Decision Support Systems 128 (2020) 113159

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

A novel approach to detect associations in criminal networks

Fredy Troncoso[a,*], Richard Weber[b]

[a] Departamento de Ingeniería Industrial, Facultad de Ingeniería, Universidad del Bío-Bío, Concepción, Chile
[b] Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile

ARTICLE INFO

Keywords:
Crime analytics
Social networks
Association
Rational choice
Criminal propensity

ABSTRACT

Understanding criminal groups as social networks has led to the design of powerful systems for decision support in criminal investigative work. Tools using the methods of social network analysis have proven particularly effective in the identification of associations between individuals whose relationships are not otherwise evident. This identification is typically based on the links between individuals and does not account for other relevant information, such as individual attributes. The present study proposes a new model for identifying criminal associations that incorporates this type of data. Built around a linear association model, this approach identifies the principal association between two individuals. Assuming one of the individuals as the crime planner, the approach can be used to maximize his/her utility function. The model is compared with an existing algorithm for identifying associations using a real dataset provided by the Public Prosecutor's Office of Región del Biobío-Chile. The results demonstrate the proposed model's effectiveness and flexibility in generating different association alternatives, a particularly useful feature that contributes to the more efficient use of criminal investigation resources.

1. Introduction

Criminal investigations must often mobilize large quantities of human and technical resources to track down the persons responsible for a crime [30]. As criminal behavior becomes more complex, such efforts demand the application of increasingly greater levels of knowledge, technology, experience, and time. Investigations typically begin with a set of suspects, and if the set is large, so is the number of investigative alternatives to be pursued. Each of these alternatives requires different types and amounts of resources, with no guarantee that any alternative will produce useful results. Given that resources are always scarce, an obvious need exists for systems that support the identification of the alternatives most likely to produce satisfactory results and the efficient employment of the available resources.

In criminal group investigations, an approach that has proven its effectiveness is to consider such groups as social networks amenable to existing methods of social network analysis [28]. Thus, criminal groups are represented as networks in which the nodes are individuals and the arcs are the links between individuals. A particular network is specified by identifying these links using data in the initial information base created when an investigation is launched [16]. Additional insights gathered during this investigative process can be used to enhance the respective network. To the best of our knowledge, the principal sources of information in the general network analysis methods developed for crime analysis to date are those that provide the necessary data for determining the links between individuals of interest. However, the aforementioned initial information bases used in criminal group investigation typically contain personal data on the suspects that are not sufficiently employed by these methods and which have the potential of enriching and complementing the analysis and thus improving investigative work. The present study develops a new decision support model of associating individuals that takes into account not only the links between individuals but also their personal information. This information is summarized and represented in terms of a value indicating each individual's propensity to commit certain types of group crimes. The proposed method offers greater support for analysis of criminal groups while adding flexibility to investigative efforts due to the ability to generate a set of alternatives by simply varying a certain key parameter. This flexibility potentially translates into more efficient use of investigative resources. The remainder of this article is organized into four sections. Section 2 provides some background on the application of social network analysis to criminal groups, including a review of the traditional tools used to identify criminal network associations. Section 3 introduces the proposed model, which uses integer linear programming to determine the best criminal association. Section 4 describes an example and an application of the proposed model, thus

* Corresponding author.
E-mail addresses: ftroncos@ubiobio.cl (F. Troncoso), rweber@dii.uchile.cl (R. Weber).

https://doi.org/10.1016/j.dss.2019.113159
Received 6 March 2019; Received in revised form 7 August 2019; Accepted 4 September 2019
Available online 06 September 2019
0167-9236/

highlighting its versatility. Finally, Section 5 presents our conclusions and some suggestions for further development.

2. Criminal groups and social network analysis

A social network may be defined as *the relational structure of a group or larger social system, consisting of the pattern of relationships among a collection of actors* [27]. The representation of a social system as a social network is a reflection of one of the most fruitful ideas in social sciences, which is that individuals are embedded in webs of social relations and interactions [3]. A criminal group can be understood as a social network in which the nodes represent criminals or suspects and the arcs are the links between individuals. These links act as channels for the transfer or flow of material and/or non-material resources [21]. In this view, criminal group members and their actions are seen as units that are interdependent rather than autonomous. The social network for any given criminal group is not explicit; therefore, its representation is a fundamental aspect of the social network approach in criminal investigation. A key element of this task is the definition of a representative link between the various suspects under consideration, which is achieved by analyzing the available data. This information can be extracted from a range of media, such as databases and written records, including suspect statements, bank account records, visual recordings, electronic mail, photographs, and mobile phone calls [11]. The collection of these data and their transformation into links is traditionally known as link analysis [6, 14, 24] and tends to be very labor intensive and time consuming [30]. Indeed, this activity is considered to be one of the main problems in social network spatio-temporal data mining [13]. The magnitude of the link between two individuals is commonly represented by a value between 0 and 1, where 0 represents no relationship between two individuals and 1 represents the strongest relationship. When specifying a representative link between two individuals, we can consider the reports of the individual actors (Self-Report), the evidence of communication or transfer of resources between individuals (Communication), the similarities in their social conduct (Similarity or Homophilia), and the observation of joint participation in events (Co-occurrence) [19].

Once a network has been obtained for a criminal investigation, the analysis focuses on identifying criminal structures, key individuals, and other important members based on the links that have been established [10]. Two approaches for extracting information from networks are *node evaluation* and *identification of associations*. Among the traditional *node evaluators* [21] are centrality measures taken from *social network analysis (SNA)* and node evaluation algorithms. The most common centrality measures are degree, closeness, betweenness, and eigenvector [20, 25].

The “ Degree of a node” is the number of its adjacent arcs. A node's “closeness” is the average distance between this node and all other nodes of the network. “Betweenness of a node” is the number of times a node belongs to the shortest path between other nodes. “Eigenvector centrality” measures a node's importance based on its adjacent nodes and their importance. The common algorithms include page rank [22], HITS [17], and topological potential [26].

The *identification of associations* reveals relationships between suspects that are not immediately identifiable but whose specification may be essential to obtaining good results. One method proposed to identify associations is the modified shortest-path algorithm [29]. This approach is based on the idea that the magnitude of a link as a value between 0 and 1 can also be considered to be a probability that two individuals are related. Thus, if two nodes are not directly connected but are associated through a path consisting of intermediate nodes and links, the probability of this association is the product of the probabilities of these links, assuming these links are independent events. The strongest association is the one that is the most probable.

To find this most probable association between two individuals using a shortest-path algorithm, Xu and Chen [29] proposed the following transformation of the links' magnitude:

$$l_{ij} = -ln\ v_{ij} \qquad 0 < v_{ij} \leq 1\ \forall\ i, j \tag{1}$$

where v_{ij} represents the probability that two individuals i and j are related.

This transformation generates a new graph of the network in which the nodes and links remain the same. The magnitude of each link, however, takes the transformed value according to Eq. (1). The new graph has the following important property: the shortest path between a pair of nodes i and j is the path with the highest value for the product of the probabilities and therefore the highest probability of all possible paths between these nodes. The respective proof can be found in [29].

Similarly, another method suggests measuring the strength of the association between two individuals through transformation of the original network into an edge-dual network in which each unique relation between two nodes is replaced by a so-called relation node [7, 8]. The strength of these relations is then calculated using the concept of k-connectivity, an indicator for measuring a network's cohesion. A network is said to be k-connected if k is the minimum number of nodes whose removal will divide the graph into two or more sub-graphs. Specifically, the authors apply a local approach to node connectivity. Given two nodes i and j, their association strength is determined by the minimum number k of relation nodes that must be removed to disconnect i and j. This implies that the greater the value of k is, the greater the strength of the two nodes' association. The connectivity problem is solved using a modification of the traditional maximum flow algorithm. Whereas the traditional version calculates the maximum node connectivity from a source node to a destination node, the modification calculates connectivity as the number of relation nodes that must be removed to disconnect the source node from a destination node.

The techniques used to establish association in a criminal network are focused mainly on relations among individuals, that is, the magnitude of the links between the nodes in the network. The approach we propose in the next section goes beyond existing models of association detection by incorporating individual information, included in the nodes, into the analysis of criminal groups.

3. A new association model integrating information on individuals and their connections

In this section, we develop a new model that incorporates individual attributes to search for associations between two nodes of a network. Subsection 3.1 provides a general view of our novel approach. Subsection 3.2 presents a new criterion to identify associations, incorporating information on links as well as on the nodes' propensity to belong to a criminal group. Based on this new criterion, Subsection 3.3 formally presents the proposed integer linear program (ILP), which determines the subset of individuals E and the set of links A_E that form the best association between two individuals. Section 3.4 presents ways to determine the propensity to belong to a criminal group.

3.1. The new association approach - a general view

Since two members in a social network may be associated through various different paths, a method is needed to identify the path that represents the “best" association between two members.

In the present context, the search for the best association can be interpreted as the process of forming a criminal group in which a decision maker or planner plans a group crime and chooses the other individuals who will participate based on their criminal abilities and their trustworthiness by maximizing some utility function. An individual's criminal ability is represented by their criminal propensity, and the trustworthiness between two individuals is represented by the magnitude of the link connecting them.

To find the best association between the planner and other individuals in the network, we assume that the planner acts rationally.

This implies that the individuals he/she chooses to be members of the group must have adequate criminal ability and provide sufficient trustworthiness to the group to ensure the crime is performed and the planner's utility is maximized.

The method we propose for identifying the best association in a network combines the strength of links and the propensity of nodes to commit crime.

To formulate the planner's rational choice of the best criminal group, we propose an ILP model that determines the group that maximizes the planner's utility function subject to a budget constraint plus additional conditions to achieve a path between the planner and other individuals in the network. A given pair of individuals is assumed to be part of a criminal group whose best association we want to find. One of the two individuals is assigned the crime planner's role and is denoted *s*. The planner chooses the group members, decides how the illicit proceeds of the crime (hereafter simply "the proceeds ") will be divided and assumes the risks involved. The second individual plays the role of the "receiver" in carrying out the crime and is denoted *d*. Consider, for example, the theft of a car. The criminal group in this case would consist of a chain of individuals in which the planner *s* plans the theft and then recruits the members who will steal and hide the car, those who will disassemble it for parts or alter its identification number, and an individual *d* who acts as the receiver or "fence ", selling the parts or the entire vehicle. Another example is the methamphetamine manufacturing and trafficking network [2, 4]. In this example, the planner manufactures methamphetamine in clandestine laboratories, and the receiver distributes methamphetamine in the retail market.

The identification of association is depicted visually in Fig. 1, where Fig. 1a represents the criminal network of the group containing *s* and *d* as defined by criminal analysis, Fig. 1b shows all the possible associations in the network between *s* and *d*, and Fig. 1c highlights, for purposes of illustration, one of the possible associations as the best, that is, the best criminal group that the planner can assemble for the crime to be carried out and to maximize his/her utility.

3.2. The planner's utility function

In this subsection, we formulate the planner's utility as a function of the links' magnitudes and the nodes' criminal propensities. The formulation of each of its components and the assumptions necessary to formulate the respective model are also described.

3.2.1. General form of the utility function

Various definitions of the utility of a criminal organization have been proposed by Becker (1968) [1], Garoupa (2000) [12], Kugler et al. (2005) [18] and Dnes and Garoupa (2010) [9]. Based on these definitions, we propose the following conceptual version of a criminal group utility function:

$$U = I\ p - C\ q - W \tag{2}$$

where $I\,p$ is the expected income from an illegal activity planned by the organization, I is the proceeds and p is the probability of carrying out the activity. $C\,q$ is the expected cost to the organization of bribes to buy the members' silence and prevent information on the planned crime from leaking, where C is the maximum bribe the organization is willing to pay and q is the probability of a leak. W is the payout from the proceeds to the members of the organization.

By adapting Eq. (2) to the context of a suspect network $G(N,A)$, where N is the set of individual suspects or nodes and A is the set of arcs joining the individuals, we express the general form of the planner's utility function in terms of the three components described above: Income, Cost of bribes, and Payout to group members. Thus,

$$U = I\ Pr(pcg_i;\ i \in E) - C\ Pr\Big(\sum_{(i,j)\in A_E} d_{ij}\Big) - \sum_{i\in E} W_i(pcg_i) \tag{3}$$

where pcg_i is the propensity of each individual $i \in N$ to belong to a criminal group and d_{ij} is the social distance between two individuals i and j, $(i,j) \in A$. $E \subseteq N$ is the set of individuals forming a particular association, and A_E: ={$(i, j) \in A$; $i, j \in E$}.

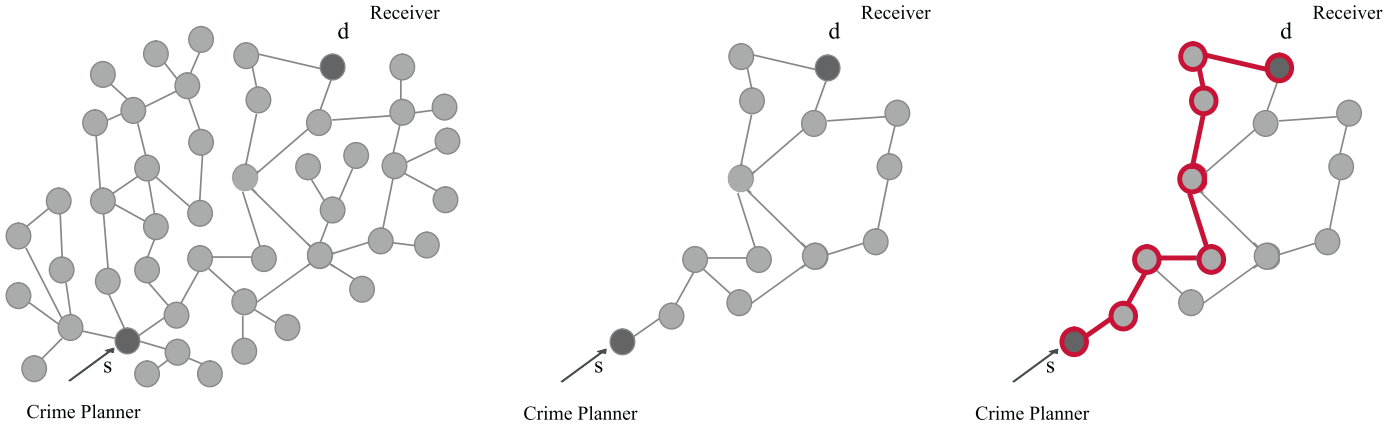
In the following three subsections, we define and justify an expression for each of the three components of this general utility function, incorporating our two measures pcg_i and d_{ij}

3.2.2. Expected income

In Eq. (3), expected income is represented by $I\ Pr(\Sigma_{\ i\in E} pcg_i)$, where $Pr(\Sigma_{\ i\in E} pcg_i)$ is the probability of carrying out a planned crime, the equivalent of p in Eq. (2). This probability depends directly on the criminal propensity $\Sigma_{\ i\in E} pcg_i$. The greater the propensity is, the greater the criminal ability of the association and the greater, in turn, the probability that a planned crime will be carried out. We now model the planner's income function using the following assumption.

Assumption 1. The probability of carrying out the crime is equal to the proportion of the criminal propensity present in the selected group.

Based on this assumption, we model the planner's income function



a) Suspect Network

b) Set of possible criminal associations between s and d

c) Best criminal association between s and d

Fig. 1. Identifying criminal associations.

as:

$$I\ Pr(pcg_i;\ i \in E) = I\frac{\sum_{i\in E} pcg_i}{pcg_{max}} \tag{4}$$

In this equation, pcg_{max} is the maximum criminal propensity that the planner can consider to carry out a crime. As a consequence of Eq. (4), when all available individuals in the network are incorporated into the criminal group (i.e., $E = N$), the planned crime is carried out with probability 1. On the other hand, when none of the individuals are incorporated, i.e., $E = \emptyset$, the probability is 0.

3.2.3. Expected cost of bribes to prevent leaks

In Eq. (3), the expected bribe cost to prevent leaks is represented by $C\ Pr(A_E)$, where $Pr(A_E)$ is the probability of a leak given the links A_E. This probability corresponds to q from Eq. (2) adapted to our case. The motivation for basing this probability is that d_{ij} is a measure of "social distance" and therefore also of mistrust (i.e., the absence of trust). Understood in this manner, total mistrust in an association is given by $\sum_{(i,j)\in A_E} d_{ij}$. Thus, we assume that the probability of a leak for a given association depends directly on the total mistrust in that association. The greater this probability is, the greater the proportion of the maximum bribe C the association will have to pay to prevent bribes. We now model the planner's bribe cost using the following assumption.

Assumption 2. The planner is willing to assign a maximum amount for preventing bribes C defined as a percentage γ of the proceeds I. The maximum amount will be paid when the maximum probability of a leak occurs, that is, in the association with the greatest mistrust between the planner s and the receiver d.

Based on Assumption 2, the planner's bribe cost function can be modeled as:

$$C\ Pr(A_E) = I\gamma\frac{\sum_{(i,j)\in A_E} d_{ij}}{d_{max}} \tag{5}$$

To obtain pcg_{max} and d_{max} used in Eq. (4) and Eq. (5), respectively, we can solve the largest path problem on the network using the social distance d_{ij} to weight the arcs. pcg_{max} and d_{max} are calculated as the sum of the solution nodes pcg_i and arcs d_{ij}, respectively.

3.2.4. Payout to criminal group members

In Eq. (3), the payout from the proceeds to the criminal group members is represented by $\Sigma_{i\in E} W_i(pcg_i)$, which corresponds to W in Eq. (2). $W_i(pcg_i)$ represents the payout each group member i receives as a function of his/her criminal ability [5], given by pcg_i.

We propose the following function for criminal group member payout:

$$\sum_{i\in E} W_i(pcg_i) = w\sum_{i\in E} pcg_i \tag{6}$$

where w is the rate the planner is willing to pay per unit of pcg. The following assumption provides a bound that is used to determine w.

Assumption 3. For any association the planner chooses, the utility he/she expects to receive is at least equal to the payment, which is proportional to his/her criminal ability given by pcg_s.

This assumption implies that for the worst association possible, the group income must be at least as much as the planner could obtain individually. This case is expressed by:

$$I - w(pcg_{max} - pcg_s) - I\gamma \geq I\frac{pcg_s}{pcg_{max}} \tag{7}$$

The left-hand side of Eq. (7) represents the proceeds less the costs incurred by the planner for the path of greatest mistrust, and the right-hand side is the planner's minimum expected utility.

The upper bound for the rate the planner is willing to pay to group members per criminal ability unit is as follows:

$$w \leq \frac{I}{pcg_{max} - pcg_s}\left(1 - \gamma - \frac{pcg_s}{pcg_{max}}\right) \tag{8}$$

If for any reason the planner chooses the path of greatest mistrust, the rates obtained for w ensure that the planner receives a utility at least commensurate with his or her criminal ability.

Assuming the planner pays the best rate possible w per criminal ability unit to ensure the chosen individuals join the group, including himself/herself, the rate is given by:

$$w = \frac{I}{pcg_{max} - pcg_s}\left(1 - \gamma - \frac{pcg_s}{pcg_{max}}\right) \tag{9}$$

3.3. Linear rational association model: the new association model

Given the elements developed in Section 3.2.4, i.e., Eqs. (4), (5) and (6) for the three components of the planner's utility function, we obtain the following utility function:

$$U = I\frac{\sum_{i\in E} pcg_i}{pcg_{max}} - I\gamma\frac{\sum_{(i,j)\in A_E} d_{ij}}{d_{max}} - w\sum_{i\in E} pcg_i \tag{10}$$

We now present an ILP model that determines the subset of individuals E and the set of arcs A_E that form the best association between individuals s (planner) and d (receiver), i.e., the association that maximizes the planner's utility, as shown in Eq. (10). First, we define the following sets of decision variables:

$$X_{ij} = \begin{cases} 1 & \text{if } (i,j) \in A_E \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

$$Y_i = \begin{cases} 1 & \text{if } i \in E \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where Eq. (11) indicates whether the link between individuals i and j is in the association and Eq. (12) indicates whether individual i is in the association.

Using these sets of decision variables, we model the planner's linear utility function as:

$$U = I\frac{\sum_{i\in N} pcg_i Y_i}{pcg_{max}} - I\gamma\frac{\sum_{(i,j)\in A} d_{ij}X_{ij}}{d_{max}} - w\sum_{i\in N} pcg_i Y_i \tag{13}$$

Substituting in the expression for w (see Eq. (9)), we obtain the final form of the planner's utility function (see also Appendix A):

$$U = \frac{I\gamma}{pcg_{max} - pcg_s}\sum_{i\in N} pcg_i Y_i - \frac{I\gamma}{d_{max}}\sum_{(i,j)\in A} d_{ij}X_{ij} \tag{14}$$

We use this planner's utility function in our ILP model to identify the best criminal group, which is the group that exhibits the best association between individuals s (planner) and d (receiver) in a network, as described above. Since $I\gamma$ is a positive constant that appears in both terms of the planner's utility function, we can divide by $I\gamma$ without altering the optimal solution. The complete formulation, denoted the linear rational association model (LiRAM), is as follows:

$$Max\ \ U = \frac{\sum_{i\in N} pcg_i Y_i}{pcg_{max} - pcg_s} - \frac{\sum_{(i,j)\in A} d_{ij}X_{ij}}{d_{max}} \tag{15}$$

s.t.

$$\sum_{j\in N} X_{sj} = 1 \tag{16}$$

$$\sum_{i\in N} X_{id} = 1 \tag{17}$$

$$\sum_{i\in N: i\neq d} X_{ij} = \sum_{k\in N: k\neq s} X_{jk} \quad \forall\, j \in N\setminus\{s, d\} \tag{18}$$

$$Y_s = 1 \tag{19}$$

$$\sum_{i \in N} X_{ij} = Y_j \quad \forall\, j \in N \setminus \{s\} \tag{20}$$

$$\sum_{i \in N} pcg_i Y_i \leq \varphi pcg_{max} \tag{21}$$

$$\sum_{i,j \in L} X_{ij} = |L| - 1 \quad \forall\ L \subseteq N \setminus \{s, d\}: |L| \geq 2 \tag{22}$$

$$X_{ij} \in \{0, 1\}\ \forall\ (i, j) \in A \tag{23}$$

$$Y_i \in \{0, 1\}\ \forall\ i \in N \tag{24}$$

Constraints (16) through (20) ensure that the planner will choose a single path to associate with the receiver. Constraint (21) represents the fact that in making this choice, the planner is willing to select a share φ of the maximum criminal propensity that the planner can consider to carry out a crime. As will be shown in our application (see Section 4), parameter φ provides a strong tool to analyze different crime scenarios. Constraint (22) eliminates solutions with subtours.

3.4. How to determine the propensity to belong to a criminal group

In Section 3.2.1 we introduced pcg_i as the propensity of each individual $i \in N$ to belong to a criminal group. pcg_i can have other interpretations depending of the particular situation. One of these interpretations is the propensity of an individual to commit crimes within a criminal group. Another meaning is the propensity of an individual to have a key role in a criminal group.

To determine pcg_i, we consider the available information regarding individual i at the time of launching the investigation. The general form to estimate pcg_i is given by:

$$pcg_i = f(s_i) \quad \forall\ i \in N \tag{25}$$

in which s_i is the set of relevant attributes of individual i and f is some function, chosen under a certain context, that transforms these attributes context-dependently into a propensity value.

When no information is available to estimate pcg_i this can be considered constant for each individual, e.g., $pcg_i = 1\ \forall i \in N$. In this case LiRAM will find the associations between individuals considering only the value of the social distance d_{ij}.

The more information is available for individual i, the better pcg_i can be estimated in a certain context. Section 4 presents two real-world applications where different interpretations of pcg_i are used. In both cases, we show ways to estimate the respective values.

4. Applications of the linear rational association model

In this section we present the application of the proposed LiRAM to identify relevant associations in a network provided by the Public Prosecutor's Office of Región del Biobío-Chile. Prior to this real-world application, we provide an example to illustrate how our model can be used.

4.1. Example application

To explain the application of the LiRAM model, we consider an example of social distance or mistrust d_{ij} and criminal propensity pcg_i as shown in Fig. 2.

In order to apply LiRAM to find the best association between the criminal planner (node 1) and the receiver (node 10), it is necessary to determine the parameters pcg_{max} and d_{max} of the objective function. These values are obtained by solving a largest path problem on the network. In this example, the maximum criminal propensity value (pcg_{max}) is 2.5 (path 1-2-6-10 and path 1-3-7-10), and the maximum mistrust d_{max} is 2.25 (path 1-2-6-10).

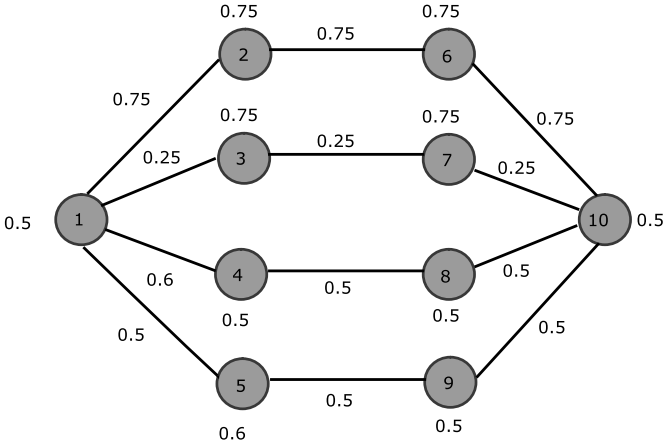


Fig. 2. Example network.

The values of pcg_{max} and d_{max} provide criteria for LiRAM to decide between greater criminal propensity and less mistrust when defining the best association. Eq. (26) shows the objective function for this example.

$$Max\ \ U = \frac{\sum_{i \in N} pcg_i Y_i}{2} - \frac{\sum_{(i,j) \in A} d_{ij} X_{ij}}{2.25} \tag{26}$$

The objective function of this example shows a slightly greater preference for criminal propensity. Using this objective function and considering the value of $\varphi pcg_{max} = 2.5$ in Eq. (21), LiRAM chooses path 1-3-7-10 as the best association. If the value of φpcg_{max} takes a value greater than or equal to 1.6 and less than 2.5, the optimality criterion of LiRAM chooses the best association from path 1-4-8-10 and path 1-5-9-10. For path 1-4-8-10, the value of the objective function is 1.4611, and for path 1-5-9-10, the value is 1.4666. Therefore, LiRAM will choose the latter path as the one with the best association. For a value of φpcg_{max} equal to 1.5, LiRAM will choose path 1-4-8-10 as the only feasible association. Path 1-2-6-10 is never considered to be a feasible association by LiRAM.

4.2. Application to The Public Prosecutor's Office of Región del Biobío-Chile dataset

The Public Prosecutor's Office of Región del Biobío-Chile is an organization that conducts the investigation of offenses and applies the corresponding actions provided by law. Much of its investigative work depends on the data associated with the historical criminal behavior of the individuals accused of an offense, called suspects.

To demonstrate the applicability of LiRAM and its effectiveness, we use a dataset provided by the Criminal Analysis Unit of the Public Prosecutor's Office of Región del Biobío-Chile. This dataset has 1666 offenses committed in the period 2002–2017 and 77 suspects. Table 1 shows the structure of the dataset.

Table 1
Structure of the dataset provided by the Criminal Analysis Unit of the Public Prosecutor's Office of Región del Biobío-Chile.

Cause code	Suspect code	Offense	Date
1700984480	CEQJ_47	Burglary in an uninhabited place	19-10-2017
1700955040	RIRR_1	Fighting in a Public Place	10-10-2017
1700920615	JMMS_18	Drug possession	01-10-2017
1700915972	SAAC_73	Injuries	24-09-2017
1700892568	ORAA_44	Injuries	23-09-2017
1700870764	CAFV_4	Burglary in an uninhabited place	18-09-2017
1700870109	BYOM_61	Theft	17-09-2017
1700854849	WDMM_20	Criminal possession of a weapon	09-09-2017
1700837824	LACS_13	Theft	06-09-2017

Table 2
Relation between suspects.

Suspect i	Suspect j	c_{ij}	d_{ij}	v_{ij}	$-\ln(v_{ij})$
AACQ_32	DEHB_70	1	1	0.25	1.386
AACQ_32	FEMC_15	2	0.5	0.5	0.693
AACQ_32	FJFR_66	1	1	0.25	1.386
AACQ_32	MAMA_25	1	1	0.25	1.386
AACQ_32	WAPM_31	3	0.33	0.75	0.287
AACQ_32	YAMP_75	2	0.5	0.5	0.693
AIAP_3	BAQV_66	1	1	0.25	1.386
AIAP_3	CASP_49	1	1	0.25	1.386
AIAP_3	FJPV_34	3	0.33	0.75	0.287
AIAP_3	JAVH_10	3	0.33	0.75	0.287
AICD_13	JIAI_57	2	0.5	0.5	0.693

In Table 1 Cause Code is a key to a criminal case investigated by the Public Prosecutor's Office. A criminal case includes one or several suspects, identified by Suspect Code and includes one or several offenses. The attribute Date is the date on which the offense was committed.

In this subsection, LiRAM will be used to identify members of a criminal group of burglary in an uninhabited place. This criminal group was already investigated and identified in 2018 by the Public Prosecutor's Office of Chile.

4.2.1. Network and determination of social distances and links

We used Cause Code and Suspect Code to establish links among individuals and construct the network. We established a link between two suspects if they have the same Cause Code. If two suspects have the same Cause Code, it means that these individuals acted jointly in one or more offenses associated with the respective cause.

In Table 2, c_{ij} is the number of criminal cases committed jointly by the suspects i and j. d_{ij} represents the social distance between suspects i and j obtained via Eq. (27).

$$d_{ij} = \frac{min\{c_{ij} > 0 \quad \forall\, i, j \in N: \quad N \quad in \quad set \quad of \quad suspects\}}{c_{ij}} \tag{27}$$

In Table 2, v_{ij} is the magnitude of the link between suspects i and j obtained via Eq. (28) and we use it in the modified shortest-path algorithm proposed by [29].

$$v_{ij} = \frac{c_{ij}}{max\{c_{ij} \quad \forall\, i, j \in N\}} \tag{28}$$

Fig. 3 shows the network obtained which is an undirected graph with 77 nodes and 374 arcs. The members of the criminal group of burglary in an uninhabited place are marked by a circle.

4.2.2. Determination of the propensity to belong to a criminal group (pcg)

In this application, we focus on the particular offense burglary in an uninhabited place. We determine the suspects' propensities to belong to the respective criminal group (pcg) based on their previous activities as follows. For each suspect i we determine two values: their overall number of burglaries in an uninhabited place and the number of such offenses during the years 2016 and 2017 as shown in Fig. 4. We interpret the first number as the suspect's overall experience and the second one as their recent experience. We then calculate the average of these two numbers overall suspects (the dotted lines in Fig. 4) leading to four quadrants

In Fig. 4 the upper right quadrant contains the suspects with the highest experience and highest level of current activity. We assigned $pcg = 1$ to these suspects. The lower right quadrant shows the suspects with low experience and a high level of current activity. We assigned $pcg = 0.75$ to these suspects. The upper left quadrant includes the suspects with high experience and a low level of current activity. We assigned $pcg = 0.5$ to these suspects. Finally, the lower left quadrant shows the suspects with low experience and a low level of current activity. We assigned $pcg = 0.25$ to these suspects. The Criminal Analysis Unit of the Public Prosecutor's Office of Región del Biobío-Chile assumes the current activity level more important than the experience in determining pcg.

4.2.3. LiRAM application and results

Once the value d_{ij} for each link (i,j) and the propensity pcg_i for each suspect i are obtained, LiRAM is applied to the network to test its effectiveness in identifying members of a criminal group of burglary in an uninhabited place. We compare the number of identified members to the number found by a modified shortest-path algorithm (SPA), such as the one discussed in Section 2, which identifies the best associations as those with the highest link weight [29].

The network of 77 suspects contains 12 members of a criminal group as indicated by circles in Fig. 3. These 12 criminals were identified by the Criminal Analysis Unit of the Public Prosecutor's Office of Región del Biobío-Chile. To validate our model, we start with the following assumption: If no previous information on criminal members is available, i.e., no suspect is identified to be a member of a criminal group, LiRAM could be used between any pair of two suspects. In our case, that would be 77*76/2 possible combinations. If, however, some members of the criminal group are already identified – as is the case in many real-world investigations - then we could use this initial information to identify additional members running LiRAM using pairs of known criminals that are not directly connected. In our case we found 86 such combinations among the 12 criminals.

The LiRAM was applied in each association for five different values of the maximum payout share φ: 0.1, 0.2, 0.3, 0.4, and 0.5. No values greater than 0.5 were used because the number of individuals included in the association grows too large and the associations are, therefore, less useful.

To measure the performance of LiRAM and SPA in each association, we used Precision rate and Recall rate defined as follows:

$$Precision = \frac{Number \quad of \quad members \quad of \quad the \quad criminal \quad group \quad in \quad the \quad association}{Number \quad of \quad suspects \quad in \quad the \quad association} \tag{29}$$

$$Recall = \frac{Number \quad of \quad members \quad of \quad the \quad criminal \quad group \quad in \quad the \quad association}{Total \quad number \quad of \quad members \quad of \quad the \quad criminal \quad group} \tag{30}$$

As an example, Table 3 shows the Precision values four out of these 86 associations, namely those between YRUR_58 and the criminal members not directly linked to him/her. The Recall values are displayed in Table 4.

In Table 3 fractional values indicate the number of members of the criminal group among the total number of suspects included in the associations. In Table 4 fractional values indicate the number of members of the criminal group among the total of members of the criminal group. For example, the association between YRUR_58 and CAAR_9 established by LiRAM, leads to a Precision of 3/6 with $\varphi = 0.2$ (see Table 3). The model includes six suspects, of which three are members of the criminal group (indicated by a red circle in Fig. 3). This result is illustrated in a continuous thick gray line in the Fig. 3. By contrast, for the same example, the modified SPA produces a Precision of 1/1. This model includes one suspect, which is a member of the criminal group as is illustrated by a dark segmented line in the Fig. 3. The association between YRUR_58 and CAAR_9 identified by LiRAM yields a Recall of 3/10 with $\varphi = 0.2$ (see Table 4). The model found three member of the criminal group out of ten members of the criminal group. The modified SPA yields a Recall of 1/10, finding one member of the criminal group out of ten members of the criminal group. By increasing his or her willingness to pay (φ), the planner can incorporate

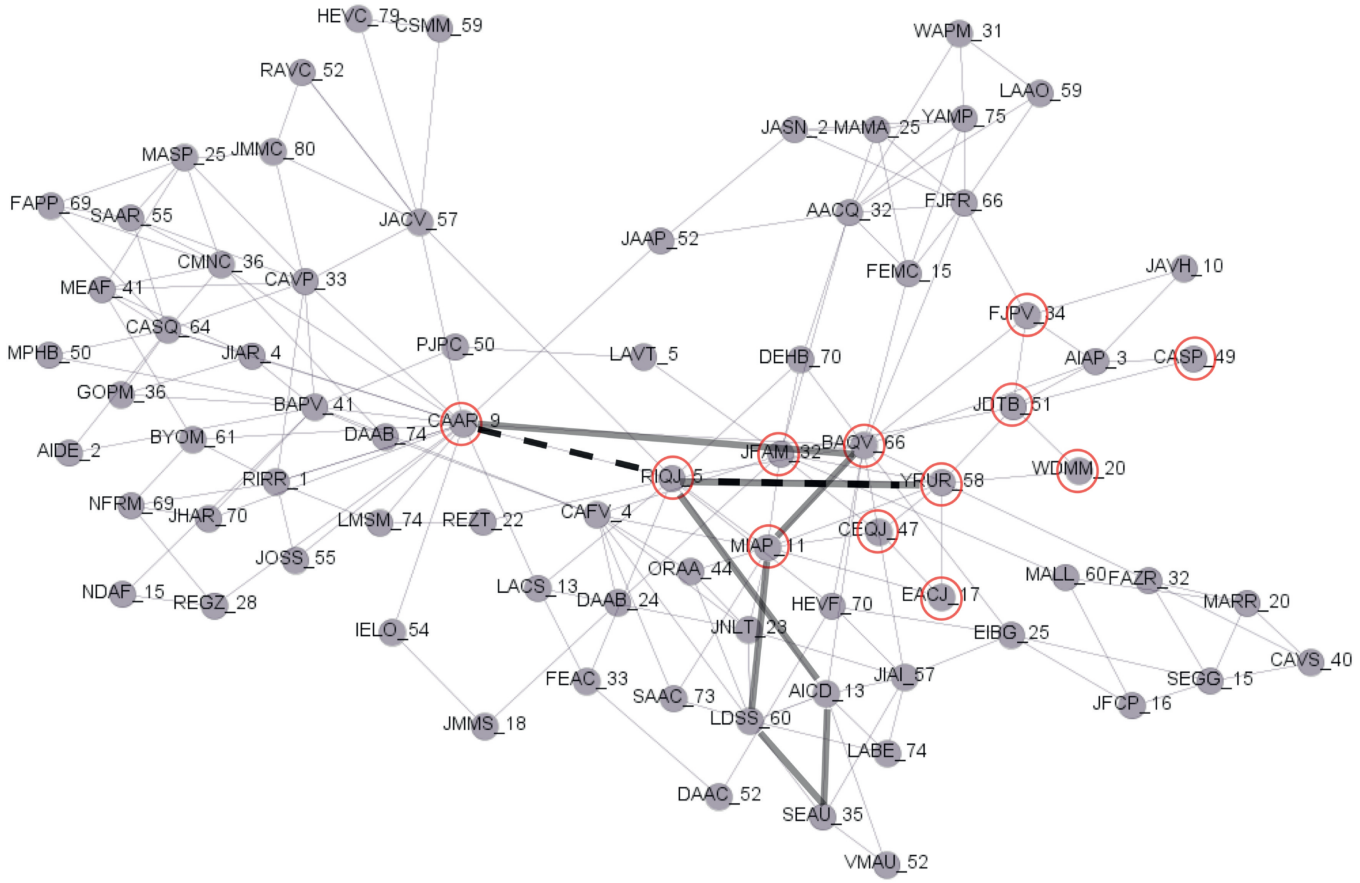


Fig. 3. Network of 77 suspects.

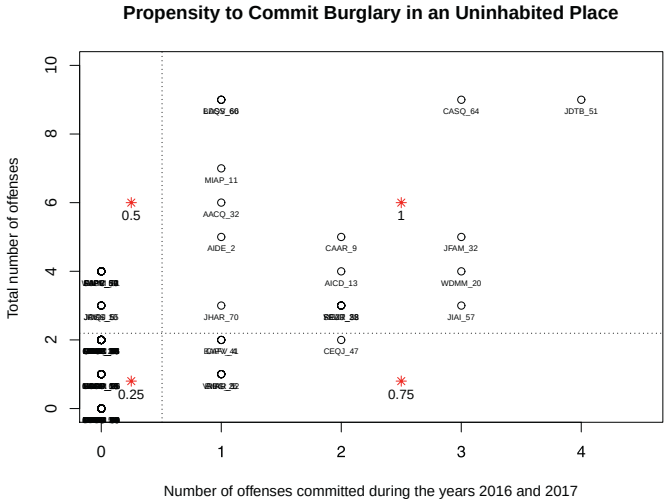


Fig. 4. Segmentation of the 77 suspects to estimate their propensity to commit robbery in an uninhabited place.

more members, and in the best case, replace members with low *pcg* by other members with higher *pcg*, given the better Recall value for higher values of φ.

The process shown above for the associations starting with YRUR_58 has been applied to all 86 associations mentioned before. The average Precision and the standard deviation of the validation process for LiRAM and Modified SPA are displayed in Table 5. The average Recall and the standard deviation are displayed in Table 6.

Table 7 shows the Average, Standard Deviation, Maximum, and Minimum of CPU-times to run LiRAM in each one of the 86 associations (considering the different values of φ). We used a 2.6 GHz Intel Core i7-6600 with 8 GB in Ram under Windows 10, using CPLEX 12.8 with one Thread.

Fig. 5 shows the average Precision and average Recall of the validation process for LiRAM (for each value of φ) and Modified SPA. The respective results have been obtained using the pcg values as determined by the above-mentioned methodology. If, however, no previous information on the suspects' criminal propensities is available, we propose using LiRAM with pcg = 1, i.e. all suspects have the same propensity to belong to a criminal group. The respective results are also shown in Fig. 5.

In the analysis of the criminal group by LiRAM and Modified SPA, Precision and Recall are relevant performance measures. We expect that LiRAM and Modified SPA will find most of the members of a criminal group in the association (high Recall) and that most of the suspects in the association will be members of the criminal group (high Precision). A problem is that Recall increases as φ is increased, while Precision decreases as φ is increased. Therefore to measure the performance in a single value that combines both measures, we use the F-measure. F-measure is the harmonic mean of Precision and Recall [23] and obtained as shown in Eq. (31).

$$F - measure = \frac{2 \quad Precision \quad Recall}{Precision + Recall} \tag{31}$$

Fig. 6 shows the average F-measure for LiRAM and SPA. It can be seen that LiRAM performed better than SPA even when using constant values for the propensities to belong to a criminal group (pcg = 1).

We prove the normality and homogeneity of variances of F-measure for the 86 associations and each value of φ in both applications of LiRAM, using the Shapiro-Wilk test and Bartlett test. Then we apply an

Table 3
Example of associations and precision values.

Precision of LiRAM and modified SPA						
Association			LiRAM			Modified SPA
YRUR_58	$\varphi = 0.5$	$\varphi = 0.4$	$\varphi = 0.3$	$\varphi = 0.2$	$\varphi = 0.1$	
FJPV_34	6/20	6/15	5/10	3/7	2/2	1/1
CASP_49	7/22	7/16	5/11	2/7	3/4	1/1
CAAR_9	6/20	6/15	6/11	3/6	2/2	1/1
BAQV_66	6/21	6/15	4/9	2/6	2/3	1/1
Average precision	30%	40.9%	48.6%	38.69%	85.4%	100%

Table 4
Example of associations and recall values.

Recall of LiRAM and modified SPA						
Association			LiRAM			Modified SPA
YRUR_58	$\varphi = 0.5$	$\varphi = 0.4$	$\varphi = 0.3$	$\varphi = 0.2$	$\varphi = 0.1$	
FJPV_34	6/10	6/10	5/10	3/10	2/10	1/10
CASP_49	7/10	7/10	5/10	2/10	3/10	1/10
CAAR_9	6/10	6/10	6/10	3/10	2/10	1/10
BAQV_66	6/10	6/10	4/10	2/10	2/10	1/10
Average recall	62.5%	62.5%	50.0%	25.0%	22.5%	10%

Table 5
Average precision and standard deviation of LiRAM and SPA.

Precision performance of LiRAM and modified SPA						
			LiRAM			Modified SPA
	$\varphi = 0.5$	$\varphi = 0.4$	$\varphi = 0.3$	$\varphi = 0.2$	$\varphi = 0.1$	
Average precision	0.357	0.428	0,510	0.557	0.874	0.871
Standard deviation	0.068	0.099	0.036	0.099	0.068	0.130

Table 6
Average recall and standard deviation of LiRAM and SPA.

Recall performance of LiRAM and modified SPA						
			LiRAM			Modified SPA
	$\varphi = 0.5$	$\varphi = 0.4$	$\varphi = 0.3$	$\varphi = 0.2$	$\varphi = 0.1$	
Average precision	0.721	0.646	0.534	0.382	0.240	0.117
Standard deviation	0.067	0.073	0.046	0.077	0.011	0.018

Table 7
Values of CPU-times of LiRAM.

Values of CPU-times of LiRAM			
Average	Standard Deviation	Maximum	Minimum
46.06 s	50.98 s	305.45 s	6.59 s

Anova Test for the confidence interval of 0.95 and found significant differences between the applications for φ greater than 0.2, as shown in Fig. 7.

If sufficient information is available to determine more specific values for pcg, e.g., as shown in the methodology proposed in Section 4.2.2, LiRAM could generate better solutions than those obtained by using constant value (pcg = 1). E.g., for values of the Maximum Payout Share to Members φ greater than 0.2 using specific values for pcg in LiRAM leads to improved results (see Fig. 7).

If the value of φ is higher, the LiRAM can choose individuals with higher pcg, that improve its overall performance. LiRAM reaching its maximum performance for $\varphi = 0.3$. The ability of LiRAM to generate

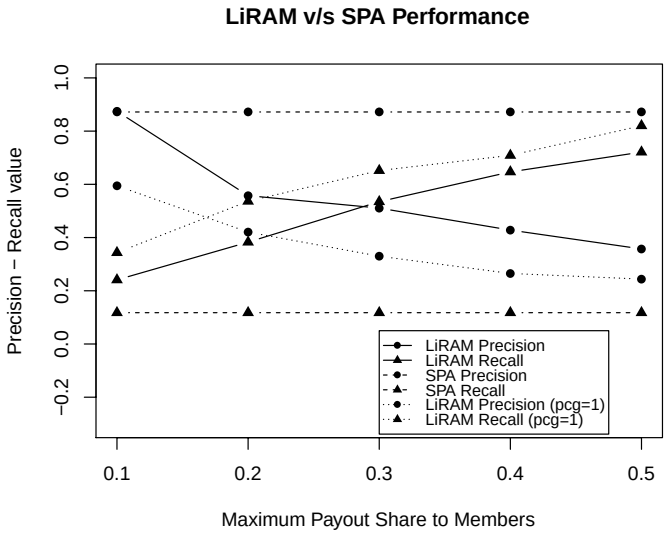


Fig. 5. Average precision and recall for LiRAM and SPA.

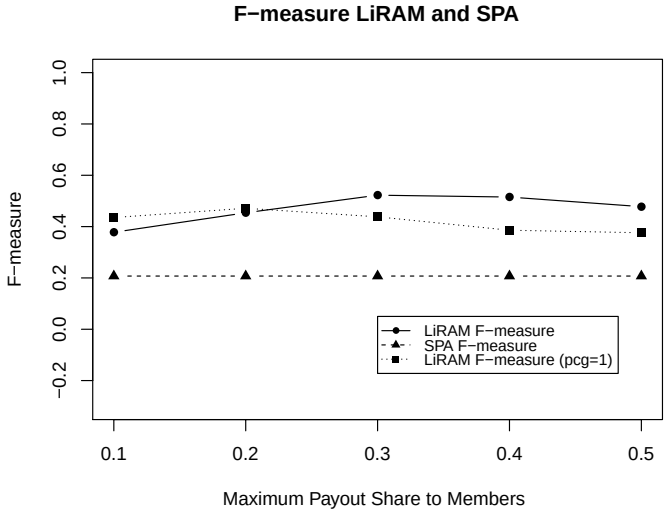


Fig. 6. Average F-measure for LiRAM and SPA.

different best groups by varying parameter φ lends a considerable measure of flexibility to the investigative process. Thus, investigations can be initiated with a small number of suspects by setting a low value of φ, which can later be increased as required, depending on the results. This characteristic has obvious potential for achieving efficiency in the assignment of investigative resources.

5. Conclusions and future research

This paper proposes a model for strengthening and enriching the existing methods of social network analysis used to determine whether

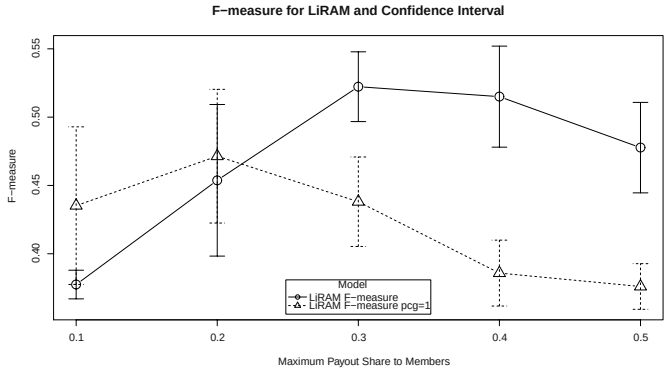


Fig. 7. Average F-measure of LiRAM for each value of φ and its Confidence Interval.

an association exists between two individuals suspected of participation in group crimes, and if so, the identity of the association's other members. Under this new approach, network nodes representing criminal group members are assigned values derived from individual attributes reflecting the members' individual criminal abilities and, therefore, their criminal propensities. The magnitudes of the links joining the members are considered to reflect their individual degrees of trustworthiness. Associations are then the result of a decision process in which a maximizing crime planner chooses the best individuals to associate with on the basis of a personal utility function that specifies the trade-off between criminal propensity and trustworthiness.

The maximization itself is determined by an ILP model denoted LiRAM (linear rational association model) that incorporates the two measures and other relevant factors. LiRAM also contains a parameter (φ) that determines the proportion of the crime proceeds the crime planner pays out to the group members as a function of their individual criminal propensities. For each parameter value, LiRAM determines the best association between any pair of suspect individuals and, therefore, the other individuals constituting the association.

The estimation of pcg in practice depends on the available information and can have other connotations as the propensity of an individual to commit crimes related to criminal groups or the propensity of an individual to have a key role in a criminal group.

The model was evaluated by application to real data. The results were compared with those obtained using a modified shortest-path algorithm. The principal findings are summarized as follows:

- LiRAM is effective in finding the best association between individuals in a social network using information about their individual attributes. The model proved to be able to find relevant associations in a network by including key individuals in the majority of the best associations identified.
- LiRAM lends flexibility to criminal investigation via its ability to generate different sets of association alternatives simply by varying the above-described parameter governing the payout (φ). Thus, an investigation can begin with a given alternative and explore others as required.
- If sufficient information is available to determine a specific value for pcg, LiRAM could generate better solutions than those obtained by using a constant value.

In a future article, we will extend the present work to include the following:

- Consider the magnitude of the link to be a probability measure expressing the likelihood that two individuals are related, assuming that these links are mutually independent events [29]. This assumption will generate a nonlinear function of the expected cost of bribes that could improve the effectiveness of the proposed model.
- If neighborhood information for suspects is also available, this information could be used to further improve the network structure and link information, e.g., as proposed in [15].
- In practice, the available information includes criminal cause, offenses, date of offenses, frequency of offenses, convictions, age, sex, among others. Considering this available information the use of logistic regression or machine learning techniques may be interesting tools to estimate *pcg*.

Acknowledgments

The authors gratefully acknowledge the support of the Santiago-based Complex Engineering Systems Institute (CONICYT - PIA - FB0816) www.isci.cl; the Anillo project ACT87 “Quantitative methods in security”; and the Ph.D. program in engineering systems at the Universidad de Chile. The first author was the recipient of a CONICYT grant number 21120226 to pursue doctoral studies in engineering systems at the Universidad de Chile. The first author acknowledges the Criminal Analysis Unit of the Public Prosecutor's Office of Región del Biobío-Chile by the dataset provided under an Internship Agreement. The second author also acknowledges financial support by FONDEF project ID16I10222, CONICYT.

Appendix A

The final form of the planner's utility function, given in the main body of the text as Eq. (14), is derived from Eq. (13) as follows:

Recall that Eq. (13)) is

$$U = I\frac{\sum_{i\in N} pcg_i Y_i}{pcg_{max}} - I\gamma\frac{\sum_{(i,j)\in A} d_{ij}X_{ij}}{d_{max}} - w\sum_{i\in N} pcg_i Y_i \tag{A1}$$

and assuming the planner is willing to pay the highest possible rate per unit of criminal ability to ensure the chosen individuals join the group, w is given by

$$w = \frac{I}{pcg_{max} - pcg_s}\left(1 - \gamma - \frac{pcg_s}{pcg_{max}}\right) \tag{A2}$$

Substituting w into the utility function, the latter becomes

$$U = I\frac{\sum_{i\in N} pcg_i Y_i}{pcg_{max}} - \frac{I\gamma}{d_{max}}\sum_{(i,j)\in A} d_{ij}X_{ij} - \frac{I}{pcg_{max} - pcg_s}\left(1 - \gamma - \frac{pcg_s}{pcg_{max}}\right)\sum_{i\in N} pcg_i Y_i \tag{A3}$$

This function can be rewritten as

$$U = I \sum_{i \in N} pcg_i Y_i \left(\frac{1}{pcg_{max}} - \frac{1}{pcg_{max} - pcg_s} + \frac{\gamma}{pcg_{max} - pcg_s} + \frac{pcg_s}{pcg_{max}(pcg_{max} - pcg_s)}\right) - \frac{I\gamma}{d_{max}} \sum_{(i,j) \in A} d_{ij} X_{ij} \tag{A4}$$

$$U = I \sum_{i \in N} pcg_i Y_i \left(\frac{pcg_{max} - pcg_s - pcg_{max}}{\sum_{i \in N} pcg_i pcg_{max} - pcg_s} + \frac{\gamma}{pcg_{max} - pcg_s} - \frac{pcg_{max} - pcg_s - pcg_{max}}{\sum_{i \in N} pcg_i pcg_{max} - pcg_s}\right) - \frac{I\gamma}{d_{max}} \sum_{(i,j) \in A} d_{ij} X_{ij} \tag{A5}$$

Simplifying, the utility function finally becomes

$$U = \frac{I\gamma}{pcg_{max} - pcg_s} \sum_{i \in N} pcg_i Y_i - \frac{I\gamma}{d_{max}} \sum_{(i,j) \in A} d_{ij} X_{ij} \tag{A6}$$

References

[1] S. Becker Gary, Crime and punishment: an economic approach, Journal of Political Economy 76 (2) (1968) 169–217.
[2] G. Bichler, A. Malm, T. Cooper, Drug supply networks: a systematic review of the organizational structure of illicit drug trade, Crime Science 6 (1) (2017) 2.
[3] S.P. Borgatti, A. Mehra, D.J. Brass, G. Labianca, Network analysis in the social sciences, Science 323 (5916) (2009) 892–895.
[4] D.A. Bright, Using social network analysis to design crime prevention strategies: a case study of methamphetamine manufacture and trafficking, Crime Prevention in the 21st Century, Springer, 2017, pp. 143–164.
[5] J.-J. Chang, H.-C. Lu, M. Chen, Organized crime or individual crime? Endogenous size of a criminal organization and the optimal law enforcement, Economic Inquiry 43 (3) (2005) 661–675.
[6] H. Chen, K. Lynch, Automatic construction of networks of concepts characterizing document databases, Systems, Man and Cybernetics, IEEE Transactions on 22 (5) (1992) 885–902.
[7] L. Ding, B. Dixon, Using an edge-dual Graph and K-connectivity to identify strong connections in social networks, Proceedings of the 46th Annual Southeast Regional Conference on XX, ACM, New York, NY, USA, 2008, pp. 475–480.
[8] L. Ding, D. Steil, B. Dixon, A. Parrish, D. Brown, A relation context oriented approach to identify strong ties in social networks, Knowledge-Based Systems 24 (8) (2011) 1187–1195.
[9] A.W. Dnes, N. Garoupa, Behavior, human capital and the formation of gangs, Kyklos 63 (4) (2010) 517–529.
[10] R. Dreżewski, J. Sepielak, W. Filipkowski, The application of social network analysis algorithms in a system supporting money laundering detection, Information Sciences 295 (2015) 18–32.
[11] E. Ferrara, P.D. Meo, S. Catanese, G. Fiumara, Detecting criminal organizations in mobile phone networks, Expert Systems with Applications 41 (13) (2014) 5733–5750.
[12] N. Garoupa, The economics of organized crime and optimal law enforcement, Economic Inquiry 38 (2) (2000) 278–288.
[13] L. Getoor, C.P. Diehl, Link mining: a survey, ACM SIGKDD Explorations Newsletter 7 (2) (2005) 3–12.
[14] R. Hauck, H. Atabakhsb, P. Ongvasith, H. Gupta, H. Chen, Using Coplink to analyze criminal-justice data, Computer 35 (3) (2002) 30–37.
[15] J.R. Hipp, C.T. Butts, R. Acton, N.N. Nagle, A. Boessen, Extrapolative simulation of neighborhood networks based on population spatial distribution: do they predict crime? Social Networks 35 (4) (2013) 614–625.
[16] D. Keatley, Pathways in Crime: An Introduction to Behaviour Sequence Analysis, Springer, 2018.
[17] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, September, J. ACM 46 (5) (1999) 604–632 ISSN 0004-5411.
[18] M. Kugler, T. Verdier, Y. Zenou, Organized crime, corruption and punishment, Journal of Public Economics 89 (9) (2005) 1639–1663.
[19] H.W. Lauw, E.-P. Lim, H. Pang, T.-T. Tan, Social network discovery by mining spatio-temporal events, Computational & Mathematical Organization Theory 11 (2) (2005) 97–118.
[20] J.M. McGloin, D.S. Kirk, Social network analysis, Handbook of Quantitative Criminology, Springer, 2010, pp. 209–224.
[21] J.S. McIllwain, Organized crime: a social network approach, Crime, Law and Social Change 32 (4) (1999) 301–323.
[22] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web. previous number = SIDL-WP-1999-0120 Stanford InfoLab, 1999.
[23] Y. Sasaki, The truth of the F-measure, Teach Tutor mater 1 (5) (2007) 1–5.
[24] M.K. Sparrow, The application of network analysis to criminal intelligence: an assessment of the prospects, Social Networks 13 (3) (1991) 251–274.
[25] R.C. van der Hulst, Introduction to Social Network Analysis (SNA) as an investigative tool, Trends in Organized Crime 12 (2) (2009) 101–121.
[26] M. Wang, W. Pan, A comparative study of network centrality metrics in identifying key classes in software, Journal of Computational Information Systems 8 (24) (2012) 10205–10212.
[27] S. Wasserman, Social Network Analysis: Methods and Applications, 8 Cambridge university press, 1994.
[28] J. Xu, H. Chen, Untangling criminal networks: a case study, Intelligence and Security Informatics, Springer, 2003, pp. 232–248.
[29] J.J. Xu, H. Chen, Fighting organized crimes: using shortest-path algorithms to identify associations in criminal networks, Decision Support Systems 38 (3) (2004) 473–487.
[30] J.J. Xu, H. Chen, CrimeNet explorer: a framework for criminal network knowledge discovery, ACM Transactions on Information Systems 23 (2) (2005) 201–226 ISSN 1046-8188.