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 [15]. 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 [43] [14] [29], such as a map or graph, to make data easier for the human brain to understand and pull insights from [8], 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 Social Network Analysis (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 [26] [6]. Networks form graph structures, and the properties of these structures represent the properties of social relations. According to Sage [38], 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 [20]. In these cases, the processing was done manually and with several stages of data refinement and evaluation. According to Klerk [22], 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 [23] 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 [31], [36], and [40] 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 [3] [4] [18] [32] [33]. Studies such as Malm's [28] 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 [2], money laundry [9] [39], police corruption [24], and youth gangs [30]. There are also lines of research in the discipline related to cybercrime [10] [11] [12]. 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 [16]). 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 [43] [14] [29]. 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 Framework - Public Prosecutor's Office of 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 [37]. 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.

Line segments between pairs of nodes link people together and represent the case(s) they have in common. The thickness of the link will be directly proportional to the number of cases in common between a pair of people. There are nodes that will be isolated in the graph, this does not mean that they are not involved in cases, but that there may not be relationships for the search filter that is used in that particular view.

Suppose that a person "A" is associated with 8 criminal cases, a person "B" with 4 and a person "C" with 2 cases. Let's add that people "A" and "B" are related to each other, because they are in 3 cases in common (cases 1, 2 and 3). On the other hand, people "A" and "C" are also related, because they have a case in common (case 4). A graphical representation of this situation is shown in Figure 1, and the double size can be observed between node "A" and node "B", precisely representing the difference in cases between both nodes (8 and 4 cases). Also seen with the naked eye is the thickness of the link between "A" and "B" three times greater than the link between "A" and "C" (3 cases in common between the first pair of nodes, and only one case for the last mentioned pair of nodes).

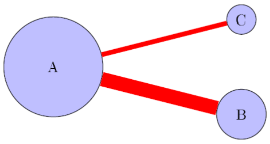


Fig. 1. Example of relationship between three people.

This is, in the first instance, a characterization of the importance of individuals in the network. However, more complex analytics could be applied.

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

Table 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.

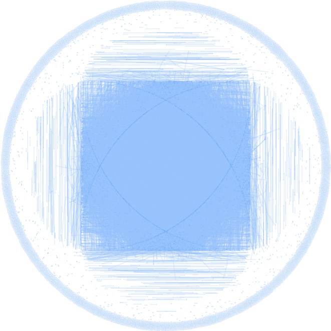
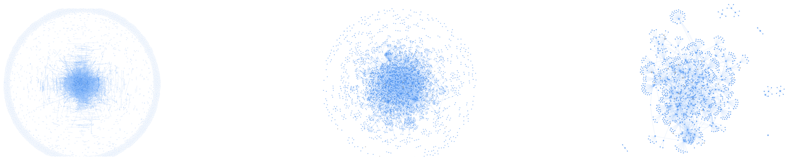


Fig. 2. More than 30000 people with more cases and their relationships.

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.

Fig. 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.

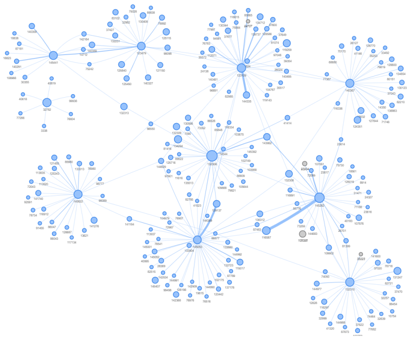


Fig. 4. 10 people with more cases in Coirón, with their relationships

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 [7] 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 [42] 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 reliable 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` [41], 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 Identification of possible Criminal Gangs

In this section, we describe our problem, some of the existing practical approaches used by law enforcement, and our approach based on graph theory with features primarily generated by the data distribution described above.

Existing methods Let us remember that in this work our main interest is the assisted identification of criminal gangs and their qualities.

People usually move between known places or nodes (home, work, supermarket, restaurant) and along the same streets or routes. The theory suggests that when a crime occurs it is because criminals and victims cross paths within some of these activity zones (node, route). From the analysis of the crime scene, different types of victims and criminals who frequent it can be determined, understand why they go to that place and what makes the criminal-victim duo meet. It is a structured way of knowing and investigating behavior patterns.

On the other hand, it can be deduced that criminals behave the same as the rest of the people, they carry out daily activities, they move along known routes to go from home to work, or to some other place they frequent. That is, they maintain a certain routine in their lives. An offender will tend to commit a crime somewhere that is within or near his daily commute from home to work, from work to a place of recreation, or other usual place.

Both approaches seek to find the greatest number of occurrence patterns among various events of similar criminality and hourly patterns, as well as the geographical areas where they occur.

The nature of the ties of the members of a criminal gang is a variable that provides information on the characteristics and similarities of the members of the group, according to specific criteria: family, cultural, proximity ties (they come from the same neighborhood), have shared prison, specialization (criminal skills), experience or other abilities, and other types of bond.

Our focus Given the theoretical and practical approaches studied previously, our own software development, which allows us to graphically display the relationships between criminal actors in the Criminal System of the Province of Chubut, is promoted as a vital support tool in decision-making criminal investigation of criminal gangs.

Being able to visualize relationships between the people involved in criminal cases helps specialists to detect triangulations, transitivities and of course centralities in the Network. All this, added to the research evidence and the subject's own expertise complete an analysis tool to determine certain gangs or highly related groups.

In 2019, there were investigations linked to repeated thefts of LCD televisions at homes [21], as well as a series of consecutive events linked to the theft of safes in companies in the industrial park of the city of Trelew.

The UAC (Criminal Analysis Unit), an auxiliary agency of the Attorney General belonging to the Chubut Public Prosecutor's Office, served as a support team in the investigation of both modus operandi, making use of all the information from the tax files, general consultations and specific information contained in the Coirón System. The information referring to each person's *groups of belonging* was of vital use, but it became an arduous job crossing information from people, to find the alleged criminal gangs behind these events.

These investigations served as an initial kick to carry out this work and to be able to provide the information already contained in the criminal management system, in another way, in a more direct and visual way when investigating, which directly serves as support for decision-making. decisions in criminal gang investigations.

Below you can see a visualization extracted from this work, using as search filters two people (nodes 116587 and 145262) with many cases and relationships in the system, in order to find if there is any kind of direct relationship between them, and in turn. time if there are nodes that produce transitivities or are in turn central to other groups. From the visualization, thanks to the link with the people identification office system, photographs were added to be placed in the nodes and make this work an even more powerful tool. Those people who have not been identified in court will not have a photograph. For legal reasons, the photographs have been blurred and identifiers have been placed instead of the real names of the people involved.

As can be seen, in the central part of the image there are many people who are criminally related to both nodes in question. In this way, actions can be taken with respect to these people in order to find patterns of occurrence that link them to the possibility of identifying them as a supposed criminal gang.

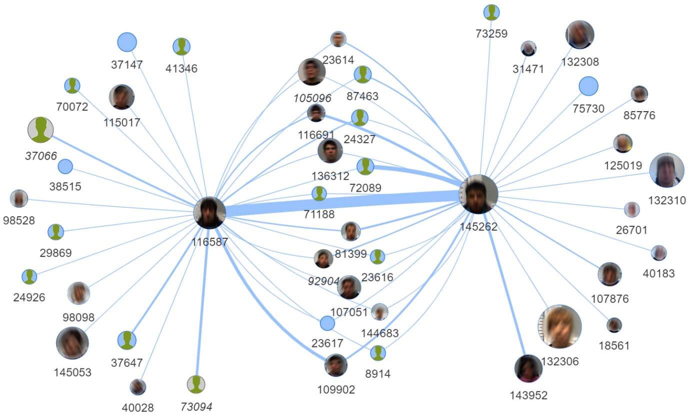


Fig. 5. Relationship between two people. Photos are added.

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 [27]. 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 [13].

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 [5]. 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 [35]. PageRank is intuitively based on the concept of *random walks* over graphs [19]: 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 [17]. 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 [25].

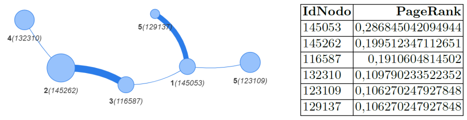
$$\pi_j = \frac{1-d}{n} + d\sum_{i=1}^{n} \pi_i h_{ij} \tag{1}$$

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

Fig. 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 [27]. 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 7 the same visualization that has been presented previously in Figure 4, 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.

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

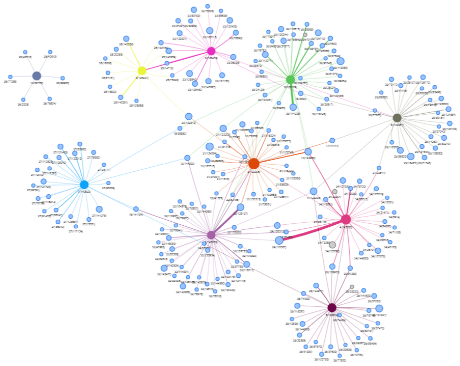


Fig. 7. 10 people with more cases in Coirón, with their relationships. Community detection by color is added.

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

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

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 [1].

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

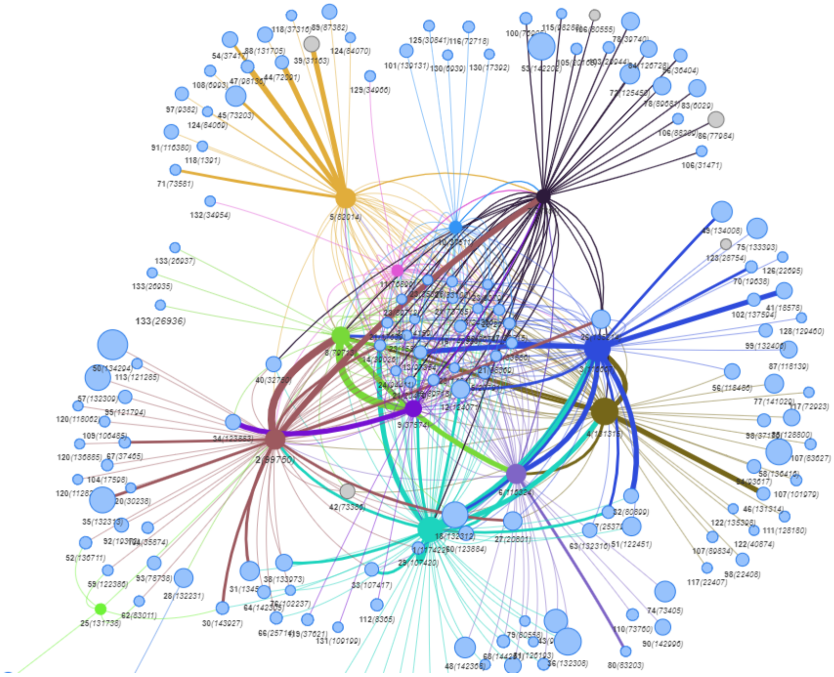


Fig. 8. PageRank execution result for the real case of LCD theft.

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

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

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).

Thanks to this improvement over the original PageRank formula, it is then possible to give greater consideration in the formula to those links that have more weights than others.

We proceeded to adapt our formula in order to contemplate the weights in the links and in this way carry out certain simple tests that verify that the modification made manages to adjust our model in an improvement with respect to the original PageRank algorithm.

Below in Figure 9 shows the example of 6 nodes shown previously where you can see the result weighting the links.

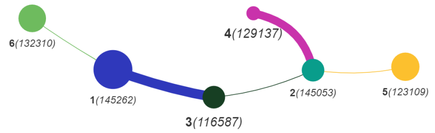


Fig. 9. PageRank run result for 6 nodes with WLRank modification.

It can be seen with the naked eye that now, having weight in the relationships in the formula, the result is much more precise, there are many common cases between 145262 and 116587, which justifies the change in the general ranking. Node 145262 has 1 in common with 132310 and 18 with 116587, a total of 19 relations to that node. On the other hand, the node that was in second position in the ranking (145053), has a total of 15 links coming from 13 cases in common with node 129137 and 1 in common with node 116587 and with node 123109.

Improvements to PageRank - Weight to nodes As mentioned in the previous point, the PageRank algorithm in its original version does not assess the weight of the links (a previously proposed and resolved issue), nor the weight of the nodes. This last point is another very important point when evaluating the individuals in the resulting graphs, as well as when it comes to the possible discovery of a supposed criminal gang.

It is almost impossible to evaluate and classify as a criminal gang individuals who have a low degree of relationship with each other, and who also each have a small number of criminal records to their credit.

To do this, we proceeded in the same way as with the weights in the links, to modify the original PageRank formula in order to obtain a more significant result and give more importance in the graph to those nodes that have more cases (greater size in the graph). graph), when they are related to others with the same number of relationships. The main objective of this modification is to generate a ranking value tiebreaker for those nodes that obtain the same score in the original pagerank algorithm. In a situation of equal ranking, the node that has a larger size will have a higher weighted value.

No dedicated bibliography was found that provides data to apply to the formula, so an own modification was made based on what has already been studied.

To begin with the weighting of the nodes, the node with the greatest weight in the graph is first identified, and then, based on the result of the PageRank formula, a new empirical value is added that arises from dividing the size of each node by the size with the highest weight already calculated, and then multiply it by the previously calculated PageRank value. Each new ranking value on each node is overwritten and becomes the final value. The SQL update algorithm is detailed below.

```
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
```

Where $Node$ is the Node table, $nodeValue$ is the weight of each node, $NodeMax$ is obtained as the maximum weight value from the Node table, and $rank$ is the previous PageRank value. When carrying out the SET, each Ranking is updated for each node.

Below is shown in Figure 10 the same example of 6 nodes shown previously where you can see the result weighting only the nodes.

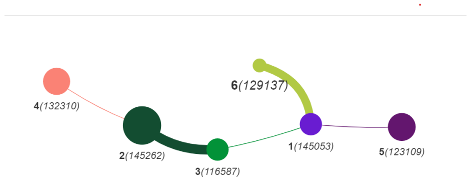


Fig. 10. PageRank execution result for 6 nodes with modification in weight of Nodes.

In it, it is possible to see, by comparing the execution of PageRank, that it maintains the main result of the original formula, but in the cases of nodes with a ranking "tie", it will now weight the one with the highest weight with the highest value. The nodes 129137 and 123109, which in the original algorithm shared fifth place in the ranking, can now be differentiated by their size (number of criminal cases for each person). The 123109 node has $68associatedcriminalcases$, being in fifth place, while the 129137 node is linked to $47criminalcases$, for which reason it is relegated to sixth place in the ranking of this graph.

Finally, Figure 11 shows the same example, but now with both modifications applied to the original PageRank formula.

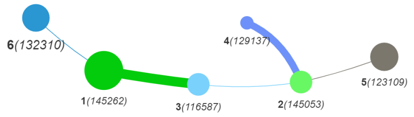


Fig. 11. PageRank execution result for 6 nodes with WLRank modification and own modification in Node weight.

7 Conclusions and future work

A study proposal of the current techniques and methodologies of intelligent data analysis and visualization for assistance in criminal investigation has been presented. All this from the records of criminal activities, their authors and the data relationships that can be derived from them. The identification of illegal networks, such as criminal gangs or criminals, was of special interest in order to promote intelligent criminal prosecution.

From the proposed study, as explained in the previous chapters, a software module was developed as a graphic tool to visualize the network of groups belonging to criminal actors, incorporating a PageRank algorithm to reflect those important people and Community detection by assigning colors. to each group/cluster of nodes.

After that, it was possible to identify that the PageRank algorithm was not sufficient when evaluating possible criminal gangs, since it does not take into account the own weights of each node and each link. In order to improve the algorithm, the necessary weightings were incorporated into the original formula both to assert the weight of the nodes as well as that of the links.

The software development was implemented in the same Public Prosecutor's Office, from which the data was taken to generate the test datasets. In this way, it has been possible not only to study the exposed techniques and methodologies, but also to achieve the production and use of the tool, by the research actors themselves.

The first impressions of those criminal investigation specialists have been very satisfactory and allow us to evaluate this work as the beginning of future visual developments to support decision-making in criminal investigation.

Currently, work continues on the development of the visualization application. It is intended to modify the original PageRank formula, enriching it with additional information according to court records. For example, the possibility of giving a higher initial ranking to those nodes that have a greater weight than others according to the records. It would also be interesting to do the same with the weight of the links between nodes, since a relationship of 2 criminal cases in common between two people is not the same as one of 18 cases in common. In this way we would also be able to give more ranking to those nodes that are related to others, in a greater number of criminal cases. This is certainly relevant to criminal investigation based on criminal records. It will also seek to delve into various centrality algorithms. There are some notions that are relevant for identifying the importance of a person in the network induced by criminal cases. For example, *betweenness centrality*, which models the extent to which a particular node is between other nodes in a network, or *closeness centrality*, which is the inverse of the sum of shortest paths (geodesics) that connect a particular node to all other nodes in a network [34]. Similarly, *eigenvector centrality*, is another way of assigning centrality to a network actor based on the idea that if a node has many central neighbors, it should also be central.

References

1. Baeza-Yates, R., Davis, E.: Web page ranking using link attributes. In: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters. pp. 328–329 (2004)
2. Bichler, G., Malm, A.: Small arms, big guns: a dynamic model of illicit market opportunity. Global Crime **14**(2-3), 261–286 (2013)

3. Bouchard, M., Amirault, J.: Advances in research on illicit networks. Global crime **14**(2-3), 119–122 (2013)
4. 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)
5. 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)
6. Burcher, M.: Social network analysis and law enforcement: Applications for intelligence analysis. Springer (2020)
7. Carley, K.M.: Destabilization of covert networks. Computational & Mathematical Organization Theory **12**(1), 51–66 (2006)
8. 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)
9. Colladon, A.F., Remondi, E.: Using social network analysis to prevent money laundering. Expert Systems with Applications **67**, 49–58 (2017)
10. Décary-Hétu, D.: Information exchange paths in irc hacking chat rooms. Crime and Networks. New York: Routledge pp. 218–230 (2014)
11. Décary-Hétu, D., Dupont, B.: The social network of hackers. Global Crime **13**(3), 160–175 (2012)
12. Décary-Hétu, D., Dupont, B.: Reputation in a dark network of online criminals. Global Crime **14**(2-3), 175–196 (2013)
13. Dubinsky, E.: Mathematical structures for computer science. by judith l. gersting. The American Mathematical Monthly **91**(6), 379–381 (1984)
14. 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)
15. 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`
16. Fiscal, M.P.: Página web. `https://www.mpfchubut.gov.ar/`
17. Franceschet, M.: Pagerank: Standing on the shoulders of giants. Communications of the ACM **54**(6), 92–101 (2011)
18. 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)
19. Göbel, F., Jagers, A.: Random walks on graphs. Stochastic processes and their applications **2**(4), 311–336 (1974)
20. Harper, W.R., Harris, D.H.: The application of link analysis to police intelligence. Human Factors **17**(2), 157–164 (1975)
21. 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`
22. 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)
23. Krebs, V.E.: Mapping networks of terrorist cells. Connections **24**(3), 43–52 (2002)
24. Lauchs, M., Keast, R., Yousefpour, N.: Corrupt police networks: uncovering hidden relationship patterns, functions and roles. Policing & society **21**(1), 110–127 (2011)
25. Lin, J., Dyer, C.: Data-intensive text processing with mapreduce. Synthesis Lectures on Human Language Technologies **3**(1), 1–177 (2010)

26. 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)
27. Ma, X., et al.: Exploring sharing patterns for video recommendation on youtube-like social media. Multimedia Systems **20**(6), 675–691 (2014)
28. 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)
29. 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)
30. McGloin, J.M.: Policy and intervention considerations of a network analysis of street gangs. Criminology & Public Policy **4**(3), 607–635 (2005)
31. Medina, R.M.: Social network analysis: a case study of the islamist terrorist network. Security Journal **27**(1), 97–121 (2014)
32. Morselli, C.: Hells angels in springtime. Trends in organized crime **12**(2), 145–158 (2009)
33. Morselli, C.: Assessing vulnerable and strategic positions in a criminal network. Journal of Contemporary Criminal Justice **26**(4), 382–392 (2010)
34. Newman, M.E.: A measure of betweenness centrality based on random walks. Social networks **27**(1), 39–54 (2005)
35. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)
36. 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)
37. Rua, G., González, L.: Sistemas judiciales: Una perspectiva integral sobre la administración de justicia. Análisis criminal en América Latina **23** (2020)
38. Scott, J., Carrington, P.J.: The SAGE handbook of social network analysis. SAGE publications (2011)
39. Soudijn, M.R.: Using strangers for money: A discussion on money-launderers in organized crime. Trends in organized crime **17**(3), 199–217 (2014)
40. 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)
41. visjs: Página web. `https://visjs.org/`
42. Wasserman, S., Faust, K., et al.: Social network analysis: Methods and applications (1994)
43. Xu, J., Chen, H.: Criminal network analysis and visualization. Communications of the ACM **48**(6), 100–107 (2005)