CLUSTERS IN SOCIAL NETWORKS WITH INCOMPLETE INFORMATION:

I. Caridi [*] and C.O. Dorso [*]
[*] *Department of Physics,*
University of Buenos Aires, Argentina.

P. Gallo [†] and C. Somigliana [†]
[†] *Argentine Forensic Anthropology Team*

We have developed a method to recognize communities in networks which lack of some information. It has been conceived to be applied to the problem of getting information about people who were disappeared in the Argentine Province of Tucuman during the period 1975-1983.

Keywords: Complex-networks; social networks; community structures; classification methods.

1. Introduction

Since 1985 the Argentine Forensic Anthropology Team (EAAF) has worked on identifying the human remains of those people who were killed or disappeared during the period from 1975 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 disappearance: where they were taken, when they were moved, where their remains are. To solve this puzzle, the EAAF has gathered a lot of data from very diverse sources. The undertaking of organizing the data to route the searches is a hard problem on which the EAAF has been working for more than 24 years. They have built a database in which every missing individual has certain attributes, including identity, workplace, address, place and date of disappearance, political affiliation, profession, etc. These investigations have resulted in more than 300 identifications, and they have found that the determination of groups of closely related people has turned out to be crucial in defining the circuit these people followed, and in the possible identification of their remains. The underlying hypothesis is that strongly correlated subsets of people ended up in the same CCD (clandestine detention center), and afterwards they were probably disappeared (buried, cremated, etc.) in the same place. This is the reason why, from the very beginning, it has been fundamental to establish some kind of connection among different cases of disappearance. Determining these strongly related sets leads to prioritizing some places over others in the context of the search. Additionally, it has been possible to take into account reports on the few individuals who had been seen in each clandestine detention center.

As stated above, the task at hand is to detect strongly correlated subsets of individuals, given the data gathered by the EAAF. We have decided to map the data on a complex network to accomplish such a task.

The usage of complex networks in social sciences has a long history. A point of inflection was attained with the work of S. Milgram, who studied what the web of people's connections within a community was like, and conducted an experiment which later inspired many investigations into the area of complex networks [Milgram, 1967]. Later, the study of complex networks in the real world greatly profited from the work by Watts and Strogatz [Watts and Strogatz, 1998], who constructed a model that captures the surprising properties of small world observed in real networks. In the following years, various sciences such as physics, sociology, biology and ecology paid close attention to the relationships that give rise to networks of very different complex systems: social, communication, biochemical and ecosystem networks. These researches resulted in the identification of a set of magnitudes and statistical properties that are typical tools for studying a complex network, as for example the structure of communities. For a recent review on complex networks see [Boccaletti *et al.*, 2006])

In the next section we will describe the formalization of the problem at hand. Then we will describe a method for recognizing highly correlated subsets of nodes (individuals) when the information at hand is incomplete. Next, we will show a typical example of this kind of calculation. Finally, we will discuss our conclusions and present perspectives of this work.

2. Formalizing the problem

We treat the problem of determining correlations among people, in terms of studying communities/clusters in complex networks. The information collected by the EAAF is used to build a social network in which the missing people are represented by nodes (hereafter referred to as such). Each node is characterized by a series of attributes. Links are established between nodes based on different criteria to relate the nodes attributes (for example, there will be a link between two nodes if they have: i) the same political affiliation, ii) the dates of disappearance differ in less than a given number of days, and also, iii) the place where the disappearance took place or the address, etc. is located within, say, a distance less than or equal to 10 km). From now on, each set of criteria used to relate the attributes of the nodes, will be referred to as a *rule*.

In this case, we have formalized the question on how to detect non-obvious correlations among different nodes

as the problem of recognizing clusters or communities of highly correlated nodes within the network.
But there is another problem: the information at hand is not complete. This fact has two sources, on one hand the information collected refers only to events of disappearance and there is no information at all with respect to very short time detentions that not have been reported as disappearances. It means that the set of nodes is incomplete. On the other hand, for some cases of the disappeared the information about some attributes is not available.

In order to cope with such a set of incomplete information we have devised a methodology that we have named CGC (clusterization, growth and coalescence). This method comprises different stages of analysis using different criteria at each one. The final result is a set of clusters where nodes with different degree of information coexist.

3. The method: CGC

In this section we will describe the method of *clusterization, growth and coalescence* (CGC) to search for correlations in a database with incomplete information. The method is based on the liquid vapor phase transition.
Let $\mathcal{N} = \{n_1, n_2, ..., n_N\}$ be the set of nodes in the network. Each node i is characterized by a set of the attributes which we call the relevant ones $A_i = \{a_i, b_i, c_i, d_i, e_i, ...\}$ and these attributes might be known or partially known. This possibility will be denoted by the set $M_i = \{a_i, b_i, ...\}$, containing the set of unknown attributes for this node. The M's do not have a fixed length and their composition varies from node to node.
Instead of using a unique linking definition between nodes i and j which connects both sets of attributes A_i and A_j, the method builds clusters in a set of steps. Each of these stages is associated to a new link definition, which gives rise to new connections in the network and, consequently, a new (larger) structure of clusters arises. The basic stages of this method are:
i) clusterization: at this stage, a certain hypothesis is defined to establish connections between nodes explicitly including the complete set of A attributes in the definition, as follows:

$$L^c_{ij} = f_c(A_i, A_j)$$

where L^c_{ij} can take the value 0 or 1. This is a possible definition for the function:

$$\begin{aligned} f_c &= \delta(a_i - a_j)[1 - \Theta(|b_i - b_j| - \Delta_1)] \\ &\times [max\{\delta(d_i - d_j), \delta(e_i - e_j), \delta(f_i - f_j)\}] \end{aligned}$$

This kind of function allows us to deal with very different attributes. For this function, a link will be present between nodes i and j if these nodes have the same political affiliation (the a attribute), the disappearance took place within a temporal interval Δ_1 (the b attribute) and both nodes lived in the same place, or worked in the same place, or the place where the event of disappearance took place was the same (d, e and f attributes)).
This defines an unweighted network, and for the particular problem we are dealing with, it generates a disconnected network, which comprises clusters of different sizes (some of which might be isolated nodes).
A cluster is defined in the following way, given a node i and a cluster C, then [Strachan and Dorso, 1997]

$$i \in C \iff \exists\ j\ \in C\ /\ L^c_{ij} = 1$$

ii) growth and coalescence: clusters generated at the first stage might grow with the addition of nodes with some unknown attributes, or, after growth, may coalesce with other clusters. At this stage, the link between two nodes i and j is defined as follows:

$$L^g_{ij} = f_g(A_i, A_j)$$

L^g_{ij} can take the values 0 or 1. At this stage, we are dealing with the nodes for which the information is not complete. So, in order to relate two nodes, the condition on the known attributes should be stronger than in the previous case. This is shown in the following definition,

$$\begin{aligned} f_g &= [1 - \Theta(|b_i - b_j| - \Delta_2)] \\ &\times [max\{\delta(d_i - d_j), \delta(e_i - e_j), \delta(f_i - f_j)\}] \end{aligned}$$

where $\Delta_1 > \Delta_2$, and we have assumed that the a attribute is the one with the least information available. Then, nodes whose M_i set contains the a attribute, will have the chance of relating the ones already clustered at stage 1. The network will still be unweighed, and the application of CGC to this particular problem results in an unconnected network.
After repeating and combining the different stages of this method, we obtain a cluster structure where there are nodes with unknown attributes.
The set of link definitions $\{L^c_{ij}, L^g_{ij}, ...\}$ and each of the parameters involved in the definition of the $f's$, are called a "CGC rule".
We have characterized the clusters structure obtained through a variable x_i associated to each node, which represents the cluster number which the i node belongs to. In this way, x_i may be an integer value from 1 to NC (which is the total number of resulting clusters, which must be lower than or equal to N). Both the set $\{x\}$ and the value of NC will depend on the CGC rule applied. Less stringent rules will give a lower number of clusters, while stricter rules will give rise to fewer connections and a greater number of clusters on the network.

4. One Example: Tucuman 1975-1983

We will now briefly review some relevant data regarding the phenomena of the disappeared in Argentina during

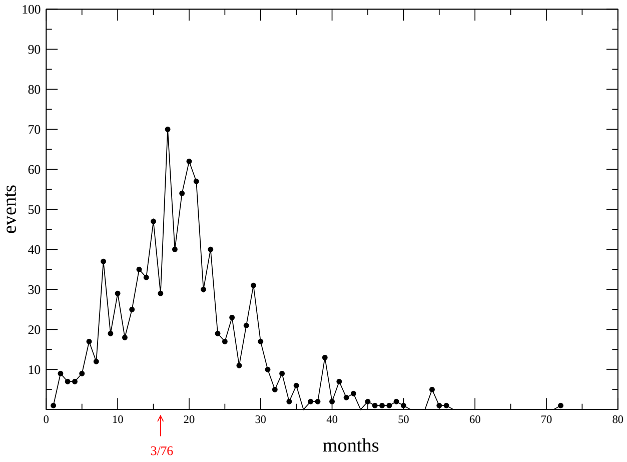

FIG. 1: In this figure we show the number of events as a function of time. Time is measured in months starting from January 1975. The military coup took place in march 1976.

the period from 1975 to 1983. In particular we will analyze the disappeared in the Tucuman area.

In Fig.1 we show the number of disappeared as a function of time in the considered period.

The total number of disappeared is (at the time being) 913. In this way we will be dealing with a network with that number of nodes.

In the following table we show the percentage of known data for some of the relevant attributes associated with each node.

Attribute	% ***known***
political affiliation	38
date of disappearance	96
address (zip code)	77
workplace (zip code)	48
place of disappearance (zip code)	99.8

In Fig.2 we show the network resulting from the application of the CGC algorithm to the disappearances which took place between July 9th, 1976 and the October 9th, 1976 comprising a total of 167 events. Clusters were determined using the following CGC rule:

$$\begin{aligned} f_c &= \delta(a_i - a_j)[1 - \Theta(|b_i - b_j| - \Delta_1)] \\ &\quad \times [\delta(f_i - f_j)\}] \end{aligned}$$

$$\begin{aligned} f_g &= [1 - \Theta(|b_i - b_j| - \Delta_2)] \\ &\quad \times [\delta(f_i - f_j)\}] \end{aligned}$$

with Δ_1 equal to 7 days while for f_g, Δ_2 takes the value of 2 days.

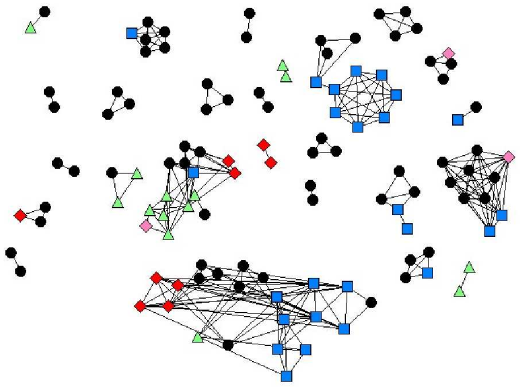

FIG. 2: In this figure we show the clusters (of more than one node) resulting after the analysis using the CGC method on the EAAF database, applying the rule described in the text. In this figure, circles denote individuals with unknown political affiliation. Other shapes are associated with different political organizations like, for example, squares denote individuals belonging to the Montoneros organization, and up pointing triangles denote individuals belonging to the PRT party.

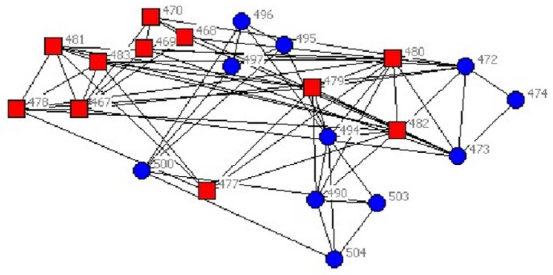

FIG. 3: In this figure we detail the large cluster detected by this CGC rule in Fig.2). This historical event has been fully confirmed through the analysis of different historical sources. In the figure, red squares denote those individuals who were seen in a given CCD (called Guerrero). While blue circles were not been seen in this CCD. Moreover, the nodes with labels 477,478,479,480,481,482,483 and 496, were seen at a second CCD (called Arsenal)

In Fig.3 we show a zoom of the largest cluster observed at the bottom of the Fig.2. This is a very interesting case because it involves different political affiliations, a rather large number of individuals, but the number of CCD's in which these individuals were seen is only two (see caption for details). There were 16 CCD in this geographical region.

5. Conclusions and perspectives

We have mapped the database gathered by the EAAF onto a network. We have devised a method that allows us to detect clusters of highly correlated individuals even when the information at hand is incomplete. Preliminary results are quite encouraging allowing us to detect, among others, a large cluster in which different political affiliations coexist, more important this large cluster can be traced to an historical event which completely justifies it existence.

Acknowledgments

C.O.D is a member of the Carrera del Investigador, CONICET. I.C. is a postdoctoral fellow CONICET C.O.D acknowledges partial financial support form grant PIP5969 CONICET

References

Boccaletti, S.Latora, V., Moreno, Y., Chavez, M., & Hwang, U. [2006] "Complex networks: Structure and dynamics," *Physics Reports* **424**, 175-308.

Milgram, S. [1967],"The small world problem," *Psychology Today* **2**, 60.

Strachan, A. & Dorso C. O. [1997], "Fragment Recognition in Molecular Dynamics"; *Phys.Rev.* **56**, 995.

Watts, D., & Strogatz, S. [1998], "Collective dynamics of small world networks" *Nature* **393**, 440.