Physica A 390 (2011) 1662–1676

Contents lists available at ScienceDirect

Physica A

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

A framework to approach problems of forensic anthropology using complex networks

Inés Caridi [a,*], Claudio O. Dorso [a], Pablo Gallo [b], Carlos Somigliana [b]

[a] Department of Physics, FCEN, University of Buenos Aires, Argentina
[b] Argentine Team of Forensic Anthropology, Argentina

ARTICLE INFO

Article history:
Received 18 August 2010
Received in revised form 11 November 2010
Available online 6 January 2011

Keywords:
Social networks
Complex networks
Forensic anthropology

ABSTRACT

We have developed a method to analyze and interpret emerging structures in a set of data which lacks some information. It has been conceived to be applied to the problem of getting information about people who disappeared in the Argentine state of Tucumán from 1974 to 1981. Even if the military dictatorship formally started in Argentina had begun in 1976 and lasted until 1983, the disappearance and assassination of people began some months earlier. During this period several circuits of Illegal Detention Centres (IDC) were set up in different locations all over the country. In these secret centres, disappeared people were illegally kept without any sort of constitutional guarantees, and later assassinated. Even today, the final destination of most of the disappeared people's remains is still unknown. The fundamental hypothesis in this work is that a group of people with the same political affiliation whose disappearances were closely related in time and space shared the same place of captivity (the same IDC or circuit of IDCs). This hypothesis makes sense when applied to the systematic method of repression and disappearances which was actually launched in Tucumán, Argentina (2007) [11]. In this work, the missing individuals are identified as nodes on a network and connections are established among them based on the individuals' attributes while they were alive, by using *rules* to link them. In order to determine which rules are the most effective in defining the network, we use other kind of knowledge available in this problem: previous results from the anthropological point of view (based on other sources of information, both oral and written, historical and anthropological data, etc.); and information about the place (one or more IDCs) where some people were kept during their captivity. For these best rules, a prediction about these people's possible destination is assigned (one or more IDCs where they could have been kept), and the success of the prediction is evaluated. By applying this methodology, we have been successful in 71% of the cases. The best rules take into account the proximity of the locations where the kidnappings took place, and link events which occurred in periods of time from 5 to 7 days. Finally, we used one of the best rules to build a network of IDCs in an attempt to formalize the relation between the illegal detention centres. We found that this network makes sense because there are survivors' testimonies which confirm some of these connections.

1. Introduction

In recent years, networks emerged as an invaluable tool for describing complex systems from very diverse areas of science, which include communication, ecosystems, biochemical, and social systems. Diverse studies suggest that properties

* Corresponding author.
E-mail address: ines@df.uba.ar (I. Caridi).

0378-4371/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2010.11.042

of networks, such as community structure and hierarchical organization, could help to explain the behavior of those systems. For example, groups of strongly connected nodes are found to correspond to known functional units on the network [1–4], or the community structure of school friendship networks is a key issue to characterize this real system [5], etc. Sometimes, links on networks correspond to explicit interactions among nodes; other times, they try to capture the implicit interactions, or the underlying connections among individuals in a system. A usual problem when working with a real database is the lack of information, which results in networks with missing links. Recently, Clauset et al. [6] developed a general technique for inferring hierarchical structure from network data, and they used it to predict missing connections on a network in the real world. While this work combines networks with statistical inference techniques, other works combine networks with information theory to characterize the community structure of the net [7,8]. The objective of this work is to provide a framework of complex networks for the analysis of data related to a forensic anthropological problem.

We will be dealing with the data collected and previously analyzed by the Argentine Team of Forensic Anthropology (EAAF [9]) regarding the disappearance of people during the dictatorship that ruled Argentina de facto during the period 1976–1983; however, it should be kept in mind that those disappearances had started more than a year before the military coup. In this work, the data is mapped on to a network, and clusters of strongly correlated individuals (nodes) are identified.

Since 1985 the EAAF has worked on identifying the human remains of those people who were assassinated or disappeared from 1974 to 1983 in Argentina. Although this has been their primary objective, they have also been trying to find out what happened to them after their kidnapping: where they were kept during their captivity (illegal detention centre, henceforth IDC), when they were relocated to other centres, when they were murdered, and where their human remains are now. A further issue EAAF has dealt with is which IDCs made up a single interconnected circuit. To solve this puzzle, EAAF has gathered abundant data from very diverse information sources. The organization of the data to guide the searches is a complex problem which EAAF has been working on for more than 25 years. They have built a database in which every missing individual has certain attributes, including identity, workplace address, residence address, place and date of disappearance, political affiliation, profession, etc. Many of these attributes are likely to relate individuals to one another. EAAF has found that the determination of groups of closely related people is crucial in searching the destination of these people, and in the possible identification of their human remains.

The underlying hypothesis driving this work is that strongly correlated subsets of people ended up being kept in the same IDC during their captivity, and afterwards were probably murdered in the same place and at the same time, a fact which might indicate that their human remains could have shared the same final destination. This hypothesis is based on the fact that the repressive forces were organized in well-defined military areas, each of them responsible for a single IDC circuit and with well-defined areas of influence. Thus, the determination of strong correlations between people leads to prioritizing certain places over others when carrying out the searches. This research is also based on information arising from the testimonies of survivors who have stated where they were kept during their captivity, and who else they saw in each of the IDCs. These investigations carried out by EAAF have resulted in almost 500 identifications of human remains [10].

In this work we intend to trace the underlying network of relations amongst disappeared individuals until the time of their kidnapping. Nodes represent individuals, and the network is built by means of the definition of "rules" which relate ("link") nodes according to the values of a subset of attributes which we consider to be the most relevant ones. In this way a set of rules is defined, the corresponding network is built and clusters are recognized.

The task is to find the most useful rule or rules in the determination of the sets of correlated individuals. For this purpose, different quality measures are proposed for and applied to the data. To this end, we used previously obtained results from the anthropological point of view and testimonial data about the place where certain people were kept captive. Once the best sets of rules had been chosen, we made a prediction about the place of captivity and we compared these results with actual data.

This paper is divided into seven sections. Section 2 will describe the available data, that is to say, the database including the attributes of disappeared people, solved cases in Tucumán, and the information on the place where some of these people were kept. Section 3 will present the applied methodology: the construction of social networks using rules which link people's attributes, and the way to determine the cluster structure on the net. Section 4 will describe the methodology to infer the best rules using knowledge and information which was not previously used, through the application of a framework of information theory and hypothesis testing. Also, we will predict the place where some people were kept (IDC) by using one of the best rules, and we will estimate the success rate of these results. Section 5 will describe the application of this methodology to the Tucumán problem, and the analysis and validation of the results obtained. Section 6 will deal with the network of IDCs of Tucumán, related to the possible circuits of illegal detention centres. Our conclusions will be presented in Section 7.

2. The actual problem at hand

In this section we will describe all the information available. The data we have analyzed correspond to the events that took place in the provincial state of Tucumán between 1974 and 1981, including $N = 1036$ cases. The set of data we worked on includes individuals who had been kidnapped, or worked, or lived, or had been born in Tucumán.

2.1. The attributes characterizing the individuals

The information collected by EAAF is used to build a social network where the missing people are represented by nodes. Each node is characterized by a series of attributes, including identity, workplace, address, place and date of kidnapping, political affiliation, etc. However, this information is not complete. On the one hand, not all the repression cases were reported (specially those of illegal detention for very short periods of time); on the other hand, for the actually reported cases, not all the relevant attributes are known.

Let i denote the nodes with $i = 1, \ldots, N$, and N, the number of nodes under analysis. For each node i, its attributes are denoted by $\overrightarrow{a_i} = (a_i^1, a_i^2, \ldots, a_i^J)$. $a_i^j = 0$ denotes the absence of information about the attribute j for the node i. Thus, J is the number of relevant attributes. We have chosen the following concepts as relevant attributes: *political* (a_i^1: political affiliation), *temporal* (a_i^2: date of kidnapping or murder) and *geographical* (for example: a_i^3: the zip code of the home address, a_i^4: the zip code of the workplace, and a_i^5: the zip code of the place of kidnapping). This information keeps changing as EAAF team collects new data. For the specific example which we describe in this work, the percentage of known information for each of the relevant node attributes is shown in Table B.1.

2.2. Information on the illegal detention centres (IDCs)

The individuals who were victims of the *Terrorismo de Estado* during the period previously mentioned can be classified into three main sets. One set corresponds to individuals who were kidnapped and eventually set free (a small percentage of the cases), another set is that of individuals who were murdered after their disappearance and whose remains were returned to their families; finally, the third and main set includes individuals who were victims of the so-called *traslados* (relocations). This means that they were killed after their kidnapping, but the final destination of their remains is still unknown in most cases. This population makes up the group who is historically known as ''the disappeared people''.

Information on the IDCs can be obtained from the testimonies of released people, who might also have revealed who else was held with them. An individual may not necessarily have been kept in a single IDC. However, the information collected is partial and, in fact, the actual number of IDCs where people were kept might never be known.

We have found it useful to associate each node i a vector $\overrightarrow{y_i} = (y_i^1, y_i^2, \ldots, y_i^{T_{idc}})$ which contains the information regarding the IDCs, i.e. if a missing individual was seen in a given IDC (say the k IDC) or if this individual was set free and acknowledged having been held in the said IDC, the vector will have $y_i^k = 1$. This vector has as many components as IDCs have been detected in this region. The possible values of k are $k = 1, \ldots, T_{idc}$ (T_{idc} being the number of IDCs in the studied area). This vector is known (i.e., it has at least one component different from 0) for only a subset of the nodes, and for the time being, only about 42% of the nodes belong to this subset. It should always be kept in mind that, as there is missing information, a null vector only means that there is yet no information corresponding to that particular individual. The number of nodes with information on IDCs is denoted by $\mathcal{N}^{idc}$.

It will be useful to define vector $\overrightarrow{N^{idc}}$:

$$\overrightarrow{N^{idc}} = (N_1, N_2, \ldots, N_{T_{idc}})$$

where N_k is the number of nodes corresponding to individuals who have been seen in the k-th IDC:

$$N_k = \sum_{i=1}^{N} y_i^k.$$

2.3. Previous results from an anthropological point of view: reference clusters

As we have mentioned in the introduction, EAAF has found that the determination of groups of closely related people is crucial. In this sense, EAAF has been developing an interdisciplinary research method to identify some of these *''succession of kidnapping episodes''* which were probably linked taking into account both their knowledge of the dynamics of the systematic method of repression and disappearances practised in Argentina [11], and diverse information sources. In brief, the sources used in this process can be classified into two categories: the written ones (declassified documents, journalistic information, judicial files, state documents such as death certificates and cemetery records), and the oral ones (survivors' testimonies, interviews to government officers, victims' family members, friends, and comrades).

For the sake of our study, these *''succession of kidnapping episodes''* will be referred to as ''reference clusters'', which will denote a set of individuals whose disappearances were (probably) strongly correlated. The EAAF has been able to identify 12 of those ''reference clusters'' comprising a total of 64 individuals as taking place in Tucumán. In the framework of this study, we will consider that such reference clusters correspond to real historical events, and as such, they can provide useful validation information to develop this new methodology.

The information about the reference clusters is contained in vector $\overrightarrow{r} = (r_1, r_2, \ldots, r_N)$, in such a way that $r(i) = r_i$ is the reference cluster to which the i node belongs to. $r_i = 0$ means that the i node does not belong to any reference clusters. The possible values that r_i can take are: $\{0, 1, 2, \ldots, R_T\}$ where R_T is the total number of reference clusters. In this particular case, $R_T = 12$, and the total number of nodes belonging to these reference clusters is 64.

Analogous to Section 2.2, it will be useful to define vector $\overrightarrow{N^r}$:

$$\overrightarrow{N^r} = (N_1^r, N_2^r, \ldots, N_{R_T}^r)$$

where N_j^r is the size of the j-th reference cluster (this information will be explicitly used in Section 4.1).

3. The methodology

We will build a social network in which the individuals are represented by nodes and the relationships among them by links. In this first approach, the links will be undirected and unweighted. The presence of links between two nodes will depend on the known attributes of the nodes. The existence of the link is then determined by "rules".

3.1. Building the network

There will be a link between a pair of nodes i and j if the corresponding attributes of the nodes described by $\overrightarrow{a_i}$ and $\overrightarrow{a_j}$ satisfy a *rule*, that is to say, a condition which involves attributes of the following nature: (a) political (a_i^1); (b) temporal (a_i^2). (Not only the disappearance dates of the people involved, but also those relevant to the historical context of the region under study.) (c) geographical (a_i^3, a_i^4, a_i^5). (Not only the zip codes of each of the locations in the region, but also relevant information on the distances between locations (a matrix of adjacencies of the zip codes).)

Rule R is defined by specifying a set of functions $F_{ij}, F'_{ij}, F''_{ij}$, where each of the functions defines if there will be a link between a pair of nodes (i and j) according to how much is known about the political attribute (a_i^1, a_j^1, the attribute with the highest degree of missing information in terms of the general bulk of information). When a_i^1 and a_j^1 are known, function F_{ij} will be applied; if one is known and the other one is not, function F'_{ij} will be applicable; when both are unknown, function F''_{ij} will be used. The main idea, previously developed in Ref. [12] is based on the construction of sets of links at different stages of the process by applying rule R: the greater the ignorance of the political attribute, the stricter the conditions with regard to the temporal attribute.

For example, given a pair of nodes i and j for which all the relevant attributes are known, a link exists between i and j if the two nodes share the same political affiliation, and the dates of disappearance differ in less than Δ_1 days and they meet a given condition regarding the geographical attributes. Formally, a link exists if F_{ij} attains value 1,

$$F_{ij} = \delta(a_i^1 - a_j^1) \times \Theta(\Delta_1 - |a_i^2 - a_j^2|) \times F_{ij}^g \tag{1}$$

where $\delta(x)$ is the Kronecker's delta which takes value $\delta(x) = 1$ if $x = 0$, and $\delta(x) = 0$ in other cases. $\Theta(x)$ is the Heaviside step function, which becomes $\Theta(x) = 1$ if $x \geq 0$ and $\Theta(x) = 0$ in other cases. F_{ij}^g is a function which takes into account the geographical information. For example, we have found it useful to use $F_{ij}^g = \max\{\delta(a_i^3 - a_j^3), \delta(a_i^4 - a_j^4), \delta(a_i^5 - a_j^5)\}$, that is to say, $F_{ij}^g = 1$ if at least one of the zip codes of these nodes is the same; but we have also analyzed other geographical functions, such as $F_{ij}^g = \delta(a_i^l - a_j^l)$ with $l = 3$–5, which means that a link exists when a specific zip code l is the same.

If for one element in a pair of nodes, there is no information about the political affiliation attribute while all the other relevant attributes for both are known, we require a stronger condition than in the previous case to establish a link between these two nodes. Formally, there will be a link between the two nodes if $F'_{ij} = 1$,

$$F'_{ij} = \Theta(\Delta_2 - |a_i^2 - a_j^2|) \times F_{ij}^g$$

where Δ_2 should be less than Δ_1. Finally, there will be a link between two nodes for which the a^1 attribute is not known, according to

$$F''_{ij} = \Theta(\Delta_3 - |a_i^2 - a_j^2|) \times F_{ij}^g$$

with $\Delta_3 < \Delta_2 < \Delta_1$. In all the stages, the same geographical function F_{ij}^g is used.

At this point it must be stated that there are special cases in which a link is established between two nodes taking into account only $\Theta(\Delta_1 - |a_i^2 - a_j^2|) = 1$. This is so when there is another source of strong correlation like, for example, that these two nodes are close relatives.

According to this, the rules will differ in the specific definition of F_{ij}^g and in the value of the parameters Δ_3, Δ_2, Δ_1.

We have also explored the possibility that rules are heterogeneous in time, that is to say, temporal parameters Δ are different before and after a certain historically relevant date. (y sacar despues.)

Up to this point, all the rules displayed use the equality of (sacar the) zip codes to establish a link, but the geographical distance between neighboring locations has also been used. Regarding neighborhoods, we have defined the matrix of adjacencies A^l, whose elements meet the condition that $A_{ij}^l = 1$ if a_i^l and a_j^l are less than $u = 17$ km apart; and $A_{ij}^l = 0$ otherwise. The index l indicates that the zip code used to define the adjacency matrix is that corresponding to attribute l (for example, A_{ij}^5 represents the adjacency matrix taking into account the zip codes where the kidnappings of nodes i and j took place). This 17-kilometre threshold distance has been chosen because Tucumán state is rather small, and for all relevant locations, the nearest neighboring town lies within this distance. An example of possible geographical functions which used information of neighborhoods is $F_{ij}^{g} = A_{ij}^{l}$ with $l = 5$.

All functions and parameters defined so far can be combined at the different stages to generate rule R.

3.2. Building the clusters

Once the network is built, clusters are recognized. For this purpose a simple clustering algorithm is used. Node j will belong to cluster C if there is a node i (belonging to C) which it is linked to.

$$j \in C \iff \exists i \in C / \max\{F_{ij}, F'_{ij}, F''_{ij}\} = 1.$$

When a given rule is applied, we get a cluster structure which can be characterized through vector $\overrightarrow{b^R} = (b_1^R, b_2^R, \ldots, b_N^R)$ where $b^R(i) = b_i^R$ denotes the index of the cluster which node i belongs to. b_i^R can take values $\{1, 2, \ldots, T_b^R\}$, where T_b^R is the total number of clusters resulting for rule R.

Using the above-mentioned information, it is possible to calculate size $S(j)$ of cluster j.

$$S(j) = \sum_{i=1}^{N} \delta(b_i^R - j).$$

4. Inferring the best rules

Taking advantage of the knowledge of the reference clusters and the information the nodes with known IDC, we will look for the best rules.

As stated above, given a rule, a set of links results and a cluster structure can be recognized. Many rules can be built and the best ones should be looked for. In order to recognize the best rules, we use a set of tests and functions as explained below. The information available can be summarized as follows

$\overrightarrow{a_i}$: the set of attributes of node i.

$\overrightarrow{y_i}$: the IDC vector associated with node i, whose components are associated to each IDC and contain the information of whether the node was seen in such IDC.

r_i: reference cluster of node i.

As described in Section 3, the information $\overrightarrow{a_i}$ (for all the nodes) is used to define the connections between the nodes on the net, together with a rule; the cluster structure is thus generated. In this section, we will use part of the information included in vector $\overrightarrow{y_i}$, together with $\overrightarrow{r}$, to find the best rule or rules to define the net.

4.1. Using the reference clusters

We will establish that a good rule should be such which properly recognizes the reference clusters, i.e. not only should all members of a reference cluster remain together in a given cluster recognized by the clustering method, but also, no member of other reference clusters should belong to it. In other words, a good rule neither breaks reference clusters nor joins them. Because the information at hand is incomplete, the recognized clusters might contain a reference cluster plus extra nodes. In this way, the reference clusters can only appear as a subset of nodes of a recognized cluster and the recognized cluster should not contain nodes belonging to more than one reference cluster.

The quality functions which will be used are based on the conditional entropy $H(z \mid x)$, given two random variables, x and z [13,14]. In Appendix A we will detail some information about the conditional entropy and about variables r and b^R. If b^R is the variable associated with the cluster of the network resulting after implementing rule R, and variable r is associated with the reference cluster which a node belongs to, then:

(a) $H(b^R \mid r) = 0$ means that the nodes belonging to a given reference cluster appear together in a single cluster resulting from the application of rule R. In this case, the reference clusters are not broken.
(b) $H(r \mid b^R) = 0$ means that nodes that belong to the same cluster resulting from the application of rule R belong to the same reference cluster (if they do belong to any of the reference clusters). In this case, the reference clusters do not join.

Thus, we define the following functions F_1 and F_2 to be minimized at the same time.

$$F_1 = H(r|b^R); \qquad F_2 = H(b^R|r).$$

4.2. Using the IDC information

When we look at the nodes in a cluster and we focus on those nodes which have information about the IDCs, we expect that, according to the hypothesis that nodes in a cluster share the same final destination, they have been seen in the same IDC or the same circuit of two or more IDCs. The task is to recognize when this occurs for a given cluster, and to identify this cluster as a *"succession of kidnapping episodes"*. In order to associate a cluster with a *"succession of kidnapping episodes"* we need the information regarding IDC in a cluster to be meaningful.This means that the total number of nodes with the same

IDC is highly unlikely to be obtained using random sampling. This allows to associate an IDC to cluster j, and this IDC label can eventually be used as a prediction for those nodes, belonging to cluster j, whose IDC is unknown. A destination prediction will be assigned only to those nodes belonging to the clusters which have been recognized as a succession of kidnapping episodes on the basis of the hypothesis test. We should keep in mind that the presence of nodes with no information about IDC should not be penalized, as it only reflects our lack of knowledge. The diagram in Fig. B.1 shows the tests to be implemented to obtain meaningful clusters.

Detailed descriptions of the implementation of these tests are presented in Appendix B.

To perform this calculation only 77% of the information on the IDCs was used (this percentage was randomly selected, and the remaining information was used in the final validation test).

4.3. The predictive power of a rule

As our objective is to predict the IDCs where the disappeared individuals were held, we must be able to quantify the prediction capability of a rule. For this purpose we find it convenient to define the following variables which depend on the implemented rule, and on the known information about IDCs for the nodes. At this stage of the work, it is important to remember that the information used corresponds to 77% of the cases involving known IDCs.

First, we define N_p, the number of nodes which the rule makes an IDC prediction for (either a single IDC or a circuit). Next, we define N_c (correct nodes), which is the number of nodes which the IDC prediction has been correct for (that is to say, the IDC where the individual was seen is the one predicted for that cluster). On the other hand, when the prediction is wrong, the node will be one of the N_i (incorrect) nodes. Finally, N_u stands for the number of nodes which an IDC prediction is made by means of the application of the rule, but for which we lack information on the IDC (either because the information on the IDC has been saved for the next validation step of this work, or because it is truly unknown). Then, $N_u = N_p - N_i - N_c$.

One variable of interest is $P_r\,|_R$, i.e. the prediction rate for rule R:

$$P_r\,|_R = \frac{N_p}{N}.$$

We now define $S_r\,|_R$ as the success rate of the rule, understood as the number of nodes correctly predicted, with respect to the total number of nodes with known information for which the rule makes a prediction. Variable $S_r\,|_R$ reflects the fraction of nodes whose prediction is correct after the application of a given rule R with respect to the total number of nodes belonging to clusters recognized as *"successions of kidnapping episodes"* and with known information on IDCs. Then, $S_r\,|_R$ will be:

$$S_r = \frac{N_c}{N_c + N_i}.$$

5. Analysis of events in Tucumán

We will now apply the methodology described in the previous sections to the data corresponding to Tucumán, the geographical region of interest.

In Fig. B.2, we show the number of disappearances, murders, etc, in the region of interest as a function of time.

5.1. Homogeneous rules in time and subsequent analysis

The rules applied to define the networks combined both geographical functions as described in Table B.1, and temporal parameters $\Delta_2 = 2, \ldots, 11$ and $\Delta_3 = 2, \ldots, 11$, with the restriction $\Delta_1 > \Delta_2 > \Delta_3$.

First, rules will be analyzed in terms of reference clusters. Fig. B.3 shows functions F_1 and F_2 for rules which are homogeneous in time.

As defined in the previous section, F_1 is a quality function which detects if the rule joins nodes belonging to different reference clusters, and F_2 detects if the rule separates nodes belonging to the same reference cluster. We have limited the region of acceptable rules to those that meet the condition $F_1 < 0.2$ and $F_2 < \Delta F_2 = 0.45$. This restriction imposed on the rules can be interpreted as follows:

- $\Delta F_2 \simeq 0.45$ is the value that F_2 takes when only one node among half of the reference clusters (those of bigger size) is wrongly classified.
- the condition that $F_1 < 0.2$ guarantees that at most two clusters (each being size 6 maximum) be joined.

It is possible to note that the rules which do not use geographical data gather unrelated reference clusters (F_1 is large), regardless of the temporal parameter used. This implies that the geographical information must be used to define the links on the net as it is a relevant attribute in the Tucumán region.

As seen in Table B.2, rules of type 1, 2 and 4 do not satisfy the reference clusters test for any temporal parameters. Among the rules which actually satisfy that criterion (namely rules 3, 5 and 6), rule number 6, which incorporates the usage of the adjacency matrix with a threshold of 17 km, seems to be the most robust one because it tolerates greater variability in terms of temporal parameters.

In Fig. B.4, we show the success rate (S_r) vs the prediction rate (P_r) for rules type 3, 5 and 6 using the temporal parameters displayed on Table B.2. We can see that the general trend is that for a given kind of rule the larger value P_r is, the lower S_r is. Looking at the figure, it is easy to realize that the triangles (rules type 6) are the best ones, because they attain the highest values of S_r while attaining very high values of P_r. As a conclusion, we choose those which use the geographical function of proximity as the set of the best rules, and among these, those of temporal parameters $\Delta_1 = 5$, $\Delta_2 = 3$, $\Delta_3 = 1, 2$. It is important to remark that when $\Delta_3 = 1, 2$, the same values of S_r and P_r arise although the cluster structure is not the same. In all cases, a value of $\alpha' = 0.17$ is considered when implementing the tests. The act of rejecting the independence leads us to consider this cluster as a *"succession of kidnapping episodes"*, and to assign it a prediction, that is to say an IDC destination to all cluster members. It should be kept in mind that this information will be further analyzed by the anthropologists of EAAF, who will apply other sources of information and their own expert knowledge to pursue the analysis.

In Fig. B.5 we show the performance of the homogeneous rule R of temporal parameters $\Delta_1 = 5$, $\Delta_2 = 3$, $\Delta_3 = 2$ and the geographical function of proximity, with respect to the reference clusters. In that Figure, only a network of 105 nodes is shown. This set of nodes comprises all the nodes belonging to the reference clusters (color circles, 64 nodes) and those nodes which, though not belonging to any reference clusters, are part of network clusters in which there is at least one node belonging to some reference cluster (black circles).

5.2. Exploring inhomogeneous rules

Also we have explored inhomogeneous rules which change the temporal parameters and geographical functions at a given date: $T_1 = 1/1/76$ because of its historical relevance (date when Bussi took over the repressive operations in Tucumán). The reason why we have used inhomogeneous rules is that we wanted to contemplate the possibility that the repressive methodology had changed at that moment, and that in turn, the rule to define connections on the network had accordingly changed.

We have used the same geographical function which gave the best results in the case of homogeneous rules. We now impose the same conditions as in the homogeneous case, i.e. $F_1 < 0.2$ and $F_2 < 0.45$. The results of such a calculation are displayed in the inset of Fig. B.3. We now perform the success and prediction rate analysis for these rules. We can see in Fig. B.6 that upward-pointing triangles, which represent rules with $\Delta_1^b = 7$, are as least as good as the best homogeneous rules. We have chosen this rule which, before T_1, uses the temporal parameters $\Delta_1^b = 7$; $\Delta_2^b = 4$; $\Delta_3^b = 2$; and after T_1, $\Delta_1^a = 5$; $\Delta_2^a = 3$; $\Delta_3^a = 2$ with the same geographical function. This is one of the rules which is (in S_r and P_r graph) at the minimum distance from the point $S_r = 1$ and $P_r = 1$.

Even though we have been able to find an inhomogeneous rule which turns out to be the best one, this rule is only slightly better than the best homogeneous one. This slight improvement has been obtained at the expense of doubling the number of parameters to be varied, and such an effort might not be justifiable in general. Nevertheless, further studies are needed to reach a final conclusion regarding this point.

5.3. Evaluation of the prediction using reserved information

We now evaluate the quality of the prediction of IDC assignment using the subset of nodes which included IDC information and had been reserved for this validation stage. This subset includes 100 nodes, i.e. 23% of the IDC information available.

We have evaluated which IDC had been assigned by the method and we later compared it with the one which was known for the individual in question. For this purpose, we have defined some magnitudes to study the obtained results. Let us suppose both that the number of nodes for which we reserve IDC information is made up of N_d nodes, and that the method makes a prediction for N_p. Then the prediction rate and the success rate will be:

$$P_r = N_p/N_d$$
$$S_r = \sum_{N_d} \sum_{k=1}^{12} \delta(b_i^R(i) - y_i^k)/N_p.$$

For the heterogeneous rule in time with temporal parameters $\Delta_1^b = 7$; $\Delta_2^b = 4$; $\Delta_3^b = 2$ before T_1; $\Delta_1^a = 5$; $\Delta_2^a = 3$; $\Delta_3^a = 2$ and after T_1, with the same geographical function type 6, we obtained a success rate (S_r) of 71%, and a prediction rate (P_r) of 41%. On the other hand, when the same calculation is performed applying the homogeneous rule which uses the same geographical function type 6 and temporal parameters $\Delta_1 = 5$; $\Delta_2 = 3$; $\Delta_3 = 1, 2$ (described in subsection B), we obtain $S_r = 67\%$ and $P_r = 36\%$.

Finally, we compared the results obtained through this methodology using the best inhomogeneous rule to two different strategies of random assignment of IDC destination. This random assignment was performed on those nodes whose IDC information had been reserved. In the first case, we applied a uniform distribution of probability to make the assignments (i.e. $p_k = \frac{1}{12}$, where $k = 1, \ldots, 12$ is the probability to assign the k IDC to each of the nodes whose IDC information had been reserved). In the second case, the probability distribution to make assignments was not uniform, given by $p_k = \frac{N_k}{\mathcal{N}^{idc}}$, where

N_k is the number of individuals who were seen in the k-th IDC, and $\mathcal{N}^{idc}$ is the total number of individuals with known IDC information (for this purpose, the set of 77% of the information of IDC was taken into account). After running 200 realizations of both strategies, we concluded that the results of randomly uniform assignment coincided with the IDC assigned to the cluster in (0.09 ± 0.04)% of the predicted cases (i.e. those N_p nodes for which the best rule made a prediction). And when we used a non-uniform distribution strategy to make predictions, the results coincided in (0.26 ± 0.03)% of the predicted cases. It is clearly seen that the correlations recognized by the method described in this paper render different assignments of the IDCs compared to those provided by the random label one.

6. Clandestine circuits network

Using the rule referred to in previous sections, we analyze correlations among IDCs. We will say that if multiple destinations including two or more IDCs are assigned to a given cluster, then these IDCs are correlated, and then we can show this relation as a graph of linked IDCs.

This network may help to answer the question about which IDCs made up a single interconnected circuit (see introduction). Nodes represent IDCs, and we use the previously obtained results to establish the connections on the IDCs network. Whenever a cluster identified as a *succession of kidnapping episodes* includes these two IDCs as its destination, we propose a connection between the said IDCs as a "clandestine circuit".

In Fig. B.7(a) we present the resulting network where we use the complete information on IDCs to construct it. When there is a connection between two IDCs, it is very likely that there was a circuit between these two IDCs during a particular period, that is to say, the link may represent an operational and functional connection, which may have caused individuals to be moved from one centre to the other.

The connections on the net are undirected because the IDC information is atemporal (the moment when the individual was kept in that IDC is unknown). It is possible, however, for the EAAF researchers to assign an arrow to the connections on the net based on some other kind of information. For example, they have clues which suggest that during certain periods the moves involved transporting individuals from rural to urban areas.

It is very interesting to compare EAAF's knowledge on the clandestine circuit (which is shown on the network of Fig. B.7(b)) with that of the IDC network which stems from the use of this methodology and the information already available.

Information on the Tucumán IDCs obtained by EAAF:

It is possible to divide all the IDCs recorded in the state of Tucumán into two subsets according to geographical characteristics. On one hand, a number of rural IDCs were set up to centralize the repressive operations against both rural movements and the guerrilla acting in the jungles of Tucumán. On the other hand, a number of urban IDCs were identified in the area of San Miguel de Tucumán and its neighbouring villages. These illegal urban centres aimed at maximizing the repression on and the kidnapping of the individuals living in this area, mainly students and factory workers.

While the IDCs *Jefatura de Policía, Arsenal* and *Reformatorio* were located in Tucumán City, *Escuelita de Famaillá*, *ex Ingenio Nueva Baviera* and *Base Ex ingenio Santa Lucia* were located in the rural area. Moreover, *Famaillá and Ex Ingenio Nueva Baviera* were actually in Famaillá itself, on National Road 38, approximately 40 km from San Miguel deTucuman. *Ex Ingenio Santa Lucía*, on the other hand, was not located on the above-mentioned road, but somewhere in the area of the pre-jungle, some 10 km from Acheral (which is 15 km from Famailla). It is easy to see that the Circuit obtained using the new methodology is much richer than the one proposed by EAAF. In particular the main rural IDCs previously mentioned, turn out to be "linked" (i.e. correlated). On the other hand, an IDC located in a neighboring state (Jujuy) turns out to be part of the circuit. It is also interesting to notice that only one IDC remains unconnected (ex Ingenio Lules), but this was an IDC which was active for a short period of time and there is not much information available on it.

7. Conclusions

This work deals with a real social problem from the point of view of complex networks. Its objective is to obtain information to help answer both specific questions about disappeared people, such as the place where they were kept in Tucumán, and general questions regarding the circuits of illegal detention centres (IDCs) operating in Tucumán.

In order to answer these questions, it is necessary to detect the set of nodes (individuals) which are strongly connected, once the network is formalized. The fundamental hypothesis is that these people shared the same specific destination, and their human remains might be located in the same place. This is why detecting strongly related sets becomes crucial in the difficult task of searching for the place where these people were kept captive, and collecting the necessary information which might lead to the identification of their remains. This project has become an attempt to formalize the methods which anthropologists of EAAF have formulated in the framework of complex networks to systematically explore the rules which link nodes, and to analyze the rules which are the most suitable for this context in order to apply new data.

We have developed a methodology which allows us to use the available information to detect the subsets of strongly connected people as clusters on a network. The information used to build the net involves a number of attributes which individuals bore until the time of their disappearance. In this way, nodes represent individuals, and links arise from manipulating (relating) their *life attributes* by applying *rules*. There is other available information which we have used to find the best rules to define the network: the information on the cases involving people whose IDC was known, and the reference clusters (which summarize the anthropologists' knowledge on solved cases). Thus, we may say that this work is interdisciplinary not only because it deals with a real social problem, but also because it uses previously obtained results from anthropological studies in order to obtain the best solution to the problem. Discovering the best rules to define the network will lead to make the best possible prediction.

We have shown that among the rules which take into account a single zip code (be it the one corresponding to the home, the workplace, or the kidnapping location), the ones which use the home or workplace zip code are not acceptable. Nevertheless, there are rules which take into account the three zip codes to define the connection (function type 5) and which are acceptable for a few parameters. But, if we consider the success rate, these rules turn out to be clearly worse than those which use only the zip code of the area where individuals were kidnapped (function type 3), or those which use the neighboring ones (function type 6). This means that it is the location where the disappearance took place, rather than the work or home ones, which is the most important piece of geographical information to link facts.

We have explored two kinds of rules: homogeneous and inhomogeneous in time. For the first type of rules, we have chosen those with temporal parameters $\Delta_1 = 5$, $\Delta_2 = 3$, $\Delta_3 = 1, 2$ and with the geographical function which considers the proximity of the locations where the kidnappings took place. For the second type, we have chosen those with temporal parameters $\Delta_1 = 7$; $\Delta_2 = 4$; $\Delta_3 = 2$ before T_1; and $\Delta_1 = 5$; $\Delta_2 = 3$; $\Delta_3 = 2$ after T_1, with the same geographical function; for this region, T_1 is a relevant date from the historical point of view.

The methodology offers a new viewpoint of the problem, and generates new hypothetical relations to be explored. Additionally, it is possible to make a prediction about the missing information on a specific individual (such as the IDC or IDCs where he or she was kept), and to formalize specific questions which EAAF may ask some people about in future interviews.

Another question of interest in this problem is which IDCs belong to the same circuit of detention centres in Tucumán. The results obtained through this methodology were used to build an IDC network representing the possible circuits between IDCs. We have found that various testimonies and sources of information provided by EAAF show that the connections on this network make sense in the context of the events which took place in Tucumán.

This methodology makes a prediction whenever there is enough information regarding a given individual and his or her social context. This is why it also allows us to discover which sectors of the database require further collection of information through interviews, because if more information were obtained, it would be possible to define whether a certain cluster is a *succession of kidnapping episodes*. This has to do with shedding light on what information can lead us to make more predictions.

This work opens a new perspective in the analysis of datasets with incomplete information and mixed metrics (as is the case of the disappeared people in Tucumán), and can probably be used in other regions where the same kind of repressive methodology was applied. In particular we are planning to extend this kind of analysis to other regions in Argentina during the same period of time where the political and geographical conditions were very different, namely the city of Buenos Aires.

Moreover, we believe that this approach can be extended to any problem of detection of structures in datasets with missing information if the problem at hand is characterized by a set of attributes in which the absence of information can be overidden by putting more stringent conditions on the relations among other attributes. For example, in our work, when there is no information on the political affiliation of a given node, a link with another one can be established if the events under consideration took place very close to one another in space and time.

Acknowledgements

We thank P. Balenzuela and A. Chernomoretz for their enriching comments on this work.

Appendix A. Using information about reference clusters

If variable x can take the values $\{x_1, x_2, \ldots\}$ with probabilities $\{p(x_1), p(x_2), \ldots\}$, and the variable z can take the values $\{z_1, z_2, \ldots\}$ with probabilities $\{p(z_1), p(z_2), \ldots\}$, the conditional entropy $H(z \mid x)$, given these two random variables, x and z [13,14] is

$$H(z \mid x) = \sum_i p(x_i) H(z \mid x = x_i)$$

where $H(z|x = x_i)$ is the average Shannon information associated to the random variable z, given a specific value of $x = x_i$.

$$H(z \mid x = x_i) = -\sum_j p(z = z_j \mid x = x_i) \log_2 p(z = z_j \mid x = x_i).$$

In this problem, two variables are associated to each node: r_i (which indicates which reference cluster, node i belongs to), and b_i^R (which indicates which cluster of the net, node i belongs to). Variable r_i depends on the information known from the anthropological point of view; and variable b_i^R depends on both rule R and the information on the attributes of the nodes. Variable r_i can take values $\{1, 2, \ldots, R_T\}$ with probabilities $\left\{\frac{N_1^r}{64}, \frac{N_2^r}{64}, \ldots, \frac{N_{R_T}^r}{64}\right\}$, as 61 is the total number of nodes belonging to some of the reference clusters. It is important to point out that the cases in which r_i takes the value 0 are not considered, as it is useful to apply only the information which is known about the reference clusters. On the other hand, the random variable b_i^R is associated to vector $\overrightarrow{b^R} = (b_1^R, b_2^R, \ldots, b_N^R)$, and takes values $\{1, 2, \ldots\}$, with probabilities $\{p(1), p(2), \ldots\}$, where

$$p(j) = S(j)/N.$$

The previous equation defines the probability of finding a node belonging to cluster j for rule R.

Appendix B. Using IDC information to implement the tests

A rule will be more informative if, in the clusters resulting from its application, there is a considerable number of nodes for which $y_i^k = 1$, at least for a certain value of k (IDC label). This number will be meaningful if it favourably compares with the null hypothesis that IDC destinations are randomly assigned, regardless of the cluster structure obtained by means of this methodology. As we have explained in Section 4.2, 77% of the information on the IDCs, which was randomly selected, is used to recognize the best rules.

B.1. Multiple test for each cluster

It will be useful to define two observables:

(i) the number of nodes in cluster j which were seen in the k IDC: $w_j^k\,|_R$,

$$w_j^k\,|_R = \sum_{i=1}^{N} \delta(b_i^R - j) y_i^k$$

where it is clear that $w_j^k\,|_R$ depends on rule R.

(ii) the size of the set of relevant IDCs for this cluster. The IDC k is a relevant IDC for cluster j if $w_j^k\,|_R > 1$ (i.e. at least two nodes in cluster j were seen in IDC k). The number of IDCs relevant to cluster j for specific rule R (which we will call $\mathcal{K}_j^R$) is obtained as follows:

$$\mathcal{K}_j^R = \sum_{k=0}^{T_{idc}} \Theta(w_j^k\,|_R - 2).$$

We have implemented a multiple hypothesis test for each cluster j which tries to reject the null (H0): "the classification of the nodes of the j-th cluster is independent of the fact that the nodes were seen in the k-th IDC".

For each cluster, a number of $\mathcal{K}_j^R$ independent tests have been carried out. α' is the significance of the multiple test. For each value of the significance, we have implemented the Bonferroni correction, obtaining $\alpha = \alpha'/\mathcal{K}_j^R$ as the significance of each individual test [15].

It is important to note that we are testing each cluster separately, regardless the rest of the clusters, and this fact is an approximation. We define $P(w_j^k = i|N, N_k, S(j))$ as the probability of finding i nodes for which $y_j^k = 1$ when taking a sample size $S(j)$, from a total of N nodes, from which N_k nodes have been seen in the IDC k. In the case H0 is true, $P(w_j^k = i|N, N_k, S(j))$ is given by hypergeometric probability distribution, with parameters N, N_k and $S(j)$.

$$P(w_j^k = i|N, N_k, S(j)) = i\frac{\binom{N_k}{i}\binom{N-N_k}{S(j)-i}}{\binom{N}{S(j)}}.$$

Let us introduce the usual p-value, which we will note as $p_j^k\,|_R$-value, specifying that the test was carried out for cluster j, using rule R when we implemented the test for IDC k.

$$p_j^k\,|_R = \sum_{i=w_j\,|_R}^{\max\{S(j),N_k\}} P(i|N, N_k, S(j)). \tag{B.1}$$

The test is one-sided as the alternative hypothesis is that there is a positive correlation between belonging to cluster j and having been seen in IDC k. This is why the sum in Eq. (B.1) ranges from $w_j\,|_R$ to $S(j)$. The top limit of the sum corresponds to the configuration in which all the components of the cluster have been seen in this IDC. This will be possible only if $S(j) < N_k$; and when this does not happen, the top limit of the sum will be N_k (the maximum number of nodes from cluster j which could have been seen in IDC k).

As we have explained, this allows one to associate an IDC to the cluster j, and this IDC label can eventually be used as a prediction for those nodes, belonging to cluster j, whose IDC is unknown. A destination prediction to those nodes belonging to the clusters which have been recognized as a *succession of kidnapping episodes* consists in an IDC vector

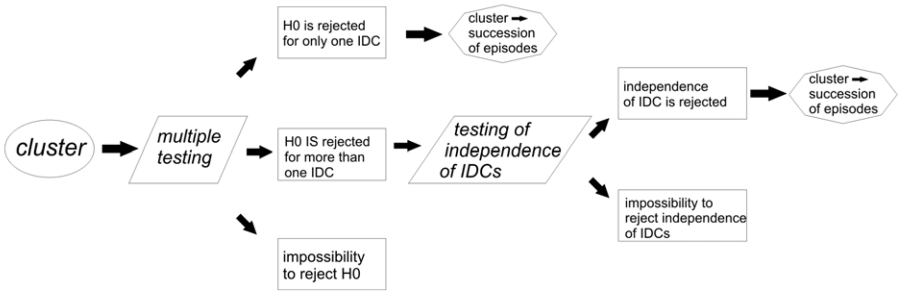

Fig. B.1. This flow chart summarizes the implementation of the use of IDC information to define a cluster on the network as a *succession of kidnapping episodes* by means of hypothesis testing. We have implemented a multiple hypothesis test for each cluster j which tries to reject the null hypothesis (H0): "the classification of the nodes of the j-th cluster is independent from the fact that the nodes were seen in the k-th IDC".

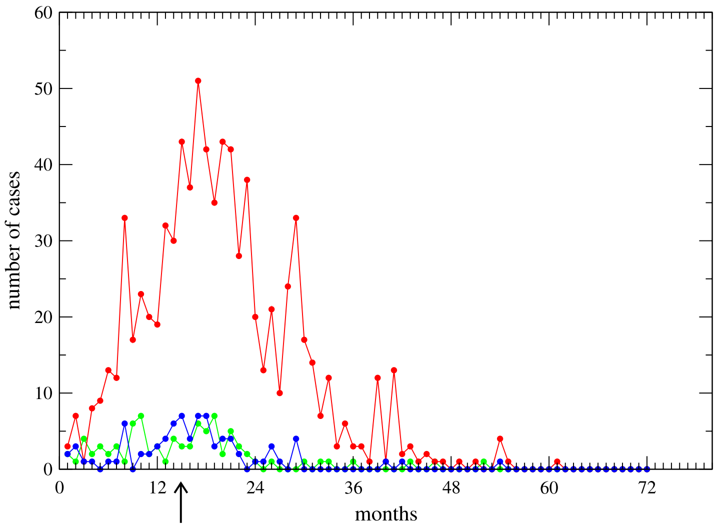

Fig. B.2. Number of disappearance cases (red), assassinations (blue) and disappearances followed by freedom (green) in Tucumán, according to month, from January 1975 to December 1980. The arrow shows the date of the coup d'etat in March 1976. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table B.1
Percentage of known information for each of the node attributes. The total number of nodes is 1036.

Attribute	% known
Political affiliation	55
Date of disappearance	92
Address (zip code)	77
Workplace (zip code)	52
Place of kidnapping (zip code)	99

$\overrightarrow{y_i}\,|_R = (y_i^1\,|_R, y_i^2\,|_R, \ldots, y_i^{Tcdc}\,|_R)$ to each node, which is defined in the same way as vector $\overrightarrow{y_i}$: the k-th component of the vector takes the value 1 if the rule predicts that the node i was held in the k-th IDC, and 0, otherwise. As a result of implementing the test for rule R (which was accepted in terms of reference clusters), the possible situations and the possible steps to be followed are represented in Fig. B.1.

If for no relevant IDC k, $p_j^k\,|_R < \alpha$ occurs, then $\overrightarrow{y_i}\,|_R = (0, \ldots, 0)$ for all the nodes i belonging to cluster j, so that $b_i^R = j$. When $p_j^k\,|_R < \alpha$ for only one IDC k, then the cluster is recognized as a *succession of kidnapping episodes* and the prediction will be $y_i^k\,|_R = 1$ for all the nodes i belonging to cluster j (so that $b_i^R = j$), and $y_i^l\,|_R = 0$ for the other components $l \neq k$ of the vector. In the following subsection we analyze the case in which more than one IDCs result significant for the cluster.

B.2. Testing circuits for more than one IDC

It may happen that more than one IDC turn out to be significant for a cluster. The fundamental assumption is that the nodes which belong to the same *succession of kidnapping episodes* shared the same destination, that is to say, that they

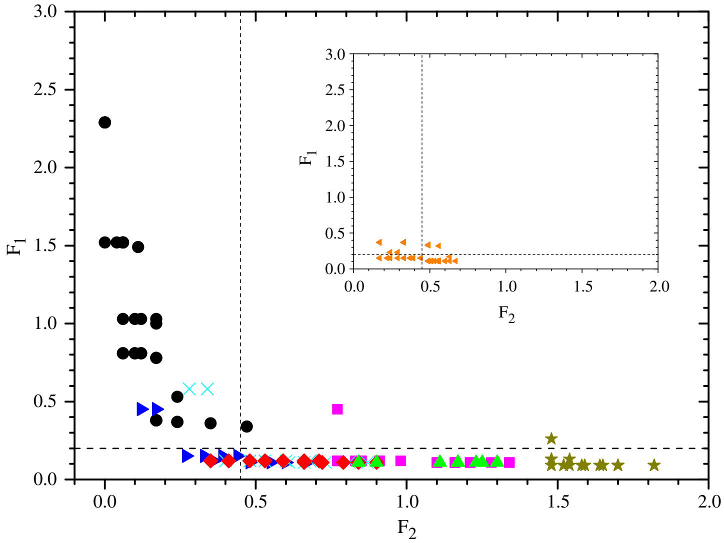

Fig. B.3. F_1 vs F_2 for homogeneous rules in time which use different geographical functions. The dotted lines limit the areas where the condition of reference clusters is met. Circle symbols: rules without geographical information; upward-pointing triangles: rules which use geographical function type 1 (home address); star symbols: rules which use geographical function type 2 (workplace); diamond symbols: rules which use geographical function type 3 (place of kidnapping); square symbols: rules which use geographical function type 4 (two out of three zip codes); (X) symbols: rules which use geographical function type 5 (at least one of three zip codes); right-pointing triangles: rules which use geographical function type 6 (proximity of the kidnapping place). The inset of the Figure shows F_1 vs F_2 for the inhomogeneous rules in time with geographical function type 6, and different temporal parameters before and after $T_1 = 1/1/1976$.

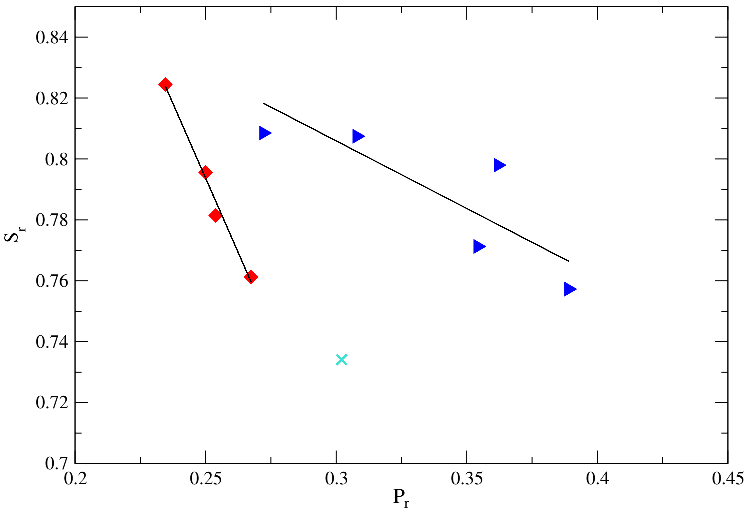

Fig. B.4. P_r vs S_r for homogeneous rules in time which are acceptable in terms of reference clusters. The right-pointing triangles represent type 6 rules (proximity of kidnapping places); the diamonds, type 3 rules (place of the kidnapping) and the X symbol, type 5 rules (one of three zip codes). In all cases, a value of $\alpha' = 0.17$ is considered when implementing the tests, and values of 0.17, 0.085, 0.057, 0.034 for α after the Bonferroni correction was implemented (depending on the total number of relevant IDCs for each cluster).

were kept in the same IDC subset ("clandestine circuit"). Likewise, as detailed in the Introduction and Section 2.2, the act of associating the destination of a single person to an IDC subset is justified because there were actually repression circuits in Tucumán. Taking this assumption into account, when a cluster turns out to be significant for more than one IDC, then we will test if we can say that there was a circuit between these IDCs by means of an exact independence test, using the information of this particular cluster [16,17].

Let us suppose that there are $w_j^{k_1}$ nodes which were seen in the IDC k_1 in the j-th cluster (whose size is $S(j)$), that $w_j^{k_2}$ nodes were seen in the IDC k_2, and that $w_j^{k_{12}}$ were seen in both IDCs k_1 and k_2. The null hypothesis is that "being held in IDC k_1 is independent from being held in IDC k_2". Then the probability that $w^{k_{12}}$ takes value x, $P(x)$, based on the assumption that H0 is true, is a hypergeometric distribution with parameters $\{S(j), w_j^{k_2}, w_j^{k_1}\}$.

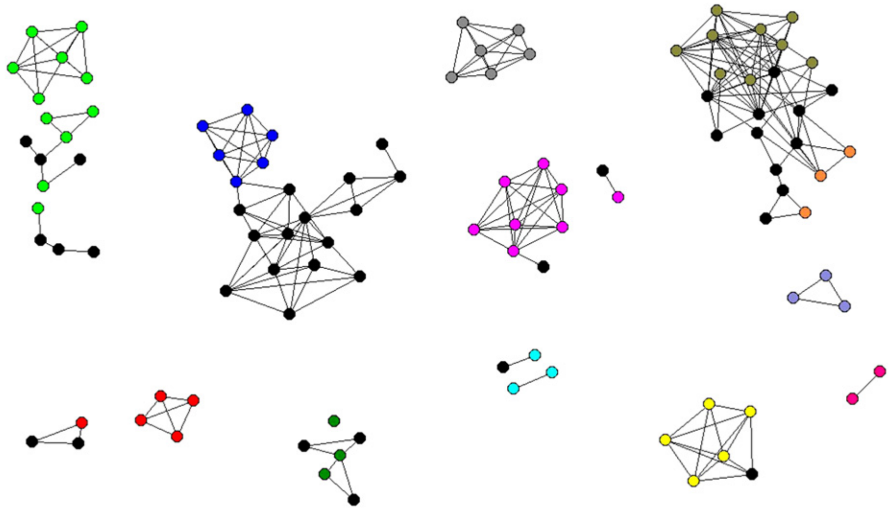

Fig. B.5. Structure of clusters obtained when implementing one of the best rules, the homogeneous rule of temporal parameters $\Delta_1 = 5$, $\Delta_2 = 3$, $\Delta_3 = 2$ and geographical function of proximity. In the Figure, only a network of 105 nodes is shown. This set of nodes comprises all the nodes belonging to the reference clusters (color circles, 64 nodes) and those nodes which, though not belonging to any reference clusters, are part of the network clusters in which there is at least one node belonging to some reference cluster (black circles) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

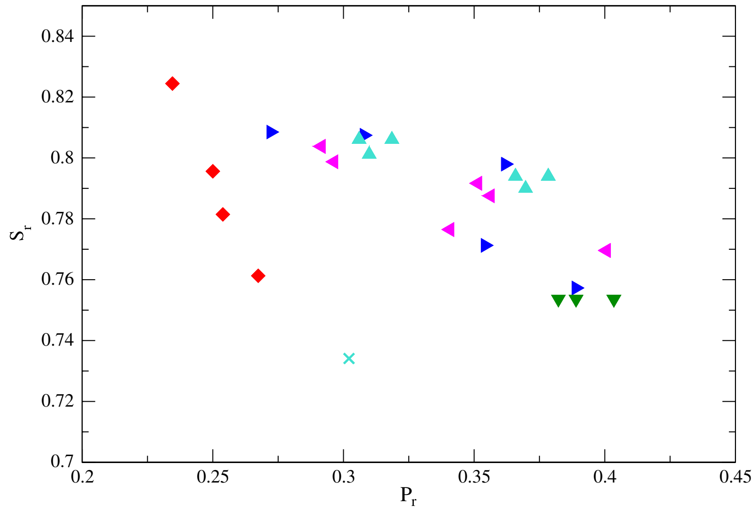

Fig. B.6. P_r vs S_r for homogeneous and inhomogeneous rules in time which are acceptable in terms of reference clusters. The right-pointing triangles represent inhomogeneous rules with geographical functions type 6 rules; upward-pointing triangles, rules where $\Delta_1^a = 7$; downward-pointing triangle symbols: rules where $\Delta_1^a = 5$, left-pointing triangle symbols, rules where $\Delta_1^a = 9$. The diamond symbols represent homogeneous rules using geographical function type 3 (place of kidnapping) and the X symbol, homogeneous rules type 5 (one of three zip codes). In all cases, a value of $\alpha' = 0.17$ is considered in the implementation of the tests.

The $p_{\text{-}value}$ of this test is

$$p_{\text{-}value} = \sum_{x=w_j^{k_{12}}}^{\min\{w_j^{k_1}, w_j^{k_2}\}} P(x).$$

If the $p_{\text{-}value} < \alpha$, then we can reject the independence hypothesis in favour of the alternative one (associating both IDCs to the cluster). That is to say, $y_i^{k_1}|_R = y_i^{k_2}|_R = 1$ for all the nodes so that $b_i^R = j$, and in this case, it is also predicted that a subset of individuals associated to this cluster were taken through the clandestine circuit which could have been comprised by two IDCs (k_1 and k_2). On the other hand, when we cannot reject the null hypothesis, then we cannot say anything about this cluster through the implementation of this test, and therefore we cannot either make any IDC assignment or make any subsequent prediction.

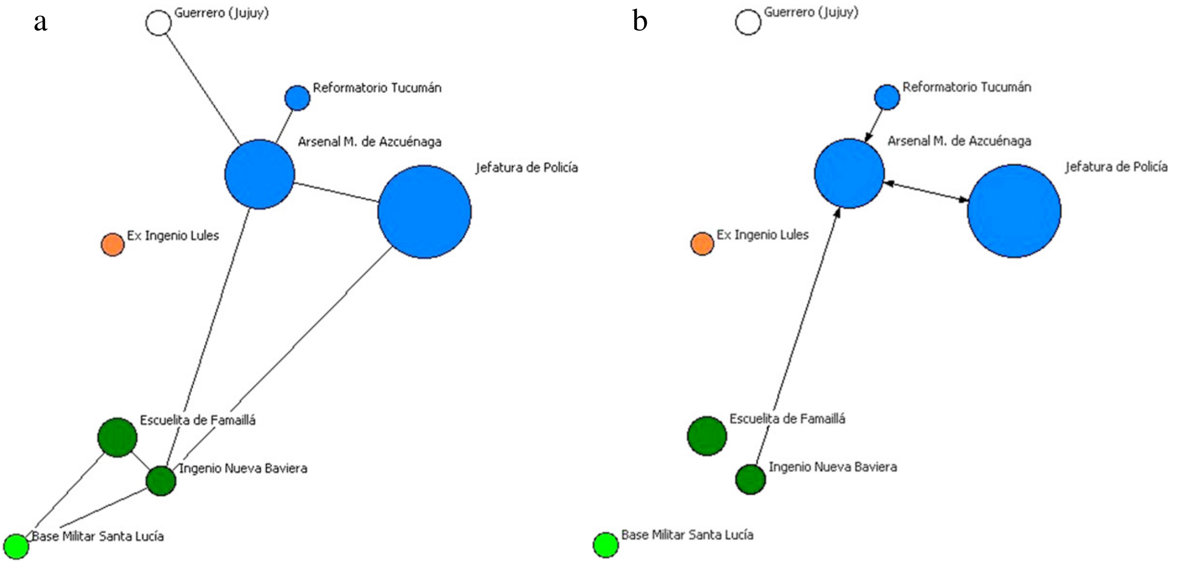

Fig. B.7. Left: IDC network which stems from implementing the inhomogeneous rule in time with temporal parameters $\Delta_1^b = 7$; $\Delta_2^b = 4$; $\Delta_3^b = 2$ before T_1; $\Delta_1^a = 5$; $\Delta_2^a = 3$; $\Delta_3^a = 2$ and after T_1 and geographical function type 6 (proximity of the place of the kidnappings). The different colors represent zip codes. The blue symbols stand for IDCs located in San Miguel de Tucumán. Right: IDC network which stems from the EAAF research. The nodes size is arbitrarily kept the same as in Fig. B.6 to ease comparison. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table B.2
Geographical functions used in the rules and the values of the temporal parameters which meet the condition about reference clusters for each one of the rules.

	Geographical function F_{ij}^g	Description	Temporal parameters which meet the condition about reference clusters
1	$\delta(a_i^l - a_j^l)$ with $l = 3$	The zip code of individuals' home address should coincide	In no case
2	$\delta(a_i^l - a_j^l)$ with $l = 4$	The zip code of individuals' workplace should coincide	In no case
3	$\delta(a_i^l - a_j^l)$ with $l = 5$	The zip code of the location where the kidnappings took place should coincide	$\Delta_1 = 9, 11$; $\Delta_2 = 5, 7$; $\Delta_3 = 1, 2, 3$
4	si $\delta(a_i^3 - a_j^3) + \delta(a_i^4 - a_j^4) + \delta(a_i^5 - a_j^5) = 2$	Two out of three zip codes should coincide (workplace, home, kidnapping)	In no case
5	$\max\{\delta(a_i^3 - a_j^3), \delta(a_i^4 - a_j^4), \delta(a_i^5 - a_j^5)\}$	At least one of three zip codes coincides	$\Delta_1 = 5$; $\Delta_2 = 4$; $\Delta_3 = 1, 2, 3$
6	A_{ij}^5	The locations where the kidnappings took place should be within 17 km	$\Delta_1 = 5, 7, 9, 11$; $\Delta_2 = 2, 3$; $\Delta_3 = 1, 2$

When a cluster turns out to be significant for three IDCs, then we test the independence of the three pairs of IDCs which are possible. If we can reject the independence hypothesis for at least two of the pairs, then we will assign the three IDCs as destination for the cluster because there is (at least) one path which connects them, and the cluster will then turn out to be a *succession of kidnapping episodes*.

References

[1] A.L. Barabasi, R. Albert, Review of Modern Physics 74 (2002) 47.
[2] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.H. Hwang, Physics Reports 424 (2006) 175.
[3] M.E.J. Newman, The structure and function of complex networks, SIAM Review 45 167 (2003) 256.
[4] A. Medus, C.O. Dorso, Physical Review E 79 (6) (2009) 066111.
[5] M.C. Gonzalez, H.J. Herrmann, J. Kertész, T. Vicsek, Physica A 379 (2007) 307–316.
[6] A. Clauset, C. Moore, M.E.J. Newman, Nature 453 (2008) 98–101.
[7] M. Rosvall, C. Bergstrom, A information theoric framework for resolving community structure in complex networks, PNAS 104 (18) (2007) 7327–7331.
[8] G. Gan, C. Ma, J. Wu, Data Clustering Theory, Algorithms, and Applications, American Statistical Association, Alexandria, Virginia, 2007, SIAM, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.
[9] EAAF is a non-government, non-profit institution, which has been working on the identification of the remains of missing people in over 50 countries since 1985. Web page of the Argentine Forensic Antrophological Team (EAAF). http://www.eaaf.org/.
[10] Informe Anual EAAF, años 2005, 2006, 2007. Informe Especial EAAF-ILLID (2008).
[11] This systematic method has been developed in specific literature:*"Tucumán, enero a diciembre de 1975"*, Campaign diary, Acdel Vilas; *"Poder y Desaparición"*, Pilar Calveiro, Colihue (2008); *"La guerra de Argelia y la Argentina. Influencia e inmigración francesa desde 1945"*, Mario Ranaletti, Anuario de Estudios Americanos, 62, 2, julio-diciembre, 285–308, Sevilla (España), 2005.: *Tucumán: Operativo Independencia*, Dante Fernandez (Argentina, 2007). Audiovisual material: *El azucar y la sangre: La guerrilla rural en tucumán 1966–1976*, Eduardo Anguita (Argentina, 2007); *Escadrons de la Mort:*

L'Ecole francaise, Marie Monique Robin (Francia, 2003). The method of the *Counter-Terrorism war* was dealt with in bibliography previous to its application in Argentina: "*La Guerre Moderne*", Roger Trinquier, Paris, La table ronde (1961); *"La Batalla de Argel"*, Gillo Pontecorvo (Italia-Argelia, 1965) has been developed in audiovisual material.
[12] Inés Caridi, Claudio O. Dorso, Pablo Gallo, Carlos Somigliana, International Journal Bifurcation and Chaos, 409, 2010 (special issue).
[13] D. Mackay, Information Theory, Inference and Learning Algorithms, Cambridge University Press, 2003.
[14] T.M. Cover, J.A. Thomas, Elements of Information Theory, in: Wiley Series in Telecommunications, 1991.
[15] Eric W. Weisstein, Bonferroni Correction, From MathWorld A Woldfram Web Resource. http://mthworld.wolfram.com/BonferroniCorrection.html.
[16] R.A. Fisher, On the interpretation of χ^2 from contingency tables, and the calculation of P, Journal of the Royal Statistical Society 85 (1) (1922) 87–94.
[17] Casella-Berger, Statistical Inference, in: Duxbury Advanced Series, 2001.