Analysis and Visualization of Social Networks induced by Criminal Records towards the Identification of Gangs: a real case for Argentina

Sebastián P. WAHLER[1,2], Martín L. LARREA[3], and Diego C. MARTÍNEZ[3]

[1] Departamento de Informática, Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco, Mitre 655, U9100, Trelew, ARGENTINA.
http://www.ing.unp.edu.ar/dpto-informatica.html

[2] Departamento de Informática, Procuración General, Ministerio Público Fiscal, Poder Judicial de la Provincia del Chubut, Belgrano 521, U9102, Rawson, ARGENTINA.
https://www.mpfchubut.gov.ar

[3] Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur, Av. Alem 1253, B8000CPB Bahía Blanca, ARGENTINA.
https://cs.uns.edu.ar/

(e-mail: spwahler@ing.unp.edu.ar, dcm@cs.uns.edu.ar, mll@cs.uns.edu.ar)

Abstract Social ties are essential inputs for the detection of criminal gangs. This input becomes even more important when it can be analysed and visualized using software tools. This paper presents a software development, currently in use, for the analysis and visualization of social networks created from criminal records in the province of Chubut. The identification of illegal networks, such as criminal gangs, is of special interest in order to promote intelligent criminal prosecution.

Keywords: Criminal Investigation · Data Analysis · Social Networks · Visualization.

1. Introduction

Criminal activities in a city or region can range from minor offenses such as theft and robbery to more serious ones such as protection rackets, cybercrime, sexual abuse, and homicides. Law enforcement agencies record these crimes using various methods and technical details. Criminal records typically include information such as the type of crime, date and time of occurrence, location, and identity of the suspect(s) if known.

This information supports the judicial investigation processes of each case, but over time they constitute an extensive knowledge base on which it is possible to extract valuable information for crime prevention and the search for justice. For example, it is possible to identify relationships between people based on a transitive analysis of criminal events in time and space that suggest the formation of either formal or informal criminal gangs. Relations such as the friendship

between various delinquents can also be inferred from criminal records. This social ties are of the utmost importance for crime prevention, as well for resolution of unfinished cases.

Criminal organizations are groups that operate outside the law. They carry out illegal activities for their benefit and to the detriment of other individuals or social groups [14]. These groups can be of different sizes and cover varied geographic areas. Frequently these groups conflict with each other. One of the particular characteristics of this type of organization is the anonymity and discretion of their members, being protective of each other. Criminals know that they are part of a network where the anonymity of each member depends heavily on the anonymity of the rest. This demonstrates the importance of collecting and analyzing information associated with the social ties of criminals.

In many cases, the criminal acts are perpetrated by individuals of low rank in the group, with little responsibility and motivated by immediate reward, aspirations of promotion, and higher reputation in their circle of contacts. On the other hand, the masterminds are individuals of higher positions in the hierarchy, with more responsibility in the criminal organization. Those individuals have leadership qualities, long-term interests, and a constant concern for retaining power for personal benefit. Security agencies usually have a record of the perpetrators, while the masterminds are more strenuous to identify. Additionally, the hierarchical structures of the gang, the way they operate, and the inherent culture of the socio-economic class lead to an entangled set of inner codes that generates more obstacles for the identification of the organization as a whole. In this context, we believe that the research and development presented in this article are a significant contribution to the prevention and resolution of criminal activities.

We have worked within two areas of Computer Science; Information Visualization, particularly Visualization of Large Data Sets, which translate information into a visual context [42] [13] [28], such as a map or graph, to make data easier for the human brain to understand and pull insights from [7], and the Social Networks Analysis, which investigate social structures through the use of networks and graph theory.

In this line of research, we study the application of these techniques to a real scenario, using criminal records of the Department of Justice[4] of the province of Chubut in Argentina. We developed active software components for the visualization and intelligent analysis of data, incorporating notions of graph analytics. In particular, in this work we are interested in considering the *PageRank* algorithm, contributing to the detection of relevant criminals among *communities of individuals*, in a similar way it is applied to rank web pages. In order to do this, we use real records of criminal activities through the collaboration of the Public Prosecutor's Office.

The rest of the article is structured as follows. The next section reviews the state-of-the-art in terms of visualization testing. In the subsequent sections, we continue with the presentation of the black-box and white-box testing tools for

[4] Ministerio Publico Fiscal

information visualizations. We develop a case study to illustrate both kinds of testing. The case study is based on a C# tool designed for the visualization of geological data, and it exemplifies the process of finding errors with tools and methods presented in this work. The last section presents the reached conclusions and the intended future work.

2. Análisis de Redes Sociales (SNA)

Social Network Analysis (SNA) has contributed to criminal investigations and related intelligence activities. A social network models individuals as nodes linked to each other by arcs or edges that represent the relationships between those individuals. These networks, and their properties, are relevant because they represent an abstraction of human relations that allows the highlighting of specific aspects of the ties and individuals [25] [5]. Networks form graph structures, and the properties of these structures represent the properties of social relations. According to Sage [37], there are four fundamental pillars of network analysis: recognition of the importance of social relationships between individuals, the collection and analysis of data on these relationships, the importance of visual representation of these data, and the need for mathematical and computational models that explain the connection patterns between individuals.

Several authors have addressed the benefits of studying social networks for criminal investigations. In the mid-1970s, basic models were used to establish and qualify the relationships between individuals or actors in a particular scenario by defining graphs according to the information collected [19]. In these cases, the processing was done manually and with several stages of data refinement and evaluation. According to Klerk [21], this is the first generation of network analysis in criminalistics. The second generation involved computational tools that automate part of the task of recording and structuring data. These tools also significantly increased the amount of data to be analyzed, making recording and consultation much more agile. The third and current generation establishes the definition of mathematical models and techniques for the generation of new knowledge. Such as the identification of positions of power and influence or the quality of potential witnesses or informants. Metrics like the centrality of a node in a graph are especially useful in this scenario.

Krebs [22] presented one of the most significant works in this regard; he identified a part of the terrorist network responsible for the attacks in the United States on September 11, 2001. He did it through their social ties with the pilots responsible for the hijacking. The works [30], [35], and [39] have used a similar strategy. On the other hand, the analysis of social networks has also gained interest in traditional criminal investigations such as mafia structures or drug trafficking [2] [3] [17] [31] [32]. Studies such as Malm's [27] have made it possible to identify roles in the supply chain of illicit drugs, which entails different criminal risks for each of the collaborators. There are also examples of these strategies applied in other illegal activities, such as art trafficking [1], money laundry [8] [38], police corruption [23], and youth gangs [29]. There are also lines

of research in the discipline related to cybercrime [9] [10] [11]. It is clear then that social network analysis can be applied to a wide range of criminal activities and has been shown to appropriately model characteristics of illegal organizations, assisting in crime prevention and the design of adequate policies to deal with them.

However, some difficulties still require intensive studies. The amount of information that is handled is enormous, in many cases with incomplete, contradictory, and, no less frequently, incorrect information. In addition, traditional human relationships are naturally mixed with illicit interactions between individuals, so it is necessary to properly identify their nature and consequences and determine the sensible limits of the analysed social network.

Currently, the state agencies in charge of justice and crime prevention have computerized records of the criminal activities detected and information derived from the investigation processes and proceedings. This information essentially constitutes a form of a social network. For our work, we pay special attention to the data produced for this purpose by the police forces of the Province of Chubut and its Judiciary through the Public Prosecutor's Office (MPF [15]). All this data is registered in a software system called Coirón. This data set contains tens of thousands of records and can be used to model different social networks on which to apply a mathematical and computational analysis. By doing so we can transform this data into information. This will make it possible to learn more about criminal activities and their perpetrators in the jurisdiction of Chubut.

Some tools and techniques can facilitate the analysis and exploration of these large data sets. In this sense, the area of Information Visualization, particularly the Visualization of Large Data Sets, seeks to assist users in such a way [42] [13] [28]. It is also important to study the tasks and interactions that the visualization must support since it is these interactions that enable the exploration of information visualization.

3. Marco de Trabajo - Ministerio Público Fiscal del Chubut

Coirón is the computer system that manages the administration of cases admitted to the Public Prosecutor's Office of Chubut. It is a tool that allows the registration, communication and management of activities, procedures and actions that are carried out for a criminal case, from the initial charge to its final completion. As a registration tool, it builds a database with the history of every case, as well as the people involved and those responsible for management in each office. As a communication tool, it groups information, allows cross-examination of relations, identifies links between cases, people, and their backgrounds. As a management tool, it manages the evolution of cases and the corresponding work of the officials. It allows planning, organizing, coordinating and controlling the workflow related to each case. It has been developed according to the needs of the Public Prosecutor's Office of Chubut, based on the current Criminal Procedure

Code and adapted to the strategic guidelines for the design and management of state Prosecutor's Offices defined by the federal Attorney General. Its progress, maintenance and continuous improvement is in charge of the Development Team of the Department of Information Technology of the Area of Planning and Management Control of the General Procurement.

Since Coirón is a software tool to support criminal investigation, it is important to enhance its features towards a smart provision of data. In particular, we are interested in the analysis of criminal records in order to facilitate the identification of gangs. This should be paired with correct information visualization tools, enhancing the analysis that will be carried out later by criminal analysts. In this paper we focus then on criminal gangs, applying techniques to acknowledge the *relative importance* of their members, which can be visualized properly in a graph denoting social connections. A relational profile can be build for criminal prosecution, through the identification of "criminal partners" via the social network induced by criminal records, in order to identify if they are part of a simple street gang or some larger criminal organization [**?**]. Social network is represented as graphs, which are of great visual aid when working with a large number of records. There are many variations on the graphs, but they all share the common feature of using a labeled circle for each actor in the population and line segments between pairs of actors to represent the fact that there is a link between them. The "*Group of Membership*" in the `Coirón` system refers to the direct relationship between an individual within the universe of people charged as perpetrators of crimes (either involved, indicted or sentenced) and other individuals of the same universe, with one or more criminal cases in common.

A software module "Membership Group Network" graphically displays this data, enriched with information obtained by social network analysis. Through various filters, it is possible to graphically show the relationships between a certain group of people in order to identify the formation of possible criminal gangs. In the graph a node a person involved in two or more criminal cases. There is a large number of people in the system with only one case with the role of *reported*, and for this reason they are excluded. However they could be part of the dataset to be displayed if any of them are found related to other nodes of the first group. The size of the node is directly related to the number of criminal cases in which the person is involved. The larger the size of the node, the more criminal cases it will be involved in.

Los segmentos de líneas entre pares de nodos, vinculan a las personas entre sí y representan el o los casos que tienen en común. El grosor de la vinculación será directamente proporcional a la cantidad de casos en común entre un par de personas. Hay nodos que se encontrarán aislados en el grafo, esto no significa que no estén involucrados en casos, sino que quizás no existan relaciones para el filtro de búsqueda que se utilice en esa vista en particular.

Supongamos que una persona "A" se encuentra asociada a 8 casos penales, una persona "B" a 4 y una persona "C" a 2 casos. Agreguemos que las personas "A" y "B" se encuentran relacionadas entre sí, por estar en 3 casos en común

(casos 1, 2 y 3). Por otro lado las personas "A" y "C" también se encuentran relacionadas, por tener un caso en común (caso 4). Una representación gráfica de dicha situación se muestra en la Figura 1, y puede observarse el doble de tamaño entre el nodo "A" y el nodo "B", representando justamente la diferencia de casos entre ambos nodos (8 y 4 casos). También se ve a simple vista el grosor del enlace entre "A" y "B" tres veces más grande que el enlace entre "A" y "C" (3 casos en común entre el primer par de nodos, y sólo un caso para el último par de nodos mencionado).

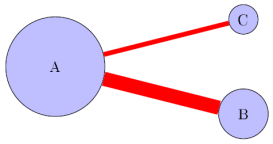


Figura 1. Ejemplo de relación entre tres personas.

Esta es, en primera instancia, una caracterización de la importancia de los individuos en la red. Sin embargo, analíticas mas complejas podrían aplicarse.

4. Data Overview

In this section, we describe our criminal case dataset and associated network of people, as well as some interesting features to mention.

Crime Dataset Our data set consists of criminal cases, actions (criminal process events log), crimes, people, elements (reported and kidnapped), all of them related; registered between October 2006 and May 2022 in the Judicial District of Trelew - Chubut. This dataset includes places (relating to people and criminal acts), dates, procedural statuses of cases and people, as well as the links between all the aforementioned datasets. In Table 4, we summarize some of the most important characteristics of the data set.

Characteristic	**Total quantity**
Cases	105586
People	132950
People in Cases	183348
Crimes	113010
Nodes	33178
Links	16964
Node/Link Relations	60513

Cuadro 1. General data set totalizers

Network Properties From the criminal case data, we were able to build the network of *Belonging Groups*. In this network, the nodes of those people whose roles are not referred to criminal actors are eliminated, such as: whistleblowers, victims, victims, etc. Figure 2 shows a visualization of the network. In it, a graph made up of more than 30000 people with more Criminal Cases registered in the Coirón Management System can be observed (with the following criteria: involved in more than one case with the role of accused, suspected or denounced; deceased persons are included , minors and legal persons). In addition, all the relationships that exist between these people and their groups of belonging are displayed in the figure.

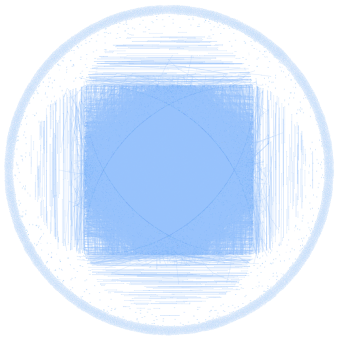
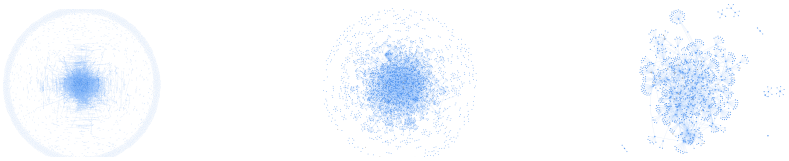


Figura 2. More than 30000 people with more cases and their relationships.

For the practical purposes of criminal investigation, a visualization with so many nodes and relationships is not representative or leads to any type of detection of criminal gangs, but it is a clear example of the universe of data that is available in the dataset used, as well as the power of the visualization tool. In Figure 3 you can see examples in which the membership groups of each node to be displayed have been taken into account, that is to say, the people with their particular membership groups are displayed (according to parameters of search selected). In image (a) 10000 people are shown, in (b) 1000 and in (c) 100 people with more cases and their related membership groups. By analyzing



(a) Top 10000. (b) Top 1000. (c) Top 100.

Figura 3. People in more cases, with the inclusion of their particular membership groups.

the composition of the network obtained, we can observe the relationships that

exist between the nodes and how the graph is "balanced", making those nodes with few or no relationships remain on the periphery of the graph. In addition to this, the measure of centrality of those nodes that are surrounded by their related ones is also appreciable. A clearer approximation to denote the measure of centrality can be seen reflected in Figure 4, where only the 10 people with the most Cases and their belonging groups are displayed. Clearly, these 10 main nodes are surrounded by their membership groups and transitivities between them can be observed through nodes that are part of the membership group of more than one main node.

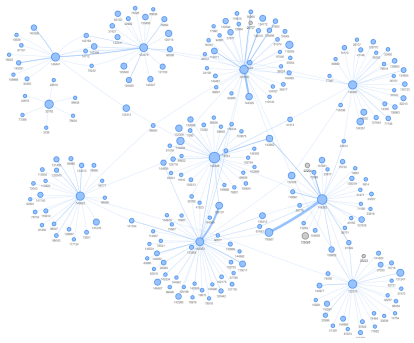


Figura 4. 10 people with more cases in Coirón, with their relationships

Degree Centrality Degree centrality is one of the simplest measures of centrality. In this, the number of links or connections that a node has with the other nodes belonging to a graph is measured. When an analysis of this type is applied, different measures can be determined. For example, in social networks we can measure the degree of entry of a node as the popularity or preference it has and the exit define it as an indicator of sociability. In our case study, members of criminal gangs dynamically modify their relationships with other members of the network, resulting in a change in their role and importance. A number of degree centrality measures can help identify these changes. These statistics can be used to filter the view of the network based on the value of a specific node and highlight its position within the network. The degree of centrality in our graph will then be defined as the number of direct links that an offender has. A node with a high degree can be seen as a "hub", an active and important node in the [6] network.

Transitivity The clustering coefficient (transitivity) of a graph measures the degree of connection of a network. High clustering coefficients mean the presence of a high number of triangles in the network. It is well known in the literature [41] that social networks show high clustering coefficient values when they reflect the underlying social structure of contacts between friends/acquaintances. Furthermore, high values of the local clustering coefficient are considered a re-

liable indicator of nodes whose neighbors are very well connected and between which a substantial amount of information can flow.

Visualization Development To carry out the visualization of the data set obtained from the previously described intelligent analysis, `Vis.js` [40], a library or dynamic visualization library based on the Javascript language, was used. It is designed to be easy to use, to handle large amounts of dynamic data, and to allow manipulation and interaction with the data. The library consists of the DataSet, Timeline, Network, Graph2d, and Graph3d components.

In our particular case we use the "Network" component, which allows us to display networks in graphs. The visualization is easy to use and supports shapes, styles, colors, sizes, images, etc. It works seamlessly in any modern browser for up to a few thousand nodes and edges. To handle a larger number of nodes, Network has clustering support. The grid uses HTML canvas for rendering.

Vis.js provides implementations of Force-directed graph drawing algorithms. These force-directed algorithms attempt to position nodes by considering the forces between two nodes (attractive if connected, repulsive otherwise). They are generally iterative and move nodes one by one until improvement is no longer possible or the maximum number of iterations is reached. The links are more or less the same length and have as few cross links as possible. Connected nodes move closer together while isolated nodes move further to the sides.

5. Identificación de posibles Bandas Delictivas

En esta sección, describimos nuestro problema, algunos de los enfoques prácticos existentes utilizados por las fuerzas de la ley y nuestro enfoque basado en la teoría de grafos con características generadas principalmente por la distribución de los datos anteriormente descrita.

Métodos existentes Recordemos que en este trabajo nuestro principal interés es la identificación asistida de bandas delictivas y sus cualidades.

Las personas nos movemos habitualmente entre lugares conocidos o nodos (hogar, trabajo, supermercado, restaurante) y por las mismas calles o rutas. La teoría sugiere que cuando ocurre un delito es porque se cruzan delincuentes y víctimas dentro de algunas de estas zonas de actividad (nodo, ruta). A partir del análisis del lugar del delito se pueden determinar distintos tipos de víctimas y delincuentes que lo frecuentan, entender por qué concurren a ese lugar y qué hace que se encuentre la dupla delincuente-víctima. Es una manera estructurada de conocer e investigar patrones de comportamiento.

Por otro lado se puede deducir que los delincuentes se comportan igual que el resto de las personas, realizan actividades diariamente, se mueven por rutas conocidas para ir de la casa al trabajo, o a algún otro lugar que frecuenten. Es decir, mantienen una cierta rutina en sus vidas. Un delincuente tenderá a cometer un delito en algún lugar que se encuentre dentro o cerca del recorrido

que realiza diariamente para trasladarse desde la casa al trabajo, del trabajo a algún lugar de recreación u otro lugar habitual.

De ambos enfoques se busca encontrar la mayor cantidad de patrones de ocurrencia entre diversos hechos de similar criminalidad y patrones horarios, como así también las zonas geográficas en donde se producen.

La naturaleza de los vínculos de los integrantes de una banda delictiva es una variable que aporta información sobre las características y similitudes de los miembros del grupo, atendiendo a criterios concretos: vínculo familiar, cultural, de proximidad (provienen del mismo barrio), han compartido prisión, de especialización (habilidades delictivas), la experiencia u otras capacidades, y otros tipos de vínculo.

Enfoque propio Ante los enfoques teóricos y prácticos estudiados anteriormente, nuestro desarrollo de software propio, que permite mostrar de manera gráfica las relaciones entre actores delictuales en el Sistema Penal de la Provincia del Chubut, se potencia como una herramienta vital de apoyo en la toma de decisiones de la investigación penal de bandas delictivas.

Poder visualizar relaciones entre las personas involucradas en casos penales ayuda a los especialistas a detectar triangulaciones, transitividades y por supuesto centralidades en la Red. Todo ello, sumado a los indicios de investigación y la propia expertís en la temática completan una herramienta de análisis para determinar ciertas bandas o grupos altamente relacionados.

En el año 2019 existieron investigaciones vinculadas a reiterados robos de televisores LCD en domicilios [20], como así también una serie de hechos consecutivos vinculados al robo de cajas fuertes en empresas del parque industrial de la ciudad de Trelew.

La UAC (Unidad de Análisis Criminal), organismo auxiliar de la Procuración General perteneciente al Ministerio Público Fiscal del Chubut, sirvió como equipo de apoyo en la investigación de ambos modus operandi, haciendo uso de toda la información de los legajos fiscales, consultas generales y específicas contenidas en el Sistema Coirón. Fue de vital uso la información referida a los *grupos de pertenencia* de cada persona, pero devino en un arduo trabajo entrecruzando información de personas, para dar con las supuestas bandas delictivas detrás de estos hechos.

Dichas investigaciones sirvieron como puntapié inicial para realizar este trabajo y poder facilitar la información ya contenida en el sistema de gestión penal, de otra manera, de una forma más directa y visual a la hora de investigar, que sirva directamente como apoyo a la toma de decisiones en las investigaciones de bandas delictivas.

A continuación se puede observar una visualización extraída de este trabajo, utilizando como filtros de búsqueda dos personas (nodos 116587 y 145262) con muchos casos y relaciones en el sistema, a fin de encontrar si existe algún tipo de relación directa entre ambos, y a su vez si existen nodos que produzcan transitividades o sean a su vez centrales de otros grupos. Desde la visualización se agregó gracias a la vinculación con el sistema de la oficina de identificación de

personas, fotografías para colocar en los nodos y hacer de este trabajo una herramienta aún más potente. Aquellas personas que no hayan sido identificadas en sede judicial no tendrán fotografía. Por cuestiones judiciales se han desenfocado las fotografías y se han colocado identificadores en vez de los nombres reales de las personas intervinientes.

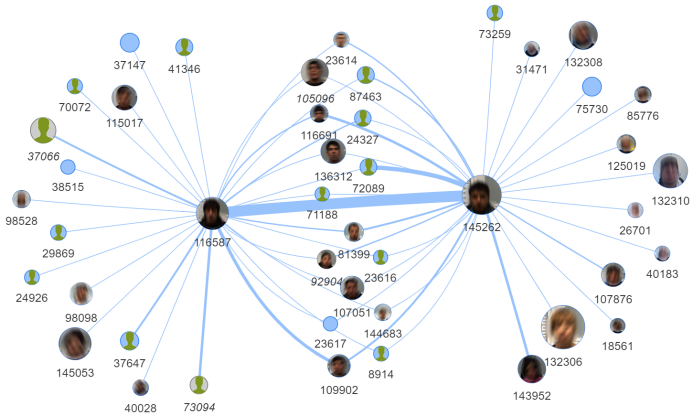


Figura 5. Relación entre dos personas. Se agregan fotografías.

Como puede verse, existen en la parte central de la imagen muchas personas que se encuentran relacionadas delictualmente con ambos nodos en cuestión. De esta manera se pueden tomar acciones con respecto a estas personas en pos de encontrar patrones de ocurrencia que los vinculen ante la posibilidad de identificarlos como una supuesta banda delictiva.

6. PageRank and community detections

As a simplified part of the structure of a community of nodes in a social network, each one represents an individual and the network has a crowd segmentation [26]. Some people are central to the community, some are on the fringes, having fewer relationships with others and therefore less influence. In this section, we present a new community discovery approach based on the PageRank algorithm to find these "important" or "most influential" criminals in our graph, in order to analyze suspected criminal gangs. Recall that a graph is a pair $G = (N, A, g)$ where N is a non-empty finite set of elements called *nodes* (vertices), A is a set of arcs and g is a function that associates each arc a belonging to A with an unordered pair (x, y), where x and y are nodes belonging to N. a is said to be an arc with endpoints x and y [12].

PageRank (PR) is a method that was implemented through an algorithm originally used by Google that assigns each web page in a given set a score that reflects its importance within the set. This score is called the *PageRank value*. Before a query, the search engine uses these scores to determine the level of relevance of the pages, and returns first those with a higher score. To calculate scores, PageRank uses the link structure of the web [4]. A web page has a high PageRank value if it is pointed to by many other pages, or if it is pointed to by pages with high rankings [34]. PageRank is intuitively based on the concept of *random walks* over graphs [18]: suppose a random browser starts browsing the web from any page. The navigator can randomly click on any of the links on the page he is currently on with probability d, which is called *damping factor*, or with probability $1 - d$ he accesses randomly to any other web page. This process is repeated indefinitely. Then, the PageRank value of a page P can be interpreted as the probability that the random browser will be in P at the end of the process. PageRank is formally defined as [16]. Let q_i be the number of outgoing links that page i has, n the total number of web pages, d the *damping factor* that generally takes the value 0.85, π a column vector named *PageRank vector*, and $H = (h_{ij})$ a square matrix of size n such that $h_{ij} = 1/q_i$ if there is a link from page i to page j , and $h_{ij} = 0$ otherwise. The value h_{ij} corresponds to the probability of accessing page j from page i in one step, after clicking on one of the links that appear on the latter. The PageRank value corresponding to page j is π_j, and is defined recursively as shown in the equation 1 [24].

$$\pi_j = \frac{1-d}{n} + d\sum_{i=1}^{n} \pi_i h_{ij} \tag{1}$$

PageRank application for criminal gangs Our dataset described above is obtained from SQL queries to the `Coirón` Database. To make use of the PageRank algorithm, it was decided to incorporate it into those SQL queries in order to obtain a result that can be used for visualization. Within the original query, a table is generated for the nodes and another table for the relationships, in this way the software made for the visualization obtains said datasets and renders the graph. To incorporate the PageRank calculation, both tables were initially adapted for the use of the formula, and certain temporary tables were needed for the calculation. On the one hand, the degree of output of each node *(Out Degree)* was computed, that is, the number of links that connect it with other nodes. Then the *damping factor* is declared, in our case 0.85, then the total count of nodes, and the initial *PageRank* of each node is calculated, to then begin the iteration seeking to comply with the sum of the formula. The *damping factor* corresponds to a probabilistic value that, applied to the web page scenario, aims to capture the possibility that a user will continue clicking on the links of a page in a continuous browsing session. Here this factor has a different meaning, reinterpreted as the factor in which the importance of an individual among their peers is diluted through a chain of arcs. We are currently studying an appropriate value based on existing data, since the damping factor

is essentially an empirical value. In this work we opted to use the eigenvalue of the original PageRank proposal.

```
INSERT INTO #OutDegree
SELECT #Node.id, COUNT(#Edge.src)
FROM #Node
LEFT OUTER JOIN #Edge ON #Node.id = #Edge.src
GROUP BY #Node.id
DECLARE @dampingFactor float = 0.85
DECLARE @Node_Num int
SELECT @Node_Num = COUNT(*) FROM #Node
INSERT INTO #PageRank
SELECT #Node.id, rank = ((1 - @dampingFactor) / @Node_Num)
FROM #Node
INNER JOIN #OutDegree ON #Node.id = #OutDegree.id
DECLARE @Iteration int = 0
```

```
WHILE @Iteration < 50
BEGIN
--Iteration Style
SET @Iteration = @Iteration + 1
INSERT INTO #TmpRank
SELECT #Edge.dst, rank = ((1 - @dampingFactor) / @Node_Num)
+ (@dampingFactor * SUM(#PageRank.rank / #OutDegree.degree))
FROM #PageRank
INNER JOIN #Edge ON #PageRank.id = #Edge.src
INNER JOIN #OutDegree ON #PageRank.id = #OutDegree.id
GROUP BY #Edge.dst
END
```

Once the development of the formula was finished, we proceeded to carry out tests that corroborate the proper functioning of the code. Examples were made for few nodes with few relationships, in such a way that verification is easy. Figure 6 is shown below, which reflects the visualization of the execution of PageRank for 6 nodes. In each node its PageRank position is shown, and in parentheses the identifier of each node. As can be seen, the central node for the calculated PageRank is the one referred to ID *145053*, whose relationships with 3 nodes weigh on the relationships that the rest of the nodes displayed in the graph have. It is interesting to see that this individual is not necessarily the one with the most criminal cases, but he is the most important among his peers *of his own social network* of related contacts.

IdNodo	PageRank
145053	0,286845042094944
145262	0,199512347112651
116587	0,1910604814502
132310	0,109790233522352
123109	0,106270247927848
129137	0,106270247927848

Figura 6. PageRank run result for 6 nodes.

Community Detection A community can be defined as a set of nodes that are more densely connected to each other than to the rest of the network. The importance of this approach lies in the fact that nodes that are contained within the same community are expected to share attributes, common characteristics or functional relationships [26]. This work takes advantage of the previously described SQL algorithm in which nodes and relationships are divided to be later visualized by the `Vis.JS` tool. In this sense, our approach is based on the search for possible people who participate in criminal gangs. From the data origin it appears that for each node all its relationships can be known, so that we can assign to each of those referring nodes a group or cluster identifier. It is then possible to verify for each pair of nodes, if they belong to a particular group (one of them will be "referring" and we will be able to identify it), and in this way assign a certain group to each relationship as well. With both tables (nodes and

relations) updated, it is possible from the visualization tool to assign colors to each cluster, and thus generate an even more practical graph to view.

Although the JavaScript language is used for visualization by the previously mentioned library `Vis.JS`, and the datasets are obtained from the System Database `Coirón` through SQL queries, the software study that takes the data from the dataset and processes it to later call the visualization library, it is developed in the C Sharp language of the .Net Framework. Below is shown in Figure ?? the same visualization that has been presented previously in Figure ??, but now with the detection of groups by color. It can be seen that each group or cluster of nodes shares the same color for internal links.

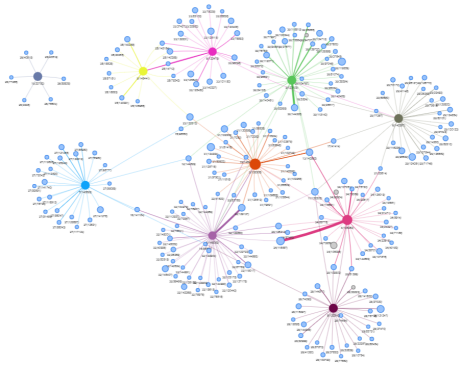


Figura 7. 10 people with more cases in Coirón, with their relationships. Community detection by color is added.

Real case studies The algorithm was also executed for real study cases resolved in the MPF, where the gangs and their leaders have been identified, and in this way validate the relevance of the implementation. As an example, Figure 8 is shown below, on the real case of LCD theft mentioned in the previous chapter. It was possible to validate that all the members of the band are in the center of the subgraph, highly related to the colored nodes, which reflect those with the highest PageRank value.

Improvements to PageRank - Link weight The idea behind PageRank is that "good" pages refer to other "good" pages. Therefore, the pages that those "good" pages refer to have a higher PageRank. Assuming that a user browses the Web randomly, such that if they are on a page, with a certain probability they either get bored and leave the page, or uniformly and randomly choose to follow one of the links on the same page on the found (removing autolinks). Therefore, the probability of being on page "p" is

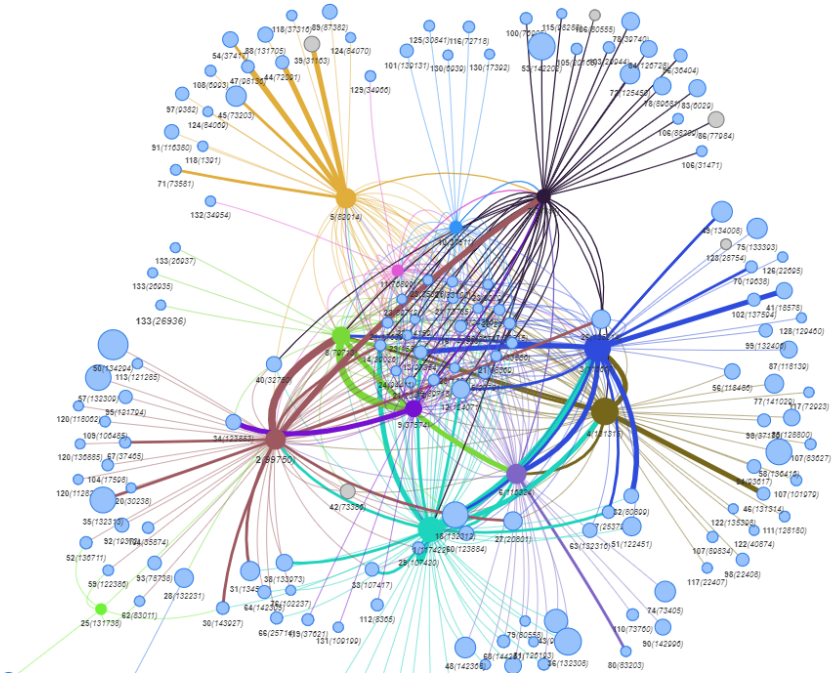


Figura 8. PageRank execution result for the real case of LCD theft.

$$PR(p) = \frac{q}{T} + (1-q)\sum_{i} \frac{PR(r_i)}{L(r_i)} \tag{2}$$

where T is the total number of pages, q is the probability of leaving page p ($q = 0:15$ is suggested in the original PageRank paper), ri are the pages that point to page p, and $L(ri)$ is the number of links on page ri. These values can be used as page rank and can be calculated using an iterative algorithm that converges quite quickly since we are interested in the rank order rather than the actual values. The term q is called the damping factor, as it exponentially decreases link spam based on sequences of links returning to a page.

From here arises a variant of Google's original PageRank algorithm, called WLRank proposed in the work of Ri Baeza-Yates and Emilio Davies [**?**].

WLRank (Weighted Links Rank) assigns the rank value $R(i)$ to page i using the following equations:

$$R(i) = \frac{q}{T} + (1-q)\sum_{j} \frac{W(j,i)R(j)}{\sum_k W(j,k)} \tag{3}$$

$$W(j,i) = L(j,i)(c + T(j,i) + AL(j,i) + RP(j,i)) \tag{4}$$

where given a link from page j to page i we have:

$L(j;i)$ is 1 if the link exists, or 0 otherwise, and c is a constant giving a basis weight to each link, $T(j;i)$ is a value that depends on the tag where the link is inserted, $AL(j;i)$ is the length of the link "anchor" text divided by a constant d that depends on estimating the average length of the anchor text in characters, and $RP(j;i)$ is the relative position of the link. link on the page weighted by a constant b.

Similar to PageRank, $R(i)$ corresponds to the probability of reaching page i while browsing the Web. If $W(j;i) = L(j;i)$ we have the original PageRank. The changes are explained below. The term $T(j;i)$ is a sequence of constants depending on the tag where the link is located. For example, if the link is inside a $< h1 >$ tag, it will have a high value of $T(j;i)$, slightly less for $< h2 >$, etc. Same for other emphasis tags like $< strong >$ or $< b >$.

The term $AL(j;i)$ gives more value to links where the creator explains in more detail what Web resource is being linked to. For example, this gives less weight to links described with home or here. Finally, the term $RP(j;i)$ gives more weight to links that are at the top of the page than at the bottom of the page (physically in the HTML code, not necessarily in the browser view).

Gracias a esta mejora sobre la fórmula original de PageRank es posible entonces darle mayor consideración en la fórmula a aquellos enlaces que tienen más pesos que otros.

Se procedió a adecuar nuestra fórmula de modo de contemplar los pesos en los enlaces y de esta forma realizar ciertas pruebas simples que comprueben que la modificación realizada consigue ajustar a nuestro modelo en una mejora respecto al algoritmo de PageRank original.

A continuación en la Figura 9 se muestra el ejemplo de 6 nodos mostrado con anterioridad en donde se puede observar el resultado ponderando los enlaces.

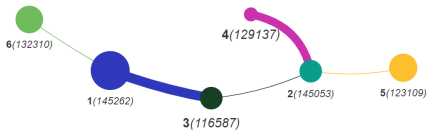


Figura 9. Resultado de ejecución de PageRank para 6 nodos con modificación WL-Rank.

Se puede apreciar a simple vista que ahora al tener peso las relaciones en la fórmula el resultado es mucho más preciso, existen muchos casos en cómun entre 145262 y 116587, lo que justifica la modificación en el ranking general. El nodo 145262 tiene 1 caso en cómún con 132310 y 18 con 116587, un total de 19 relaciones a ese nodo. Por otra parte el nodo que quedó en segunda posición en el ranking (145053), tiene un total de 15 enlaces provenientes de 13 casos en común con el nodo 129137 y de 1 caso en cómun con el nodo 116587 y con el nodo 123109.

Mejoras a PageRank - Peso a los nodos Como se mencionó en el punto anterior, el algoritmo de PageRank en su versión original, no hace valorar al peso de los enlaces (tema propuesto y resuelto con anterioridad), ni el peso de los nodos. Este último es otro punto muy importante a la hora de evaluar a los individuos en los grafos resultantes, como así también a la hora del posible descubrimiento de una supuesta banda delictiva.

Es casi imposible evaluar y calificar de banda delictiva a individuos que tengan un bajo grado de relación entre sí, y que además cada uno tenga como antecedente penal un número pequeño de casos en su haber.

Para ello, se procedió del mismo modo que con los pesos en los enlaces, a modificar la fórmula original de PageRank en pos de obtener un resultado más significativo y darle más importancia en el grafo a aquellos nodos que tengan más casos (más tamaño en el grafo), cuando se encuentren relacionados con otros con la misma cantidad de relaciones. El objetivo principal de dicha modificación es generar un desempate de valor de ranking para aquellos nodos que en el algoritmo original de pagerank obtengan igual puntaje. Ante una situación de igualdad de ranking, el nodo que tenga un mayor tamaño quedará con un mayor valor ponderado.

No se encontró bibliografía dedicada que aporte datos para aplicar a la fórmula, por lo que se realizó en base a lo ya estudiado una modificación propia.

Para comenzar con la ponderación del peso de los nodos, se identifica primero al nodo con mayor peso en el grafo, y luego sobre el resultado de la fórmula de PageRank se procede a sumar un nuevo valor empírico que surge de dividir el tamaño de cada nodo por el tamaño de mayor peso ya calculado, para luego multiplicarlo por el valor de PageRank anteriormente calculado. Cada nuevo valor de ranking sobre cada nodo es sobreescrito y pasa a ser el valor final. A continuación se detalla el algoritmo SQL de actualización.

```
1

DECLARE @NodeMax float
SELECT @NodeMax = MAX(n.nodeValue) FROM #Node n
IF(@pesoEnNodos = 1)
BEGIN
UPDATE PR
SET rank = rank + ( (#Node.nodeValue/@NodeMax) * rank )
FROM #PageRank PR
INNER JOIN #Node ON PR.id = #Node.id
end
```

En donde *Node* es la tabla de los Nodos, $nodeValue$ es el peso de cada nodo, $NodeMax$ se obtiene como el máximo valor de peso de la tabla de Nodos, y *rank* es el valor previo de PageRank. Al realizar el SET se produce la actualización de cada Ranking para cada nodo.

A continuación se muestra en la Figura 10 el mismo ejemplo de 6 nodos mostrado con anterioridad en donde se puede observar el resultado ponderando sólo los nodos.

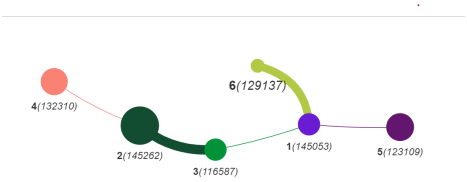


Figura 10. Resultado de ejecución de PageRank para 6 nodos con modificación en peso de Nodos.

En el mismo es posible observar comparando la ejecución de PageRank, que mantiene el resultado principal de la fórmula original, pero para los casos de nodos con "empate" de ranking, ahora ponderará con mayor valor al que posea un peso mayor. Los nodos 129137 y 123109 que en el algortimo original compartían el quinto lugar en el ranking, ahora pueden diferenciarse por su tamaño (cantidad de casos penales de cada persona). El nodo 123109 posee 68$casospenales$ asociados, quedando en quinto lugar, mientras que el nodo 129137 se encuentra vinculado a 47$casospenales$, por lo que queda relegado al sexto lugar en el ranking de este grafo.

Por último se muestra en la Figura 11 el mismo ejemplo, pero ahora con ambas modificaciones aplicadas a la fórmula original de PageRank.

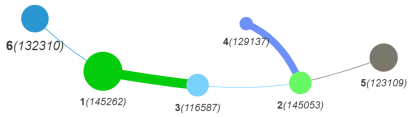


Figura 11. Resultado de ejecución de PageRank para 6 nodos con modificación WL-Rank y la modificación propia en peso de Nodos.

7. Conclusiones y trabajos futuros

Se ha presentado una propuesta de estudio de las técnicas y metodologías actuales de análisis inteligente de datos y visualización para la asistencia en la investigación criminal. Todo ello a partir de los registros de actividades delictivas, sus autores y las relaciones de datos que pueden derivarse a partir de ellas. Fue de especial interés la identificación de redes ilegales, tales como bandas delictivas o criminales para propender a una persecución penal inteligente.

Del estudio propuesto, como se explicó en los capítulos anteriores, se desarrolló un módulo de software como herramienta gráfica para visualizar la red de grupos de pertenencia de los actores delictuales, incorporando un algoritmo de PageRank para reflejar aquellas personas importantes y Detección de comunidades asignándole colores a cada grupo/cluster de nodos.

Luego de ello se pudo identificar que no era suficiente el algoritmo de PageRank a la hora de evaluar las posibles bandas delictivas, ya que en el mismo no se tienen en cuenta los pesos propios de cada nodo y cada enlace. Para mejorar el algoritmo se procedió a incorporar a la formula original las ponderaciones necesarias tanto para hacer valer el peso de los nodos como así también el de los enlaces.

El desarrollo de software se implementó en el mismo Ministerio Público Fiscal, del cual se tomaron los datos para generar los datasets de pruebas. De esta manera se ha logrado no sólo estudiar las técnicas y metodologías expuestas, sino también alcanzar la puesta en producción y uso de la herramienta, por los propios actores de la investigación.

Las primeras impresiones de aquellos especialistas de investigaciones penales han sido muy satisfactorias y permiten evaluar a este trabajo como el inicio de futuros desarrollos visuales para el apoyo a la toma de decisiones en la investigación penal. Actualmente se continúa trabajando en el desarrollo de la aplicación de visualización. Se pretende modificar la fórmula original de PageRank, enriqueciéndola con información adicional según los registros judiciales. Por ejemplo, la posibilidad de darle mayor ranking inicial a aquellos nodos que tengan un peso mayor que otros según los registros. También sería interesante hacer lo mismo con el peso que poseen los enlaces entre nodos, ya que no es lo mismo una relación de 2 casos penales en común entre dos personas, que una

de 18 casos en común. De esta manera lograríamos darle más ranking también a aquellos nodos que estén relacionados con otros, en mayor cantidad de casos penales. Esto es sin duda relevante para la investigación criminal basada en antecedentes penales. También se buscará profundizar sobre diversos algoritmos de centralidad. Existen algunas nociones que son de relevancia para la identificación de la importancia de una persona en la red inducida por las causas penales. Por ejemplo, *betweenness centrality*, que modela la medida en que un nodo en particular se encuentra entre otros nodos en una red, o *closeness centrality*, que es la inversa de la suma de los caminos más cortos (geodésicas) que conectan un nodo particular con todos los demás nodos de una red [33]. De manera similar, *eigenvector centrality*, es otra forma de asignar la centralidad a un actor de la red basado en la idea de que si un nodo tiene muchos vecinos centrales, también debería ser central.

Referencias Bibliográficas

1. Bichler, G., Malm, A.: Small arms, big guns: a dynamic model of illicit market opportunity. Global Crime **14**(2-3), 261–286 (2013)
2. Bouchard, M., Amirault, J.: Advances in research on illicit networks. Global crime **14**(2-3), 119–122 (2013)
3. Bright, D.A., Greenhill, C., Reynolds, M., Ritter, A., Morselli, C.: The use of actor-level attributes and centrality measures to identify key actors: A case study of an australian drug trafficking network. Journal of contemporary criminal justice **31**(3), 262–278 (2015)
4. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems **30**(1-7), 107–117 (1998)
5. Burcher, M.: Social network analysis and law enforcement: Applications for intelligence analysis. Springer (2020)
6. Carley, K.M.: Destabilization of covert networks. Computational & Mathematical Organization Theory **12**(1), 51–66 (2006)
7. Chen, H., Atabakhsh, H., Tseng, C., Marshall, B., Kaza, S., Eggers, S., Gowda, H., Shah, A., Petersen, T., Violette, C.: Visualization in law enforcement. In: CHI'05 extended abstracts on Human factors in computing systems. pp. 1268–1271 (2005)
8. Colladon, A.F., Remondi, E.: Using social network analysis to prevent money laundering. Expert Systems with Applications **67**, 49–58 (2017)
9. Décary-Hétu, D.: Information exchange paths in irc hacking chat rooms. Crime and Networks. New York: Routledge pp. 218–230 (2014)
10. Décary-Hétu, D., Dupont, B.: The social network of hackers. Global Crime **13**(3), 160–175 (2012)
11. Décary-Hétu, D., Dupont, B.: Reputation in a dark network of online criminals. Global Crime **14**(2-3), 175–196 (2013)
12. Dubinsky, E.: Mathematical structures for computer science. by judith l. gersting. The American Mathematical Monthly **91**(6), 379–381 (1984)
13. Feng, M., Zheng, J., Ren, J., Hussain, A., Li, X., Xi, Y., Liu, Q.: Big data analytics and mining for effective visualization and trends forecasting of crime data. IEEE Access **7**, 106111–106123 (2019)
14. Finckenauer, J.O.: Problems of definition: what is organized crime? Trends in organized crime **8**(3), 63–83 (2005), `https://doi.org/10.1007/s12117-005-1038-4`

15. Fiscal, M.P.: Página web. https://www.mpfchubut.gov.ar/
16. Franceschet, M.: Pagerank: Standing on the shoulders of giants. Communications of the ACM **54**(6), 92–101 (2011)
17. Giommoni, L., Aziani, A., Berlusconi, G.: How do illicit drugs move across countries? a network analysis of the heroin supply to europe. Journal of Drug Issues **47**(2), 217–240 (2017)
18. Göbel, F., Jagers, A.: Random walks on graphs. Stochastic processes and their applications **2**(4), 311–336 (1974)
19. Harper, W.R., Harris, D.H.: The application of link analysis to police intelligence. Human Factors **17**(2), 157–164 (1975)
20. Jornada, D.: Caso de estudio real. https://www.diariojornada.com.ar/57375/policiales/Como_era_el_trabajo_de_la_banda_de_los_LCD_que_fue_desbaratada_esta_semana_en_Trelew
21. Klerks, P.: The network paradigm applied to criminal organisations: Theoretical nitpicking or a relevant doctrine for investigators? recent developments in the netherlands. Connections **24**(3), 53–65 (1999)
22. Krebs, V.E.: Mapping networks of terrorist cells. Connections **24**(3), 43–52 (2002)
23. Lauchs, M., Keast, R., Yousefpour, N.: Corrupt police networks: uncovering hidden relationship patterns, functions and roles. Policing & society **21**(1), 110–127 (2011)
24. Lin, J., Dyer, C.: Data-intensive text processing with mapreduce. Synthesis Lectures on Human Language Technologies **3**(1), 1–177 (2010)
25. M., K.R.P., Mohan, A., Srinivasa, K.G.: Practical social network analysis with python. Computer Communications and Networks, Springer International Publishing, Basel, Switzerland, 1 edn. (aug 2018)
26. Ma, X., et al.: Exploring sharing patterns for video recommendation on youtube-like social media. Multimedia Systems **20**(6), 675–691 (2014)
27. Malm, A., Bichler, G.: Networks of collaborating criminals: Assessing the structural vulnerability of drug markets. Journal of research in crime and Delinquency **48**(2), 271–297 (2011)
28. Mathew, A., Mary Jose, A., Sabu, C., Raj, A., et al.: Criminal networks mining and visualization for crime investigation. Caniya and P, Mufeed and Raj, Asha, Criminal Networks Mining and Visualization for Crime Investigation (July 8, 2021) (2021)
29. McGloin, J.M.: Policy and intervention considerations of a network analysis of street gangs. Criminology & Public Policy **4**(3), 607–635 (2005)
30. Medina, R.M.: Social network analysis: a case study of the islamist terrorist network. Security Journal **27**(1), 97–121 (2014)
31. Morselli, C.: Hells angels in springtime. Trends in organized crime **12**(2), 145–158 (2009)
32. Morselli, C.: Assessing vulnerable and strategic positions in a criminal network. Journal of Contemporary Criminal Justice **26**(4), 382–392 (2010)
33. Newman, M.E.: A measure of betweenness centrality based on random walks. Social networks **27**(1), 39–54 (2005)
34. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)
35. Qin, J., Xu, J.J., Hu, D., Sageman, M., Chen, H.: Analyzing terrorist networks: A case study of the global salafi jihad network. In: International Conference on Intelligence and Security Informatics. pp. 287–304. Springer (2005)
36. Rua, G., González, L.: Sistemas judiciales: Una perspectiva integral sobre la administración de justicia. Análisis criminal en América Latina **23** (2020)

37. Scott, J., Carrington, P.J.: The SAGE handbook of social network analysis. SAGE publications (2011)
38. Soudijn, M.R.: Using strangers for money: A discussion on money-launderers in organized crime. Trends in organized crime **17**(3), 199–217 (2014)
39. Stollenwerk, E., Dörfler, T., Schibberges, J.: Taking a new perspective: mapping the al qaeda network through the eyes of the un security council. Terrorism and Political Violence **28**(5), 950–970 (2016)
40. visjs: Página web. `https://visjs.org/`
41. Wasserman, S., Faust, K., et al.: Social network analysis: Methods and applications (1994)
42. Xu, J., Chen, H.: Criminal network analysis and visualization. Communications of the ACM **48**(6), 100–107 (2005)