TPR. *(ii)* The TNR achieved by the Katz score in G_M and G_P is generally very large (bigger than 0.99) even if slightly smaller that achieved by local methods. Specifically, Figure 2 (Left Panel) indicates the presence of a turning point $\overline{\alpha}$ (with $\overline{a} \simeq 0.52$ in case of the G_M graph and $\overline{a} \simeq 0.44$ in case of the G_P graph) beyond which the TPR quickly drops. The Katz score thus perfectly addresses issues we highlighted above and, with a suitable choice of α, all highly-scored pair of nodes are actually tied by an edge. Such a result agrees fairly well with our model about information flow in criminal networks: criminal often do not communicate directly each other but they prefer to make use of intermediaries to convey messages, both in face-to-face meetings and in case of phone calls; however, the chain of intermediaries is generally very short for security reasons.

6.4. Accuracy of Link Prediction Methods in Montagna

As a further step of our analysis, we analyse the accuracy of the methods to calculate node similarity.

We applied 10-fold cross validation to quantify the predictive accuracy of each previous predictor. Cross-validation is a procedure used to assess the accuracy of a Machine Learning algorithm which, in the latest years, gained an astonishing popularity (Hastie et al., 2009). The main reasons explaining the popularity of k-fold cross validation

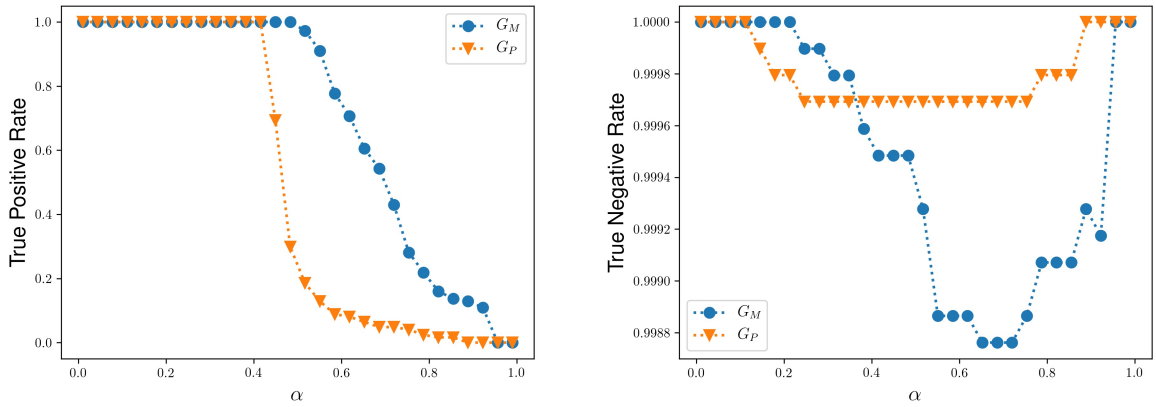

Figure 2: **Left Panel**. True Positive Rate (TPR) associated with the Katz score as function of α on G_M and G_P graphs. **Right Panel**. True Negative Rate (TNR) associated with the Katz score as function of α on G_M and G_P graphs.

are its simplicity as well as its ability of producing less optimistic accuracy assessment than methods based on the (random) division of a dataset into a training and a test part. In short, in the k-fold cross validation we randomly shuffle a dataset $\mathcal{D}$ and divide it into k groups, say, $G_1, \ldots, G_k$; common choices for k are $k = 5$ and $k = 10$. For each group G_i, we take G_i as test dataset and we use the remaining $G_1, G_2, \ldots, G_k$ groups as training set: in other words we use all groups G_j (but G_i) to fit our model; once our model has been fitted, we evaluate its accuracy on G_i. Such a procedure is repeated for each group G_i and, consequently, any sample in the original dataset is used $k - 1$ times for training purposes and one time for testing purposes. At the end of evaluation procedure we obtain k values of accuracy (one for each group used as test set); we thus take the average of the accuracy scores on each group Gi as the accuracy of the algorithm to evaluate.

The prediction accuracy is evaluated by a standard metric, the Area Under the Receiving Operating Curve (AUROC) [2]. We repeated the calculation of AUROC $n = 50$ times, thus generating a sample of the true AUROC scores. We then calculated the empirical mean m and the empirical standard deviation s of the sample above; if we denote as μ the true AUROC, it is well-known that the random variable $t = \frac{\sqrt{n}(m-\mu)}{s}$ follows a t-student distribution with $n - 1$ degrees of freedom (Ross, 2017). We then calculated the value of A for which $P(-A \leq t \leq A) = 0.95$ and we take the interval $\left(m - A\frac{s}{\sqrt{n}}, m + A\frac{s}{\sqrt{n}}\right)$ as the 95% confidence interval associated with the true AUROC score.

In Table 2 we report the confidence intervals (CI) associated with AUROC for AA, CN, PA, JC, Node2Vec, Katz and PPR methods on G_M and G_P graphs. We report the AUROC 95% confidence intervals for the Katz and PPR methods. Moreover, we considered some specified values of the parameter α (namely $\alpha = 0.1, 0.3, 0.5, 0.7$ and 0.9) and we investigated how the α parameter affected the AUROC.

In case of G_M graph, the AUROC is generally very high for all methods under investigation and the worst-performing method is PA. The Katz score and the PPR score generally outperform but their AUROC tends to slightly decrease as α increases: for instance, if $\alpha > 0.7$ the AUROC achieved by Katz score ranges from 0.893 to 0.918 while the AUROC measured for PPR ranges between 0.922 and 0.94. The JC, CN and AA methods achieve an AUROC which is slightly smaller than that of the Katz and the PPR score. In contrast, PA displays the worst performance and its AUROC is 17.51% less than that of AA and 17.25% smaller than that of JC.

On the G_P network, instead, the PA method achieves the highest AUROC and the performances of all other methods significantly deteriorate. For instance, the AA method achieves an AUROC ranging from 0.602 to 0.643, with a loss of more than 30% with respect to the G_M graph.

We are therefore able to answer research questions $\mathbf{Q}_1$ and $\mathbf{Q}_2$. As for $\mathbf{Q}_1$, we observe a few methods introduced in this paper are very and they achieve an AUROC, which, in some cases, is higher than 0.9.

[2]The AUROC is understood as the probability that a randomly chosen edge in the test set gets a higher score than a randomly chosen non-edge.

Method		*CI*	
Name	α	G_M	G_P
JC		(0.917, 0.938)	(0.57, 0.623)
CN		(0.938, 0.953)	(0.595, 0.634)
PA		(0.754, 0.789)	(0.872, 0.913)
AA		(0.939, 0.957)	(0.602, 0.643)
Node2Vec		(0.899, 0.919)	(0.577, 0.646)
$\mathcal{K}$	0.1	(0.946, 0.959)	(0.706, 0.753)
	0.3	(0.95, 0.965)	(0.711, 0.772)
	0.5	(0.939, 0.955)	(0.701, 0.754)
	0.7	(0.927, 0.946)	(0.73, 0.778)
	0.9	(0.927, 0.946)	(0.696, 0.748)
PPR	0.1	(0.939, 0.956)	(0.66, 0.724)
	0.3	(0.954, 0.968)	(0.696, 0.752)
	0.5	(0.942, 0.958)	(0.698, 0.747)
	0.7	(0.939, 0.955)	(0.687, 0.743)
	0.9	(0.922, 0.94)	(0.688, 0.75)

Table 2: AUROC Confidence Intervals for the AA, CN, PA, JC, Node2Vec, Katz Coefficient and PPR methods computed on the G_M and G_P graphs.

As for $\mathbf{Q}_2$, we report that graph topology actually plays an important role on the process of predicting edges: specifically, methods which are very accurate on G_M performs badly on G_P (and vice versa). In detail if a graph is poorly connected (with a low edge density and a small clustering coefficient) local methods are to be preferred to global ones. Vice versa, global methods as the Katz score achieve their best accuracy on graphs which display a better level of connectivity (i.e. in graph with larger edge density and clustering coefficients).

7. Robustness of the Link Prediction algorithms

In this section we aim at answering our research question $\mathbf{Q}_3$. We recall that both G_M and G_P are built upon the evidence collected by police forces and, therefore, they are an incomplete sample of true graphs G'_M and G'_P. An important discrepancy between G_M and G'_M (resp., G_P and G'_P) might significantly alter the conclusions we can draw from the analysis of G_M (resp. G_P) and, in particular, it might severely alter our ability of predicting edges between criminals.

We run our analysis in parallel for the two networks G_M and G_P and rely on the methods achieving the highest prediction accuracy in the analysis of the previous section: for G_M, we concentrate on the Katz score with different levels of parameter α; for G_P, we focus on PA.[3] Let us consider the G_M graph and our aim is to quantify the difference between $\mathcal{K}_\alpha(G_M)$ and $\mathcal{K}_\alpha(G'_M)$ At an aggregate level, we introduce the parameter $\rho_\alpha(G_M, G'_M)$ to quantify such a difference:

$$\rho_\alpha(G_M, G'_M) = \frac{||\mathcal{K}_\alpha(G_M) - \mathcal{K}_\alpha(G'_M)||_2}{||\mathcal{K}_\alpha(G_M)||_2} \tag{10}$$

Equation 10 can be applied to graphs G_P and G'_P and method PA, which yielded the highest prediction accuracy in the telephone call network. However, the equation is unapplicable in practice because we do not know the true graphs G'_M and G'_P. We can overcome this issue by assuming that missing edges – i.e., those edges in G'_M (resp., G'_P) but not observed in G_M (resp., G_P) – have been generated using a suitable *probabilistic model*.

Our probabilistic model assumes that non-observed edges in G'_M (resp. G'_P) are non-edges in G_M (resp., G_P); each non-edge in G_M (resp., G_P) is associated with a parameter ℓ, called *likelihood*, such that the higher the likelihood, the

[3] For G_M, we have also run our analysis in the case of PPR score with similar results. Due to space limitations we report only results in case of the Katz score.

p	*CN*	*JC*	*Random*
1.0	0.134	0.107	0.121
5.0	0.157	0.174	0.179
10.0	0.232	0.235	0.201
15.0	0.291	0.296	0.233

Table 3: ρ as function of p, percentage of added edges, in the G_P graph.

more likely a non-edge in G_M (resp., G_P) will correspond to an edge in G'_M (resp., G'_P). If the likelihood ℓ is specified, we can select non-edges from G_M (resp., G_P) on the basis of their likelihood and we can incrementally insert them into G_M (resp., G_P) until a pre-defined stop condition is satisfied. At the end of this procedure we obtain G'_M (resp., G'_P).

We considered multiple strategies to model the likelihood ℓ, namely: *(i)* Common Neighbors (see Equation 4), *(ii)* Jaccard's Coefficient (see Equation 3), and *(iii)* a Random model, a baseline where ℓ is distributed as a uniform random variable in the interval $[0, 1]$.

The model above resemble *network-growth models* (Newman, 2010) which describe the creation/evolution of a network (for example, a mechanism similar to the preferential attachment is at the base of the generation of Barabasi-Albert networks). However, in network-growth models we assume that new nodes arrive and join the network and the last node can decide which other nodes to connect to. In contrast, in our model, there are no new nodes that can be added to the network: this is equivalent to the simplifying hypothesis that the network is *perfectly observable* about what concerns the subjects in it (that is, the investigation has not excluded any criminal subject) and the possible lack of information only concerns the relations observed by the investigators.

Our experimental protocol consists of the following steps:

Step 1 Let T be the list of non-edges in G_M (resp., G_P) sorted by decreasing likelihood scores. We took 50% of the top elements in T, i.e., we chose half of the non-edges that have the highest likelihood. This step is necessary to create a group of potential non-edges that is sufficiently large but, at the same time, which is reliable enough because non-edges with low values of likelihood are filtered out. We call $C \subseteq T$ the set of the non-edges generated at the end of Step 1.

Step 2 We randomly choose a sample of $R(p) \subseteq C$ with size equal to p from C. In our experiments, we set $p = \{1\%, 5\%, 10\%, 15\%\}$. Of course, the larger p, the higher the number of missing edges.

Step 3 We add elements in $R(p)$ to G_M (resp., G_P), thus creating a new graph $G'_M(p) = \langle N, E \cup R(p)\rangle$ (resp., $G'_P(p) = \langle N, E \cup R(p)\rangle$).

Step 4 We calculate the relative variation ρ using Equation 10, where the graph G'_M (resp., G'_P) is replaced by $G'_M(p)$ (resp, $G'_P(p)$) .

Steps 2-4 have been repeated 30 times to avoid statistical fluctuations. The results are shown in Figure 3 for G_M, and in Table 3 and in Figure 4 for G_P.

As for the meeting network, the Random strategy clearly induces the highest values of ρ for any value of α and p. This is a largely predictable result: if the probability of the existence of a non-edge follows one of the other strategies (i.e., JC and CN), then the network structure is somehow able to predict the existence of missing edges. On the other hand, if edges were randomly placed, the network structure would not offer any insight to predict the existence of missing edges and, thus, the parameter ρ significantly grows: for instance, it suffices to set $p = 5\%$ and $\alpha = 0.2$ to obtain $\rho \simeq 1$.

The growth of α implies a growth of ρ for the Random strategy. For CN, ρ is relatively stable, only slightly increasing for higher values of α. A limit case happens when we decide to adopt JC as the likelihood function: in this case, the result is anti-intuitive because when α increases, a reduction of ρ occurs (with peaks up to 18%). In practice, if $\alpha \to 1$, the contribution of relatively long walks is not-negligible, and, thus, long walks are capable of contrasting

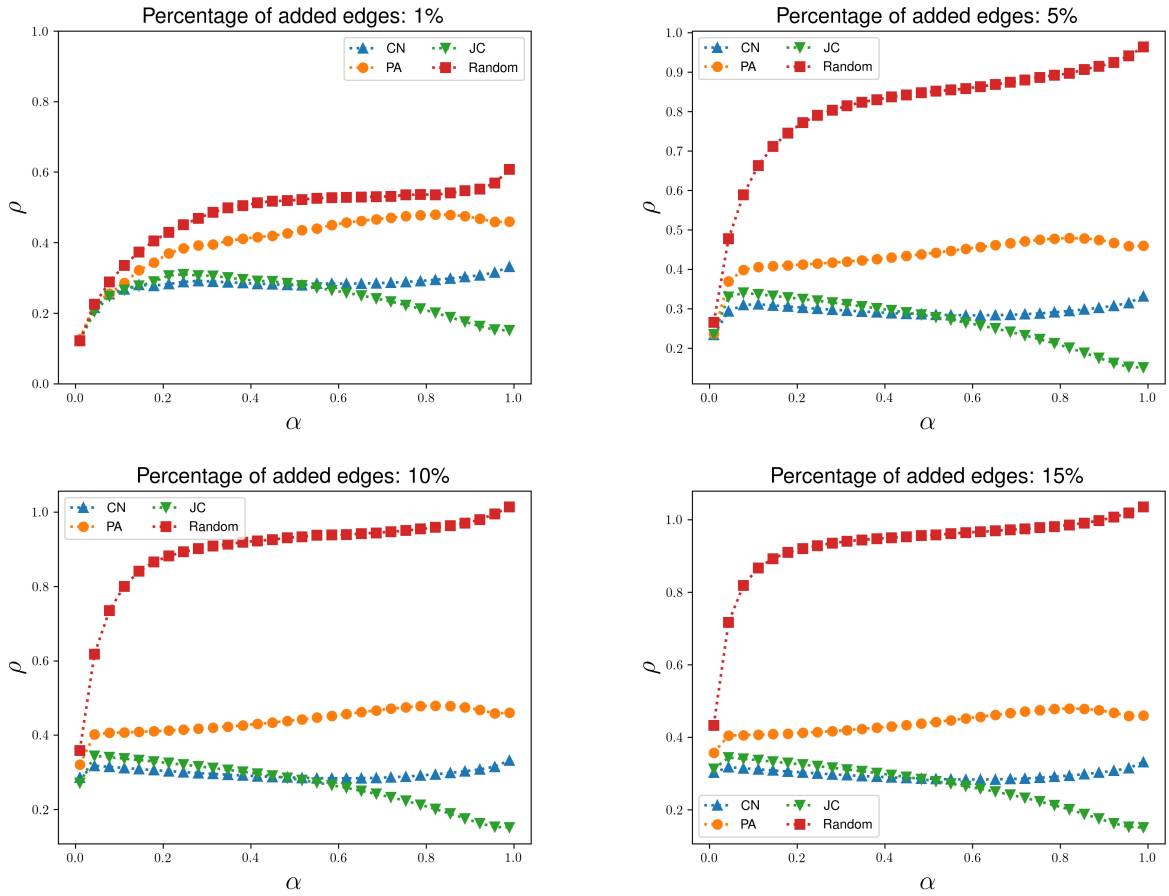

Figure 3: Variation of ρ as function of α in the G_M graph for $p = 1\%$ (**Top Left**), 5% (**Top Right**), 10% (**Bottom Left**), 15% (**Bottom Right**).

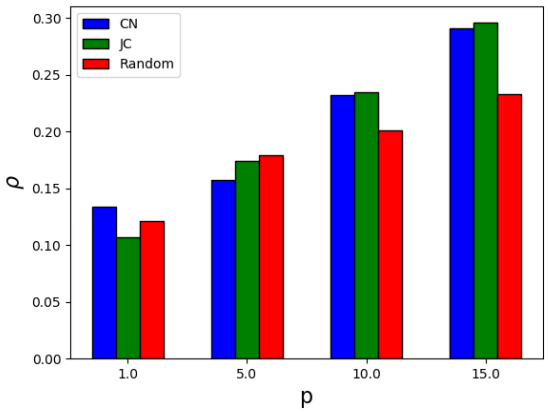

Figure 4: ρ as function of p, percentage of added edges, in the G_P graph.

high level of uncertainty associated with larger values of p. For a fixed α, the parameter p plays a key role on the value of ρ and, obviously, the higher p, the higher ρ.

We obtained totally different results in the G_P graph. The Random and JC likelihood functions are the only strategies that generate the highest value of ρ and there is a crossover point $\overline{\alpha}$ in the JC likelihood beyond which we observe a variation of ρ greater than that detected in the random generative model. The variations of ρ in the CN generative model become almost imperceptible if α gets larger than 0.1, and, thus, CN seems an unhappy choice to analyse telephone conversation flows. The trend of ρ is relatively little affected by p if $p < 15\%$; however, if $p > 15\%$ the value of ρ in all the generative models analysed undergoes significant changes.

We are now in the position of answer $\mathbf{Q}_3$: link prediction algorithms are sensitive not only to uncertainty in G_M and G_P (captured by the parameter p) but also on the type of graph they operate on. If the amount of uncertainty is relatively small and non-observed edges derive from a specified generative model, we can hope for robust edge prediction. This is confirmed by the results in G_M, where both CN and JC generate lower values of ρ than the Random strategy for different combination of of p and α. Conversely, there is no clear indication from G_P, suggesting that the network growth does not follow a specific strategy. We can thus conclude that the robustess of link prediction is not only dependent on the amount of uncertainty (i.e. p) but also on type of network and underlying relations. In the light of our studies, we recommend law enforcement agencies not only to build a detailed map of connections between criminals but also to investigate how such a map evolves over time: in this way we would be able to determine which of the likelihood functions described in this section better fit experimental observations and, if required, we could design more sophisticated likelihood functions to help law enforcement agencies to detect and prevent crimes.

8. Conclusions and Future Works

We presented a study of two criminal networks extracted from the outcome of an anti-mafia law enforcement operation called "Montagna" against individuals charged for participating in a mafia association. This study is interesting *per se* as the pre-trial detention order from which the networks under study have been extracted, namely the network of meetings and phone calls, concerns the birth and growth of a branch of Sicilian Cosa Nostra in the North-Eastern part of Sicily, a territory historically under the control of the Palermo and Catania families.

We first applied some of the most widely used similarity criteria to both networks in order to perform link prediction. The most accurate results were obtained by applying the Katz score, and our experimental finding confirm that the predictions heavily rely upon short-range interactions. This is consistent with the structure of Mafia families and the average clustering coefficient of the two networks under study.

Next, we investigated on the robustness of link prediction algorithms in presence of network uncertainties. To this end, we carried out an experiment in which the observed networks were regarded as the starting point of a growing

mechanism during which some missing edges were added. The impact of non-observed edges was measured in terms of the difference of the Katz scores.

Our experiment shows that the Meeting Graph G_M slightly differs from graphs in which up to 15% of new edges were added. On the contrary, the Phone Call Graph G_P exhibits strong differences between the observed and the new graphs. This may induce to ask whether the information relative to this network may actually be incomplete and may need some more edges which were neglected during the investigations.

Even more interestingly, the experiments we carried out clearly show that metrics of accuracy, the most widely used measures able to assess the quality of a link prediction method, should be integrated with a new measure, the stability, which takes into account the extent with which the insertion of unobserved edges modifies the network.

As for future work, we plan to apply network embedding methods to better analyse criminal networks. For our purposes, we need to resort to approaches capable of handling heterogeneous networks (Li & Tang, 2019), i.e. graphs in which nodes and edges carry specific information (e.g., nodes may be labelled with the role of an individual in a criminal group while edges specify the type of interaction between two criminals). We also plan to consider temporal information in the link prediction task, as described by Soares & Prudencio (2013).

References

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, *25*(3), 211–230. https://doi.org/10.1016/S0378-8733(03)00009-1.

Agreste, S., Catanese, S., De Meo, P., Ferrara, E., & Fiumara, G. (2016). Network structure and resilience of mafia syndicates. *Information Sciences*, *351*, 30 – 47. https://doi.org/10.1016/j.ins.2016.02.027.

Avrachenkov, K., Chebotarev, P., & Rubanov, D. (2019). Similarities on graphs: Kernels versus proximity measures. *European Journal of Combinatorics*, *80*, 47–56. https://doi.org/10.1016/j.ejc.2018.02.002.

Bahulkar, A., Baycik, N. O., Sharkey, T., Shen, Y., Szymanski, B., & Wallace, W. (2018). Integrative analytics for detecting and disrupting transnational interdependent criminal smuggling, money, and money-laundering networks. In *2018 IEEE International Symposium on Technologies for Homeland Security (HST)* (pp. 1–6). https://doi.org/10.1109/THS.2018.8574121.

Berlusconi, G. (2013). Do all the pieces matter? Assessing the reliability of law enforcement data sources for the network analysis of wire taps. *Global Crime*, *14*(1), 61–81. https://doi.org/10.1080/17440572.2012.746940.

Berlusconi, G., Calderoni, F., Parolini, N., Verani, M., & Piccardi, C. (2016). Link prediction in criminal networks: A tool for criminal intelligence analysis. *PLOS ONE*, *11*(4), 1–21. https://doi.org/10.1371/journal.pone.0154244.

Bouchard, M. (2020). Collaboration and boundaries in organized crime: A network perspective. *Crime and Justice*, *49*. https://doi.org/10.1086/708435.

Calderoni, F. (2010). Inside criminal networks. *European Journal on Criminal Policy and Research*, *16*(1), 69–70. https://doi.org/10.1007/s10610-010-9118-7.

Calderoni, F. (2012). The structure of drug trafficking mafias: the 'ndrangheta and cocaine. *Crime, Law and Social Change*, *58*(3), 321–349. https://doi.org/10.1007/s10611-012-9387-9.

Calderoni, F., Piccardi, C., & Brunetto, D. (2017). Communities in criminal networks: A case study. *Social Networks*, *48*, 116–125. https://doi.org/10.1016/j.socnet.2016.08.003.

Calderoni, F., & Superchi, E. (2019). The nature of organized crime leadership: criminal leaders in meeting and wiretap networks. *Crime, Law and Social Change*, (pp. 1–26). https://doi.org/10.1007/s10611-019-09829-6.

Campana, P., & Varese, F. (2012). Listening to the wire: criteria and techniques for the quantitative analysis of phone intercepts. *Trends in Organized Crime*, *15*(1), 13–30. https://doi.org/10.1007/s12117-011-9131-3.

Catanese, S., De Meo, P., Ferrara, E., & Fiumara, G. (2014). Detecting criminal organizations in mobile phone networks. *Expert Systems with Applications*, *41*(13), 5733–5750. https://doi.org/10.1016/j.eswa.2014.03.024.

Chen, H., Chung, W., Xu, J., Wang, G., Qin, Y., & Chau, M. (2004). Crime data mining: A general framework and some examples. *IEEE Computer*, *37*, 50–56. https://doi.org/10.1109/MC.2004.1297301.

Chen, H., Li, X., & Huang, Z. (2005). Link prediction approach to collaborative filtering. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '05)* (pp. 141–142). Denver, Colorado, USA: IEEE. https://doi.org/10.1145/1065385.1065415.

De Moor, S., Vandeviver, C., & Vander Beken, T. (2020). Assessing the missing data problem in criminal network analysis using forensic DNA data. *Social Networks*, *61*, 99–106. https://doi.org/10.1016/j.socnet.2019.09.003.

Diviák, T. (2019). Key aspects of covert networks data collection: Problems, challenges, and opportunities. *Social Networks*. https://doi.org/10.1016/j.socnet.2019.10.002.

Duxbury, S. W., & Haynie, D. L. (2019). Criminal network security: An agent-based approach to evaluating network resilience*. *Criminology*, *57*(2), 314–342. https://doi.org/10.1111/1745-9125.12203.

Everett, M., & Borgatti, S. (2013). The dual-projection approach for two-mode networks. *Social Networks*, *35*(2), 204–210. https://doi.org/10.1016/j.socnet.2012.05.004.

Fan, C., Liu, Z., Lu, X., Xiu, B., & Chen, Q. (2017). An efficient link prediction index for complex military organization. *Physica A: Statistical Mechanics and its Applications*, *469*, 572–587. https://doi.org/10.1016/j.physa.2016.11.097.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

Ferrara, E., De Meo, P., Catanese, S., & Fiumara, G. (2014). Visualizing criminal networks reconstructed from mobile phone records. *CEUR Workshop Proceedings*, *1210*.

Ficara, A., Cavallaro, L., De Meo, P., Fiumara, G., Catanese, S., Bagdasar, O., & Liotta, A. (2020). Social network analysis of sicilian mafia interconnections. In H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, & L. M. Rocha (Eds.), *Complex Networks and Their Applications VIII* (pp. 440–450). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-36683-4_36.

Freeman, L. C. (1979). Centrality in social networks' conceptual clarification. *Social Networks*, *1*(3), 215–239.

Gambetta, D. (1993). *The sicilian mafia*. Harvard University Press Cambridge, MA.

Gambetta, D., & Reuter, P. (1995). Conspiracy among the many: the mafia in legitimate industries. In *The Economic Dimensions of Crime* (pp. 99–120). Springer. https://doi.org/10.1007/978-1-349-62853-7_5.

Grassi, R., Calderoni, F., Bianchi, M., & Torriero, A. (2019). Betweenness to assess leaders in criminal networks: New evidence using the dual projection approach. *Social Networks*, *56*, 23–32. https://doi.org/10.1016/j.socnet.2018.08.001.

Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 855–864). San Francisco, CA, USA: ACM. https://doi.org/10.1145/2939672.2939754.

Hasan, M., Chaoji, V., Salem, S., & Zaki, M. (2006). Link prediction using supervised learning. In *Proc. of the SDM 2006 Workshop on link analysis, counter-terrorism and security*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hric, D., Peixoto, T., & Fortunato, S. (2016). Network structure, metadata, and the prediction of missing nodes and annotations. *Physical Review X*, *6*(3), 031038. https://doi.org/10.1103/PhysRevX.6.031038.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, *11*(2), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, *18*(1), 39–43. https://doi.org/10.1007/BF02289026.

Kim, M., & Leskovec, J. (2011). The network completion problem: Inferring missing nodes and edges in networks. In *Proc. of the SIAM International Conference on Data Mining (SDM 2011)* (pp. 47–58). Mesa, Arizona, USA: SIAM. https://doi.org/10.1137/1.9781611972818.5.

Kleemans, E., & Bunt, H. (1999). The social embeddedness of organized crime. *Transnational Organized Crime*, *5*, 19–36.

Kleemans, E., & De Poot, C. (2008). Criminal careers in organized crime and social opportunity structure. *European Journal of Criminology*, *5*, 69–98. https://doi.org/10.1177/1477370807084225.

Klerks, P. (2001). The network paradigm applied to criminal organisations: Theoretical nitpicking or a relevant doctrine for investigators? recent developments in the Netherlands. *Connections*, *24*(3), 53–65.

Krebs, V. (2002). Mapping networks of terrorist cells. *Connections*, *24*(3), 43–52.

Leicht, E., Holme, P., & Newman, M. (2006). Vertex similarity in networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, *73*(2), 026120. https://doi.org/10.1103/PhysRevE.73.026120.

Li, C., & Tang, Y. (2019). Efficient heterogeneous proximity preserving network embedding model. *Expert Systems with Applications*, *134*, 201–208. https://doi.org/10.1016/j.eswa.2019.05.044.

Liben-Nowell, D., & Kleinberg, J. (2003). The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management* CIKM '03 (p. 556–559). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/956863.956972.

Lim, M., Abdullah, A., Zaman, N., & Supramaniam, M. (2019). Hidden link prediction in criminal networks using the deep reinforcement learning technique. *Computers*, *8*, 8. https://doi.org/10.3390/computers8010008.

Marciani, G., Porretta, M., Nardelli, M., & Italiano, G. F. (2017). A data streaming approach to link mining in criminal networks. In *2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)* (pp. 138–143). https://doi.org/10.1109/FiCloudW.2017.88.

Menon, A. K., & Elkan, C. (2011). Link prediction via matrix factorization. In D. Gunopulos, T. Hofmann, D. Malerba, & M. Vazirgiannis (Eds.), *Machine Learning and Knowledge Discovery in Databases* (pp. 437–452). Berlin, Heidelberg: Springer Berlin Heidelberg.

Morselli, C. (2003). Career opportunities and network-based privileges in the Cosa Nostra. *Crime, Law and Social Change*, *39*(4), 383–418. https://doi.org/10.1023/A:1024020609694.

Newman, M. (2001). Clustering and preferential attachment in growing networks. *Phys. Rev. E*, *64*(2), 025102. https://doi.org/10.1103/PhysRevE.64.025102.

Newman, M. (2010). *Networks: an Introduction*. Oxford University Press.

Pandey, B., Bhanodia, P. K., Khamparia, A., & Pandey, D. (2019). A comprehensive survey of edge prediction in social networks: Techniques, parameters and challenges. *Expert Systems with Applications*, *124*, 164–181. https://doi.org/10.1016/j.eswa.2019.01.040.

Paoli, L. (2004). Italian organised crime: Mafia associations and criminal enterprises. *Global Crime Today: The Changing Face of Organised Crime*, *6*(1), 19–32. https://doi.org/10.1080/1744057042000297954.

Paoli, L. (2008). *Mafia brotherhoods: Organized crime, Italian style*. Oxford University Press.

Parisi, F., Caldarelli, G., & Squartini, T. (2018). Entropy-based approach to missing-links prediction. *Applied Network Science*, *3*(1), 17. https://doi.org/10.1007/s41109-018-0073-4.

Rhodes, C., & Jones, P. (2009). Inferring missing links in partially observed social networks. *JORS*, *60*, 1373–1383. https://doi.org/10.1057/jors.2008.110.

Ross, S. (2017). *Introductory statistics*. Academic Press.

Soares, P., & Prudencio, R. (2013). Proximity measures for link prediction based on temporal events. *Expert Systems with Applications*, *40*(16), 6652–6660. https://doi.org/10.1016/j.eswa.2013.06.016.

Sparrow, M. (1991). The application of network analysis to criminal intelligence: An assessment of the prospects. *Social Networks*, *13*(3), 251–274. https://doi.org/10.1016/0378-8733(91)90008-H.

Strang, S. (2014). Network analysis in criminal intelligence. In *Networks and network analysis for defence and security* (pp. 1–26). Springer. https://doi.org/10.1007/978-3-319-04147-6_1.

Van der Hulst, R. (2009). Introduction to social network analysis (SNA) as an investigative tool. *Trends in Organized Crime*, *12*(2), 101–121. https://doi.org/10.1007/s12117-008-9057-6.

Watts, D., & Strogatz, S. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918.

585 Xu, J., & Chen, H.-c. (2005). Criminal network analysis and visualization. *Commun. ACM*, *48*, 100–107. https://doi.org/10.1145/1064830.1064834.