UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

ANÁLISIS DE REDES SOCIALES PARA MEJORAR LA EFICIENCIA EN LA PERSECUCIÓN PENAL

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

VALENTINA VERÓNICA REYES POZO

PROFESOR GUÍA:
Richard Weber Haas

MIEMBROS DE LA COMISIÓN:
Juan Romero Godoy
David Salinas Fuentes

SANTIAGO DE CHILE
2024

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL
POR: VALENTINA VERÓNICA REYES POZO
FECHA: 2024
PROF. GUÍA: RICHARD WEBER HAAS

ANÁLISIS DE REDES SOCIALES PARA MEJORAR LA EFICIENCIA EN LA PERSECUCIÓN PENAL

El proceso de persecución penal, liderado por el Ministerio Público, es esencial para la administración de la justicia. El Sistema de Análisis Criminal y Focos Investigativos nace en 2015 a partir de la ley que fortalece el Ministerio Público, y tiene por objetivo robustecer la Persecución Penal en delitos de mayor connotación social. SACFI propone una nueva lógica de investigación, donde a partir del análisis agrupado de causas que forman parte de un mismo problema delictual, se logran mejores resultados en la identificación de imputados que en un inicio son desconocidos, así como en las penas otorgadas a quienes infringen la ley de forma reiterada.

Para realizar un Análisis Criminal más eficaz y eficiente, SACFI se encuentra en constante búsqueda de herramientas que permitan mejorar sus procesos. Durante el año 2023, se presentó una nueva herramienta tecnológica desarrollada por la Universidad de Chile, Universidad del Bío-Bío y la Universidad de los Andes en colaboración con el Ministerio Público, que, basándose en el modelamiento de las interacciones entre personas con historial delictual, permite identificar potenciales miembros que puedan haber participado de un delito contra la propiedad. La aplicación de esta herramienta requiere de conocer al menos uno de los imputados, y a partir de este, reconstruir la posible agrupación criminal. Sin embargo, una de sus limitaciones es que para delitos contra la propiedad, existe una baja cantidad de imputados conocidos al momento de iniciar la investigación de un delito. Lo que hace que la aplicación de los modelos ya desarrollados no sea factible.

El presente trabajo de título muestra un nuevo modelo que permite trazar líneas investigativas en causas que no cuentan con imputados conocidos en un inicio. Además, el modelo propone una forma de mejorar la eficiencia en las actividades operativas que implica la persecución penal, disminuyendo los tiempos de análisis de grandes cantidades de información, identificando patrones entre delitos, y proponiendo una forma de enriquecer la información de causas que inicialmente, no cuentan con suficientes antecedentes que permitan seguir líneas investigativas. Para esto, se utilizaron técnicas de Modelamiento y Análisis de Redes Sociales, Optimización Lineal, *Large Language Models* o Modelos Grandes de Lenguaje con el fin de facilitar el análisis y la identificación de patrones delictuales.

Los resultados del modelo validan la hipótesis que sugiere una correlación entre delitos similares y la participación de individuos en tales delitos, en el caso de robos contra la propiedad. Además, se concluye que el modelo propuesto permite reducir en más de un 80 % las actividades de selección y análisis de causas que den cuenta de un problema delictual. El modelo abre la posibilidad de encontrar potenciales individuos que hayan cometido delitos que, en un principio, no cuentan con imputados conocidos.

No es suficiente afirmar que la justicia tarda pero llega,
la justicia que no se ejerce cuando corresponde, ya es injusta.

-Pierre Dubois

Agradecimientos

Quiero agradecer a mi familia por todo el amor, preocupación y entrega a lo largo de todos estos años. A mis padres, por su esfuerzo y entrega constante, por transmitirme valores como la empatía, la responsabilidad y la perseverancia, que me han permitido llegar hasta acá. Agradezco infinitamente a Romain por ser parte de todo este proceso, por escucharme, por entenderme y por motivarme siempre en los momentos más difíciles. A mis amigos de la vida, Nicolás, Sofía y Bastián por brindarme su cariño y preocupación.

Agradezco a todas aquellas personas que me rodearon y fueron parte de mi vida estudiantil, profesores, familiares y amigos que me han inspirado y motivado a lo largo de todos estos años a seguir superándome.

A los profesores que me formaron y me entregaron herramientas valiosísimas tanto en lo académico como en lo humano. En especial, agradezco al Profesor Richard Weber por su calidad humana y académica, por ser una fuente de inspiración, por confiar y guiar mi trabajo. Agradezco al Profesor Juan Pablo Romero, quien estuvo siempre disponible a escuchar mis inquietudes y a acompañarme en esta etapa.

Agradecer a mi compañero Alex, por su plena disposición y generosidad para compartir sus conocimientos, por su compañía y por su preocupación en mi trabajo.

Agradezco profundamente al equipo SACFI, quienes me abrieron las puertas para aprender día a día, y estuvieron siempre ahí para atender mis dudas, corregir y validar constantemente este trabajo. En especial agradezco a David, por su profesionalismo y calidez humana, que hicieron de esto una experiencia muy enriquecedora.

Agradezco finalmente a todas las instituciones y organismos que me brindaron oportunidades únicas para seguir aprendiendo, en especial, a la Universidad de Chile, a l'École Centrale de Marseille y al Ministerio Público.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Delincuencia

La delincuencia es un fenómeno social complejo y multifacético que representa un desafío persistente para las sociedades en el mundo. Además, el impacto que tiene en la sociedad es significativo ya que afecta la seguridad, el bienestar y la calidad de vida de las personas. Es por esto y otros factores que este tema se ha posicionado como la mayor preocupación de los chilenos en el último tiempo.

En la encuesta 'What worries the world?', realizada periódicamente por Ipsos Global Advisor, se evidencian los temas que más le preocupan a los habitantes de cada país. En particular, se constata que Chile está entre los 5 países que consideran este tema como su mayor inquietud, bajo la pregunta: '¿Cuáles de los siguientes tres temas le parecen más preocupantes en su país?'. [1]

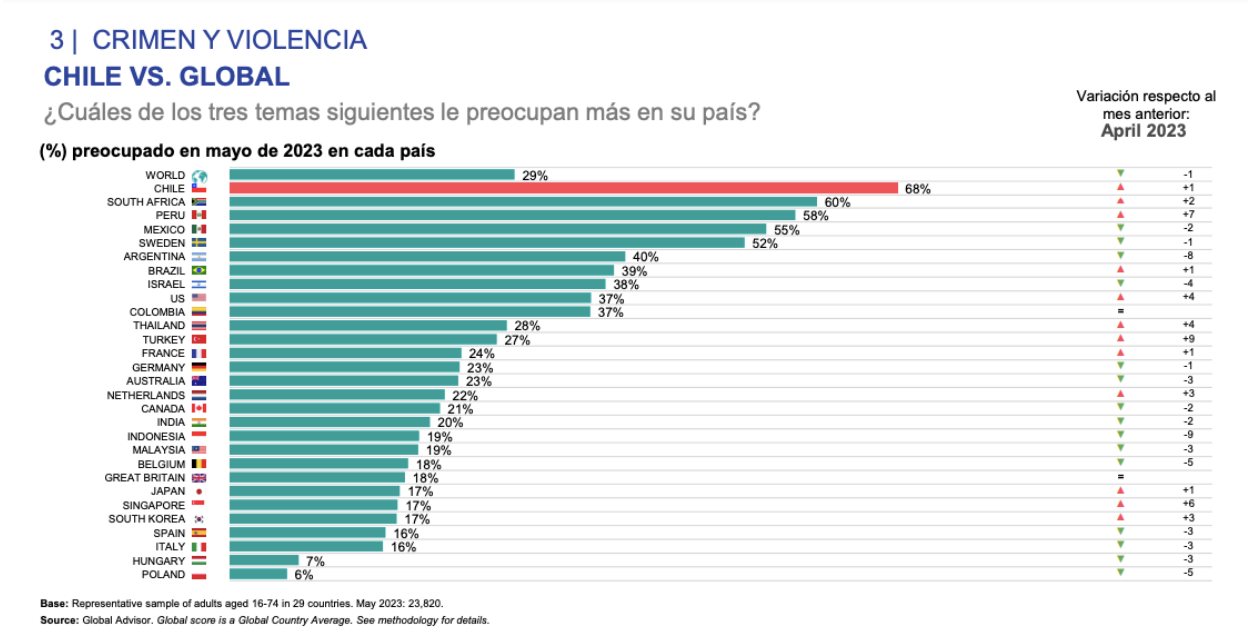


Figura 1.1: Porcentaje de preocupación por la delincuencia según cada país. [1]

Aunque Chile no sea uno de los países con las cifras más elevadas de delincuencia en el cono sur, la creciente preocupación por este fenómeno podría deberse a varios factores. Algunos autores señalan que la prensa, la agenda política y la variación de las estadísticas entre los años 2021 y 2022 podrían explicar, en parte, la preocupación por este fenómeno. En la tabla 1.2 se observa el aumento en el número de casos policiales[1] entre 2021 y 2022, para los delitos de mayor connotación social[2], según el Centro de Estudios y Análisis del Delito. [3]

	2021	2022
GRUPO DELICTUAL / DELITO		
Delitos de mayor connotación social	**288.970**	**425.342**
Hurtos	59.165	86.498
Robo con violencia o intimidación	44.947	74.028
Robo de objetos de o desde vehículo	35.227	50.410
Robo de vehículo motorizado	23.441	33.024
Robo en lugar habitado	28.440	41.398
Robo en lugar no habitado	26.381	41.543
Robo por sorpresa	16.575	27.300

Figura 1.2: Cantidad de denuncias en los años 2021, 2022 a nivel nacional. [3]

Al analizar el aumento porcentual de cada uno de estos delitos, se evidencia una de las causas de preocupación en términos de seguridad pública.

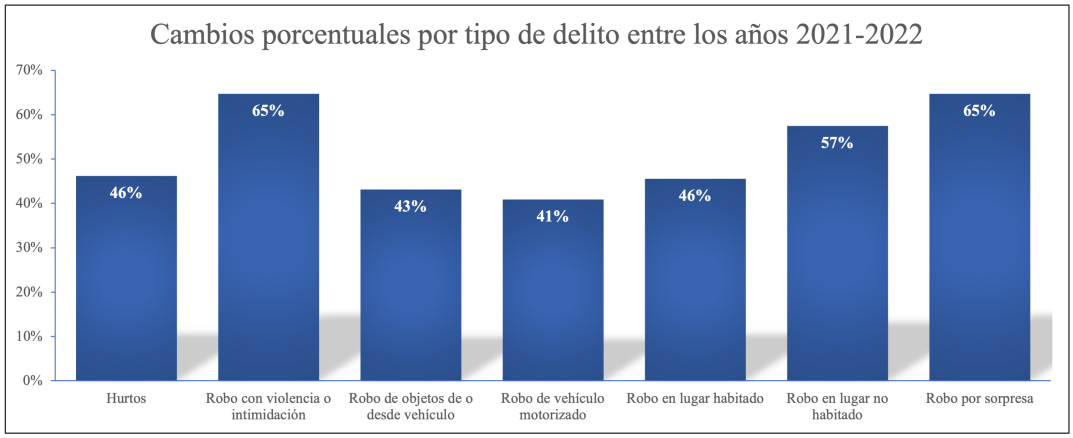


Figura 1.3: Variación porcentual de denuncias entre los años 2021, 2022 a nivel nacional. [Fuente: elaboración propia]

Es especialmente inquietante el aumento en dos tipos de delitos: el robo con violencia o intimidación y el robo por sorpresa, con un incremento del 65 %. Estas cifras generan preocupación tanto entre la población como en las autoridades, y hace necesaria la toma de medidas para abordar esta problemática de manera efectiva.

[1] El indicador de casos policiales considera denuncias de delitos que realiza la comunidad en las unidades policiales junto con las detenciones que realizan las policías ante la ocurrencia de delitos flagrantes.

[2] Los delitos de mayor connotación social son aquellos que generan mayor repercusión social y mediática ya que producen un efecto generalizado de inseguridad [2]. Entre los delitos de mayor connotación social se consideran: homicidio, lesiones, violación, robo con fuerza y robo con violencia.

1.2. Ministerio Público de Chile

En el contexto de la lucha contra la delincuencia, la persecución penal desempeña un papel fundamental, abarcando actividades como la averiguación de hechos delictivos, la búsqueda de sanciones para los responsables y la protección de los intereses de las víctimas. [4]

Desde el año 2000, la Fiscalía de Chile, oficialmente conocida como el Ministerio Público, es la institución autónoma encargada de llevar a cabo la persecución penal y la investigación de los delitos. Este organismo autónomo, que no forma parte de ninguno de los tres poderes del Estado, tiene la misión de dirigir de manera exclusiva y objetiva las investigaciones penales y ejercer la acción penal pública dentro del marco legal establecido, considerando los intereses de las víctimas y de la sociedad. [5]

1.2.1. Organigrama

El Ministerio Público consta de una Fiscalía Nacional, y 19 Fiscalías Regionales a lo largo del país. Si bien existen solo 16 regiones geográficas, se contemplan 19 Fiscalías Regionales ya que la Región Metropolitana, dada su extensión y densidad de habitantes, se divide en 4 Fiscalías Regionales.

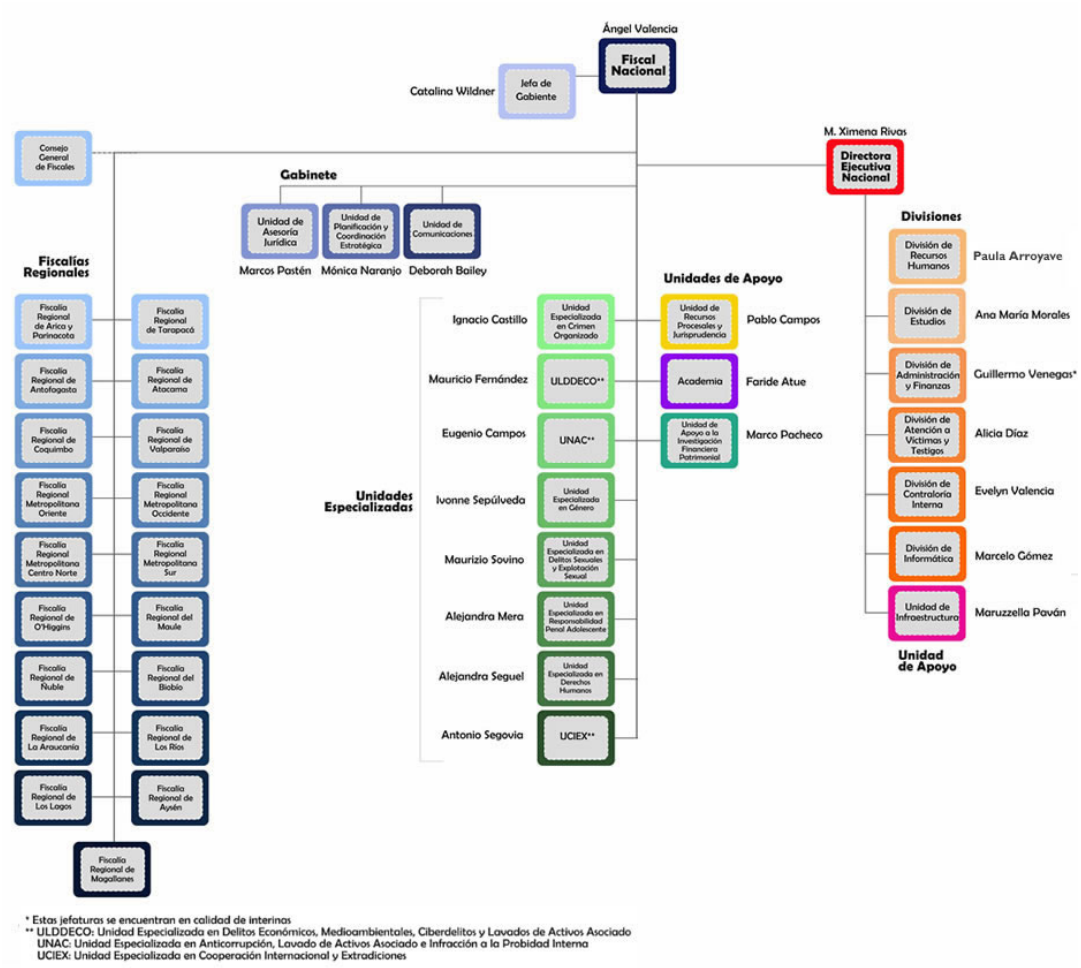


Figura 1.4: Organigrama del Ministerio Público, agosto 2023. [6]

Cada Fiscalía Regional está bajo la dirección de un Fiscal Regional, quien asume la responsabilidad de liderar las operaciones del Ministerio Público en su área geográfica asignada. Por su parte, la Fiscalía Nacional desempeña un papel fundamental como entidad coordinadora entre las diversas Fiscalías Regionales, estableciendo los lineamientos estratégicos y operativos a nivel nacional. En la Fiscalía Nacional se establecen múltiples unidades especializadas, enfocadas en áreas delictivas específicas tales como: Crimen Organizado, Delitos Sexuales y Explotación Sexual, y Derechos Humanos, entre otras, que se detallan en el organigrama de la figura 1.4. Además, se han establecido divisiones encargadas de gestionar diversos aspectos funcionales del Ministerio Público, tales como Administración y Finanzas, Recursos Humanos y Tecnología de la Información, también reflejadas en el organigrama. Tanto las unidades de apoyo como las divisiones tienen como objetivo brindar respaldo al funcionamiento eficiente de la Fiscalía y servir de apoyo al Fiscal Nacional en el cumplimiento de sus funciones.

En el marco de este trabajo, se destaca la División de Estudios que tiene como misión "Asesorar al Fiscal Nacional respecto de la gestión, mediante la evaluación y control del quehacer institucional y la realización de los estudios necesarios para ello."[7] Actualmente, la División de estudios, evaluación, control y desarrollo de la gestión tiene a su cargo al Sistema de Análisis Criminal y Focos Investigativos, en adelante, SACFI.

1.3. Sistema de Análisis Criminal y Focos Investigativos (SACFI)

Con el objetivo de fortalecer las labores del Ministerio Público, en 2015 se creó el área de Sistema de Análisis Criminal y Focos Investigativos (SACFI). Este sistema busca contribuir a la persecución penal mediante el robustecimiento de estrategias de análisis e investigación sobre estructuras de criminalidad reconocibles [8]. Gracias a la agrupación de conjuntos delictivos en áreas de investigación focalizadas, se busca mejorar la efectividad y eficiencia en la persecución penal.

Las labores de SACFI se enfocan, actualmente, en la persecución de delitos contra la propiedad y aquellos de mayor connotación social. Sin embargo, los delitos se recalifican anualmente por orden del Fiscal Nacional. A modo de ejemplo, se muestran los delitos que fueron objeto de investigación SACFI en el año 2017, en la figura 1.5. [9]

Tabla 8: *delitos calificados que son objeto de investigación SACFI*[7]

#	Delitos calificados SACFI	Resolución de calificación
1	Porte de arma prohibida (art. 14 inc. 1°)	Res. FN/MP N°2.500, de 29 de diciembre de 2016
2	Posesión, tenencia, porte armas art 9 inc 1 ley 17779	
3	Tenencia de armas prohibidas art. 13	
4	Tráfico de armas (art. 10)	
5	Hurto agravado (art. 447 código penal)	Res. FN/MP N°2.438, de 31 de diciembre de 2015
6	Hurto de bienes pertenecientes a redes de suministro público	
7	Hurto de hallazgo	
8	Hurto simple por un valor de 4 a 40 UTM	
9	Hurto simple por un valor de media a 4 UTM	
10	Hurto simple por un valor sobre 40 UTM	
11	Infracción ley 11.564 de mataderos clandestinos. Art. 1.	
12	Abigeato	
13	Receptación. Art. 456 bis a	
14	Robo con intimidación. Art. 433, 436 inc 1º 438.	
15	Robo con violencia	
16	Robo por sorpresa. Art. 436 inc. 2°	
17	Robo con fuerza de cajeros automáticos	
18	Robo de vehículo motorizado art. 443 inc. 2	
19	Robo en bienes nacionales de uso público o sitios no destinados a habitación	
20	Robo en lugar habitado o destinado a la habitación. Art. 440	
21	Robo en lugar no habitado. Art. 442.	
22	Otros delitos	Res. FN/MP N°2.514, de 28 de diciembre de 2017

Fuente: Ministerio Público

Figura 1.5: Delitos calificados que son objeto de investigación SACFI al año 2017. [9]

El Sistema de Análisis Criminal y Focos Investigativos se estructura con dos niveles organizacionales. A nivel nacional existe una Unidad Coordinadora, que se encuentra en las dependencias de la Fiscalía Nacional. Esta unidad coordinadora tiene como objetivo gestionar el actuar de SACFI a través de las distintas regiones, así como brindar asesoría al Fiscal Nacional.

Por otra parte, cada fiscalía Regional cuenta con una unidad SACFI. Cada unidad funciona con dos equipos: una unidad de Análisis Criminal compuesta por analistas, y otra de Focos Investigativos compuesta por Fiscales Adjuntos de Focos. Cada unidad regional de SACFI está encabezada por un Fiscal Jefe de Focos, que a su vez está bajo el mando del Fiscal Regional.

1.3.1. Procesos de trabajo

La metodología de trabajo que propone SACFI se basa en el concepto de "Foco investigativo", definido como "un conjunto de delitos de igual o distinta naturaleza, que forman parte de un problema delictual donde se podrían identificar una o más estructuras de criminalidad reconocible" [8]. De esta manera, se busca dejar la lógica de investigación del "caso a caso", agrupando delitos que permitan tener un mejor entendimiento del problema, facilitando la obtención de mejores resultados en la persecución penal. El diagrama 1.6 ilustra las principales etapas que contempla el proceso de trabajo de los equipos:

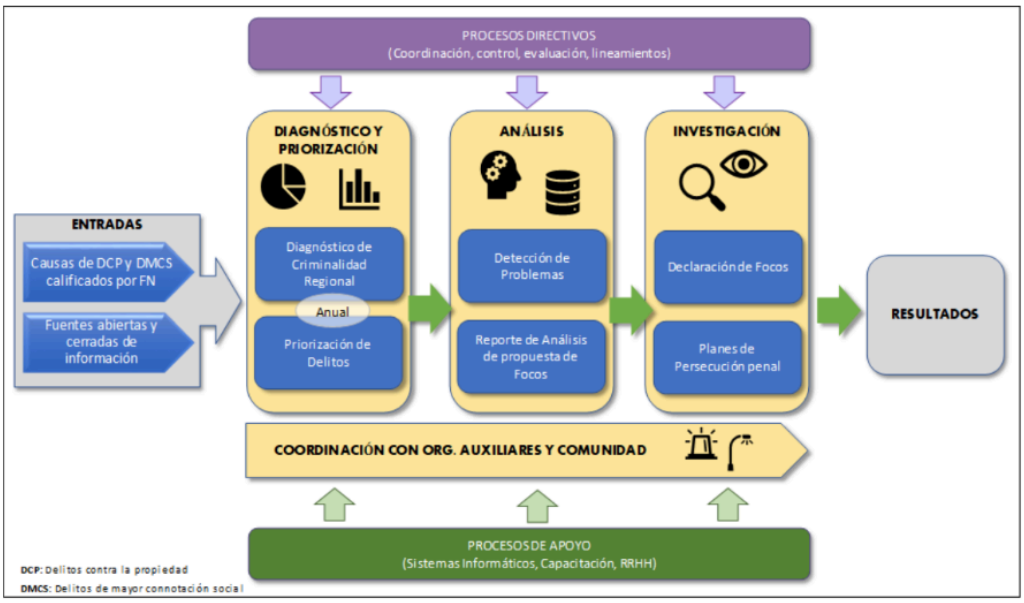


Figura 1.6: Diagrama de los procesos de trabajo llevados a cabo por los diferentes equipos SACFI a nivel regional. [Fuente: Documentación SACFI]

Secuencialmente, estos procesos de trabajo suceden de la siguiente forma:

1. El Fiscal Nacional declara cuáles son los delitos prioritarios para SACFI.
2. Una vez al año, las unidades SACFI elaboran su Diagnóstico de Criminalidad Regional, en el cual se declaran los delitos que serán priorizados en función de los lineamientos del Fiscal Nacional y la realidad local.
3. Una vez declarados, los analistas criminales comienzan a investigar para encontrar información que pueda ser valiosa en la declaración de un foco investigativo.
4. Los analistas presentan un Reporte de Análisis de Propuesta de Foco que es presentado al Fiscal Jefe de Focos de cada región.
5. El Fiscal Regional decide qué focos declarar basándose en el insumo presentado por los analistas.
6. Una vez que se han declarado los focos, se asigna a cada uno de estos, un Fiscal de Focos para construir el Plan de Persecución penal.
7. Se continúa con el monitoreo del foco a través de Reportes de Monitoreo.
8. Finalmente, se construye un informe de Término del Plan de Persecución penal que resume los resultados de la investigación y persecución penal del foco.

Dentro de este proceso, se desarrollan distintos productos, entre los cuales destacan Diagnósticos de Criminalidad Regional, Informes de Comportamiento Criminal Regional, Reportes de Análisis de Propuesta de Focos, y Planes de Persecución Penal.

1.3.2. Herramientas tecnológicas en SACFI

Desde la unidad coordinadora SACFI se han elaborado distintas herramientas tecnológicas que buscan sistematizar la información y hacerla más accesible para los equipos a nivel nacional. Entre estas, se encuentran herramientas de analítica avanzada de relatos, análisis georreferencial, y otras enfocadas al análisis de imágenes y vídeos.

Más recientemente, en el primer semestre de 2023 se presentó la nueva herramienta, "Fiscal Heredia", gracias al proyecto de investigación FONDEF "Sistema de analítica integrada de información para la persecución de delitos contra la propiedad: inteligencia artificial para detectar estructuras criminales" en el cual participaron distintas universidades, entre ellas la Universidad de Chile, Universidad del Bío Bío y la universidad de los Andes, además de la Agencia Nacional de Investigación y Desarrollo (ANID), y el Ministerio Público. Esta herramienta se basa en modelos de Redes Sociales y optimización para identificar la red criminal más probable de un individuo, y así focalizar la investigación de los Fiscales.

A la fecha, "Fiscal Heredia" cuenta con dos modelos de optimización: StRAM[3] y LiRAM[4], que operan según el número de sujetos identificados inicialmente. De forma general, los modelos buscan encontrar la banda criminal más probable que tendría un individuo que acaba de cometer un delito. Primeramente, se utilizan Redes Sociales para representar las relaciones entre individuos que registren delitos en las bases de datos del Ministerio Público. Así, cada individuo con historial delictivo se representa con un nodo, y existe un arco entre ellos si es que estos participaron de manera conjunta en un mismo delito en el pasado. Además, se determina el valor *'pcg'* o 'Propensión a pertenecer a una banda criminal' de cada individuo, calculado entre otras cosas según la actividad delictual histórica versus la actividad delictual reciente que registre cada uno. Con esta información, los modelos optimizan una función de utilidad que representa la ganancia que tendría la supuesta banda criminal, considerando un planificador que busca maximizar las ganancias entre sujetos en los cuales exista confianza y maximicen el nivel de *pcg*[5]. De esta forma, los individuos que tengan mayor actividad delictual recientemente, y que se encuentren en las capas de interacción cercanas de un sospechoso, tendrán más probabilidad de pertenecer a una banda criminal que otros individuos que tengan muchas causas o delitos históricamente, pero que en los dos últimos años no tengan causas recientes. [10] [11]

Si bien los modelos entregan resultados bastante buenos que aportan información para guiar la investigación, también se han identificado muchas áreas de mejora. En primer lugar, en la actualidad, el uso de "Fiscal Heredia" requiere conocer al menos un individuo que haya participado en el delito que se esté investigando. Además, existen espacios de mejora en la forma de calcular el parámetro *pcg*, incorporando más información de las bases de datos del Ministerio Público que logren perfilar al individuo según el tipo de delitos cometidos, o bien enriqueciendo los arcos de la red social con otro tipo de relaciones visibles en las distintas fuentes de datos.

[3] Steiner Tree Rational Association Model.
[4] Linear Rational Association Model.
[5] Propensity to belong to a criminal group.

1.4. Descripción del proyecto y justificación

1.4.1. Modelos de redes sociales y sus limitaciones actuales

Los modelos que se han desarrollado hasta la fecha se basan en algoritmos que requieren conocer al menos un imputado en una causa delictiva, para luego encontrar individuos que puedan estar relacionados, en una causa determinada.

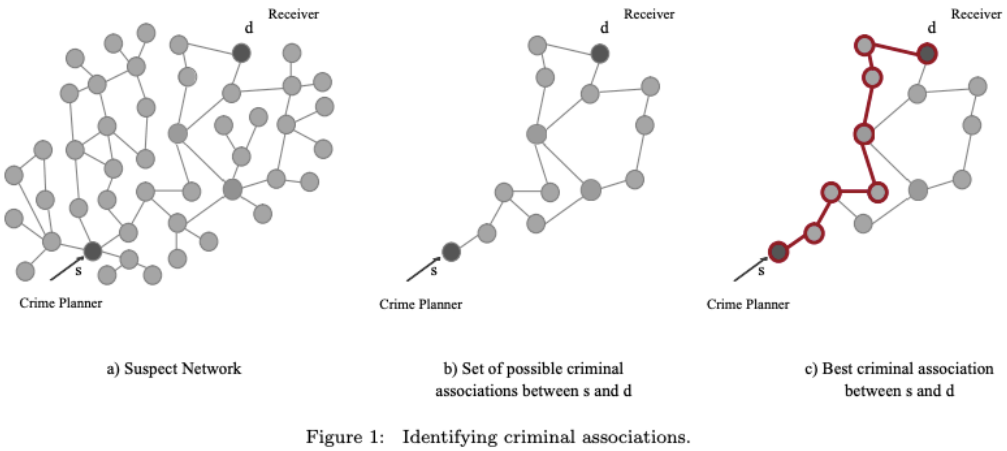


Figure 1: Identifying criminal associations.

Figura 1.7: Representación gráfica de los modelos de SNA actuales. [11]

En la figura 1.7, se puede apreciar una red social generada a partir de dos individuos iniciales conocidos, denominados 'Planificador' (*Crime Planner*) y 'Receptor' (*Receiver*). El modelo LiRAM [11], escoge el mejor camino que conecta ambos nodos, considerando los mayores valores de *pcg* que al mismo tiempo minimicen la distancia entre los sujetos. De esta forma, el modelo sugiere los individuos que podrían pertenecer a la banda criminal.

Para llegar a este resultado, el modelo optimiza una función que representa la utilidad que tendría el planificador del delito, considerando el beneficio de agregar un nuevo integrante a la banda dado su grado de cercanía y propensión a pertenecer a una banda criminal (*pcg*).

Tal como se mencionó en la subsección 1.3.2, una de las principales limitaciones de los modelos StRAM y LiRAM de 'Fiscal Heredia', es la necesidad de tener al menos un involucrado conocido para obtener los posibles integrantes de una banda criminal. Sin embargo, según el Boletín Estadístico del primer trimestre de 2023, emitido por el Ministerio Público [12], para más de la mitad de los delitos ingresados no se cuenta con imputados conocidos en un inicio. Más aún, existen delitos donde este fenómeno se incrementa, por ejemplo, en las categorías de hurtos, la cifra llega a un 74,97 % de imputados desconocidos y alcanza su nivel más alto en el delito de robos no violentos con un 93,31 %.

CATEGORÍA DE DELITOS	IMPUTADOS				
	Imputado conocido (IC)	% Conocido	Imputado Desconocido (ID)	% Desconocido	Total
CUASIDELITOS	1.964	65,73%	1.024	34,27%	2.988
DELITOS CONTRA LA FE PÚBLICA	6.798	75,52%	2.204	24,48%	9.002
DELITOS CONTRA LA LIBERTAD E INTIMIDAD DE LAS PERSONAS	33.054	61,81%	20.426	38,19%	53.480
DELITOS CONTRA LEYES DE PROPIEDAD INTELECTUAL E INDUSTRIAL	378	83,81%	73	16,19%	451
DELITOS DE JUSTICIA MILITAR	33	91,67%	3	8,33%	36
DELITOS DE LEYES ESPECIALES	10.336	73,00%	3.822	27,00%	14.158
DELITOS DE TORTURA, MALOS TRATOS, GENOCIDIO Y LESA HUMANIDAD	85	11,63%	646	88,37%	731
DELITOS ECONÓMICOS Y TRIBUTARIOS	10.517	40,54%	15.428	59,46%	25.945
DELITOS FUNCIONARIOS	308	45,70%	366	54,30%	674
DELITOS LEY DE DROGAS	4.562	49,15%	4.720	50,85%	9.282
DELITOS LEY DE TRÁNSITO	11.848	89,09%	1.451	10,91%	13.299
DELITOS SEXUALES	5.919	55,48%	4.749	44,52%	10.668
FALTAS	9.498	77,74%	2.719	22,26%	12.217
HECHOS DE RELEVANCIA CRIMINAL	5.362	14,94%	30.533	85,06%	35.895
HOMICIDIOS	452	57,80%	330	42,20%	782
HURTOS	7.979	25,03%	23.896	74,97%	31.875
LESIONES	29.451	63,16%	17.180	36,84%	46.631
OTROS DELITOS	5.140	74,36%	1.772	25,64%	6.912
OTROS DELITOS CONTRA LA PROPIEDAD	9.920	35,58%	17.961	64,42%	27.881
ROBOS	2.308	8,36%	25.298	91,64%	27.606
ROBOS NO VIOLENTOS	3.445	6,69%	48.044	93,31%	51.489
TOTAL NACIONAL	159.357	41,72%	222.645	58,28%	382.002

Figura 1.8: Delitos ingresados por categoría de delitos y tipo de imputado. [12]

Sumado a esto, muchos actores internos del Ministerio Público reconocen que uno de los resultados más relevantes del SACFI es la identificación de sujetos desconocidos [9]. Es por esto, que colaborar en esta línea es de gran relevancia para mejorar la eficacia y eficiencia en la persecución penal, y contribuir en la misión de SACFI.

1.4.2. Términos aplicados a las causas

En agosto de 2019 se emitió un Informe que buscaba evaluar el Sistema de Análisis Criminal y Focos Investigativos al año 2018, realizado por el Centro de Microdatos de la Facultad de Economía y Negocios de la Universidad de Chile [9]. El informe evalúa los resultados logrados por SACFI al año 2018, con el objetivo de reconocer las opciones de mejora de las prácticas y favorecer una mejor instalación del sistema. Dentro de los aspectos evaluados, se mencionan entre otros, el desempeño investigativo en lograr identificar imputados, el desempeño investigativo en lograr generación de información útil para la judicialización, y el desempeño investigativo en lograr términos de calidad, dado que estos dan cuenta de los objetivos que persigue SACFI.

En torno a las tres aristas de desempeño evaluadas, las cifras de imputados conocidos se evidencian en la subsección 1.4.1. Por otra parte, el desempeño en 'lograr términos de calidad' se refiere a los logros en materia penal que se obtuvieron mediante el tratamiento de causas bajo los equipos SACFI. Para evidenciar el problema u oportunidad que surge en esta área, se muestran en la figura 1.9, las cifras del primer trimestre del año 2023 en cuanto al término de causas por región, específicamente para delitos de robo:

	TIPO DE TÉRMINOS	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIV	XV	XVI	RM CN	RM OR	RM OCC	RM SUR	Total Nacional
ROBOS	SENTENCIA DEFINITIVA CONDENATORIA	109	57	48	106	289	67	62	102	85	42	5	4	15	31	15	511	259	238	229	2.274
	SENTENCIA DEFINITIVA ABSOLUTORIA	6	4	2	1	31	9	4	9	1	0	0	0	0	2	0	9	26	14	8	126
	SOBRESEIMIENTO DEFINITIVO	0	1	0	3	8	2	2	3	3	12	1	0	0	0	1	29	12	8	10	95
	SOBRESEIMIENTO TEMPORAL	6	7	3	4	15	9	10	5	5	10	0	0	1	0	1	21	11	3	12	123
	SUSPENSIÓN CONDICIONAL DEL PROCEDIMIENTO	0	1	0	6	13	2	1	10	1	13	0	0	0	4	1	28	5	8	12	105
	SOBRESEIMIENTO DEFINITIVO 240	0	0	1	2	5	2	1	1	0	3	0	0	0	1	0	0	3	0	0	19
	ACUERDO REPARATORIO	0	1	2	2	2	0	1	0	1	6	0	0	3	0	0	1	0	1	0	20
	FACULTAD DE NO INVESTIGAR	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	3	1	1	0	8
	SUBTOTAL SALIDA JUDICIAL	**121**	**71**	**56**	**124**	**365**	**91**	**82**	**130**	**96**	**86**	**6**	**4**	**19**	**38**	**18**	**602**	**317**	**273**	**271**	**2.770**
	ARCHIVO PROVISIONAL	715	664	211	768	1.868	393	375	1.206	327	316	8	13	98	477	396	8.179	2.815	2.507	2.969	24.305
	DECISIÓN DE NO PERSEVERAR	19	31	18	41	91	41	8	29	25	13	2	0	2	32	8	82	78	29	39	588
	PRINCIPIO DE OPORTUNIDAD	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	3
	INCOMPETENCIA	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	2	4
	SUBTOTAL SALIDA NO JUDICIAL	**734**	**695**	**229**	**809**	**1.960**	**435**	**383**	**1.236**	**353**	**329**	**10**	**13**	**100**	**509**	**404**	**8.261**	**2.894**	**2.536**	**3.010**	**24.900**
	ANULACIÓN ADMINISTRATIVA	0	1	0	0	7	2	0	14	6	1	0	3	0	0	1	36	10	3	7	91
	AGRUPACIÓN A OTRO CASO	25	37	44	34	104	32	24	60	25	34	0	5	18	11	19	106	117	37	36	768
	OTRAS CAUSALES DE TÉRMINO	0	2	0	0	5	1	0	1	0	1	0	1	0	4	1	7	1	2	0	26
	OTRAS CAUSALES DE SUSPENSIÓN	1	3	0	2	10	6	0	4	0	1	0	0	2	2	2	5	4	2	3	47
	SUBTOTAL OTROS TÉRMINOS	**26**	**43**	**44**	**36**	**126**	**41**	**24**	**79**	**31**	**37**	**0**	**9**	**20**	**17**	**23**	**154**	**132**	**44**	**46**	**932**
	TOTAL POR ROBOS	**881**	**809**	**329**	**969**	**2.451**	**567**	**489**	**1.445**	**480**	**452**	**16**	**26**	**139**	**564**	**445**	**9.017**	**3.343**	**2.853**	**3.327**	**28.602**

Figura 1.9: Resumen de Términos aplicados por región, en el primer trimestre 2023. [12]

De la tabla 1.9, se observa que los términos más comunes corresponden a: 'archivo provisional' y 'sentencia definitiva condenatoria', representando un 84,97 % y 7,95 % respectivamente del total de términos aplicados a nivel nacional en los delitos de robos. Más específicamente, es posible observar que el tipo de término se clasifica en tres categorías: 'salida judicial', 'salida no judicial' u 'otros términos'. Al mirar las salidas no judiciales, para menos del 0,1 % de las causas a nivel nacional se aplicó el término 'principio de oportunidad' o el término 'incompetencia'. El primero se aplica si es que el hecho no compromete el interés público. El término 'decisión de no perseverar' representa el 2,05 % del total de términos aplicados a nivel nacional, este término se puede aplicar cuando no se reúnen suficientes antecedentes para presentar una acusación. Finalmente, el 84,97 % de los términos aplicados a nivel nacional corresponden al 'archivo provisional', facultad que tiene el Ministerio Público de archivar la causa cuando no existen suficientes antecedentes que permitan continuar la investigación [13].

La aplicación temprana del término 'archivo provisional' ha sido objeto de análisis y cuestionamiento en diversas ocasiones. De cara a las víctimas y a la sociedad, la aplicación de este término, genera la sensación de impunidad y por ende disminuye la confianza en el sistema, aumentando la sensación de inseguridad. Al ahondar en las causas de la aplicación de este término, Zegarra sostiene que *'en la práctica se reciben una gran cantidad de denuncias respecto de las cuales se aportan muy pocos antecedentes, resultando difícil obtener resultados. Es necesario que los fiscales cuenten con la posibilidad de discriminar entre el gran número de casos que llegan a conocimiento del Ministerio Público de aquellos que ofrezcan perspectivas favorables para desarrollar una investigación productiva (...)'* [14]. La autora concluye que la aplicación de este término se debe en gran parte a los términos aplicados en causas que no cuentan con imputados conocidos.

Una consecuencia a la aplicación del término de archivo provisional, es la pérdida de información de las causas que terminan de esta forma. En muchas ocasiones, este tipo de causas presentan antecedentes parciales del delito pero no los suficientes para poder continuar con la investigación, sin embargo, si se investigaran estas causas con otras de forma conjunta, podrían existir cruces de información que incrementen la probabilidad de éxito en la investigación. Esta es la hipótesis que sostiene SACFI en su metodología de trabajo, ya que a través de la agrupación de causas se espera tener mejores resultados en la investigación y posterior persecución penal.

1.4.3. Soluciones posibles

En el marco del Análisis Criminal, la Ciencia de Datos emerge como una disciplina clave para generar valor a partir de los datos, y apoyar en la toma de decisiones. Esta disciplina incorpora la utilización de técnicas de recopilación, procesado, modelamiento y visualización de datos.

En este ámbito, varios actores de la academia han propuesto distintos enfoques que permitirían mejorar las técnicas del análisis criminal. Por ejemplo, algunos han enfocado sus investigaciones en la elaboración de índices de reincidencia para un determinado delito [15]. Otros han centrado sus esfuerzos en el procesamiento de datos que permiten predecir hotspots o zonas calientes de criminalidad [16]. También se han propuesto métodos para formar *Clusters* de delitos, y ver su proyección en el tiempo [17]. Otros autores han hecho propuestas mediante el Análisis de Redes Sociales (SNA bajo la sigla en inglés derivada de Social Network Analysis), ya que este enfoque permite visualizar la relación entre individuos que comparten algún atributo en común.

El uso de una u otra herramienta depende en gran parte del objetivo, la calidad y cantidad de datos con los que se cuente. En el caso del Ministerio Público, en particular de SACFI, el estudio de evaluación realizado por el centro de Microdatos y detecta que las bases de datos con las que cuenta el Ministerio Público no poseen la estructura necesaria para realizar minería de datos. [9]

Ante estas limitaciones, es necesario un sistema que permita estandarizar el análisis de información mediante la integración de diferentes fuentes de datos. Para el presente trabajo, se destacan las ventajas del enfoque de análisis de redes sociales por sobre otras soluciones, ya que este permite representar visualmente las diferentes vinculaciones que podrían existir entre entidades distintas, facilitando la comprensión de un fenómeno complejo como los que se observan en el ámbito del análisis criminal. A partir de esta representación visual mejora también la interpretabilidad de los resultados que pueda entregar el modelo.

1.4.4. Presentación del proyecto

El proyecto que se propone actualmente, consiste en elaborar un modelo que relacione las causas delictivas, en vez de los individuos como lo hace el enfoque tradicional de 'Fiscal Heredia' mediante los modelos StRAM y LiRAM. Esto permitiría la agrupación y selección de causas similares que posiblemente fueron cometidas por los mismos individuos. Mediante este enfoque, se busca seleccionar causas que sean relevantes para aumentar las probabilidades de éxito en una investigación que, a priori, no cuenta con antecedentes suficientes, y que potencialmente terminarían en el archivo provisional del Ministerio Público. Para la construcción de este modelo, se utilizará Análisis de Redes Sociales y técnicas de programación lineal que permitan optimizar una función objetivo capaz de identificar conjuntos de causas en las que hayan podido participar los mismos individuos y que contengan información suficiente para llevar a cabo la persecución penal.

1.5. Hipótesis de investigación

"Se puede establecer una correlación entre causas delictivas similares y la participación de individuos en estas actividades delictivas."

1.6. Objetivos

1.6.1. Objetivo general

Mejorar la identificación de individuos involucrados en delitos a partir del cruce de información entre causas similares, utilizando técnicas de Análisis de Redes Sociales (SNA) y programación lineal.

1.6.2. Objetivos específicos

1. **Analizar patrones delictivos:** Identificar patrones de comportamiento y relaciones entre distintas causas.
2. **Analizar Redes Sociales:** Establecer vínculos entre causas delictivas en base a una métrica de similitud.
3. **Formular un modelo de Programación Lineal:** Optimizar la agrupación de causas mediante un modelo matemático de programación lineal.
4. **Aplicar y evaluar el modelo:** Aplicar el modelo propuesto en un conjunto de datos reales y evaluar su eficacia, comparando los resultados obtenidos con los métodos tradicionales de investigación.

1.7. Métricas y evaluación de resultados

Para la evaluación de los resultados, se espera abordar distintas dimensiones que se enumeran a continuación:

1. **Comparación de causas incorporadas a un foco delictivo manualmente vs. incorporación de causas en base al modelo de redes sociales propuesto**: Actualmente, los equipos de analistas realizan consultas a las bases de datos para filtrar los relatos según palabras que se consideren claves. Por ejemplo, en un foco para investigar robos que se realicen en paraderos deben probar distintas palabras, tales como: 'paradero','robo','celular'. Sin embargo, esta búsqueda es limitada en sus resultados ya que muchos partes policiales presentan errores de tipeo, y por lo tanto hay un conjunto de causas que nunca se encontrarán de esta manera. Sin embargo, con la propuesta del modelo de red social, se logra capturar la similitud entre los modus operandi, rompiendo la barrera de coincidencia de palabras ya que se tiene un nuevo enfoque de búsqueda o similitud según el contexto del relato.
2. **Tiempo de análisis de causas de un foco:** Cuando los equipos SACFI comienzan a trabajar en un foco declarado, deben elaborar el documento de 'Plan de persecución

penal'. Para esto, luego de tener un universo de causas pertinentes, escogen un grupo acotado de causas que puedan servir para la investigación. Por tanto, se evaluará el tiempo que toma actualmente el enfoque manual versus el tiempo que toma la consulta a través del modelo propuesto.

3. **Identificación de delitos cometidos por los mismos sujetos:** Se medirá si el modelo propuesto es capaz de identificar causas similares en las que hayan participado al menos uno de los imputados.

1.8. Alcances

Dentro de los alcances del proyecto, se considera el planteamiento y desarrollo de un modelo que sea capaz de analizar un conjunto de datos de prueba, en el contexto de la familia de delitos de robos y hurtos sucedidos en un espacio geográfico y temporal que permitan la evaluación mediante la comparación de los resultados obtenidos manualmente por la unidad SACFI regional respectiva.

La implementación del modelo no se aborda en este trabajo, dado que implicaría un tiempo de desarrollo mayor al que se contempla en esta investigación. Además, para la implementación del modelo se trabajará con una medida de similitud entre causas delictivas que ya ha sido establecida mediante técnicas de procesamiento de texto con LLMs (*Large Language Models*).

Capítulo 2

Marco Conceptual

En esta sección se presentan los principales fundamentos que sustentan este trabajo. Primero se presentan conceptos relevantes en el ámbito jurídico y en el ámbito del análisis criminal. Luego, se levantan otras dos subsecciones respecto a Análisis de Redes Sociales (SNA), Problemas de Programación Lineal (PPL), algunos métodos que existen para resolver estos problemas, y se mencionan algunas técnicas de procesamiento de lenguaje natural.

2.1. Marco Jurídico

En esta subsección se consideran las definiciones propuestas por el Ministerio Público en su sitio web, artículos del Código Procesal Penal (CPP) y otras definiciones levantadas desde la documentación interna de SACFI.

- **Delito:** Es delito toda conducta descrita por la ley, que lleva aparejada una sanción penal en caso de contravención o incumplimiento. En el artículo primero del Código Penal se define como 'acción u omisión voluntaria penada por la ley'.
- **Imputado:** Según el artículo 7 del Código Procesal Penal, un imputado es la 'persona a quien se le atribuye participación culpable en un hecho punible'.
- **Investigación:** Conjunto de actuaciones encaminadas a la comprobación de un hecho que reviste caracteres de delito y la participación que puede corresponderle a una o más personas como autores, cómplices o encubridores. Por mandato constitucional le corresponde a la Fiscalía, la que actúa auxiliada por las policías y otros organismos que colaboran en las labores investigativas.
- **Foco Investigativo:** Conjunto de delitos de igual o distinta naturaleza que forman parte de un problema delictual donde se podrían identificar una o más estructuras de criminalidad reconocible a través del análisis criminal y declarado por el respectivo fiscal regional.
- **Estructura Criminal Reconocible:** disposición o modo en que interactúan los distintos hechos delictuales que conforman el fenómeno criminal o problema delictivo. Pueden responder a distintos modelos: mercado delictual, agrupación delictual, patrones o modus operandi.

- **Imputado prolífico:** Imputado que presenta múltiples cargos o acusaciones en relación con diversos delitos. Corresponde a alguien que ha sido acusado en numerosas ocasiones y puede tener un historial extenso de involucramiento en actividades ilegales.

2.2. Análisis Criminal

El análisis criminal, desde la perspectiva de SACFI, puede entenderse como el conjunto de procesos orientados a entregar información oportuna y pertinente en relación a patrones y tendencias de hechos delictuales [8]. Este proceso implica la recopilación, organización y evaluación de datos relacionados con la actividad criminal para extraer patrones significativos.

Según la Asociación Internacional de Analistas Delictuales (IACA), un patrón delictual puede definirse como un grupo de dos o más delitos o incidentes reportados o descubiertos que es único, porque cumple cada una de las siguientes condiciones:

- Comparten al menos una coincidencia en el tipo de delito o comportamiento de los delincuentes o las víctimas; características del agresor, víctimas o blancos; bienes o especies afectadas (propiedad) o el lugar de ocurrencia.
- No existe relación conocida entre víctimas e infractores (es decir, es(son) delito(s) entre desconocidos).
- Los elementos comunes hacen del conjunto de delitos una configuración identificable y distinta de otras actividades delictuales que se producen en el mismo período.
- La actividad delictiva es generalmente de duración limitada, que puede ir desde semanas a meses.
- El conjunto de delitos relacionados es tratado como una unidad de análisis y abordado a través de tácticas y de la acción policial focalizada.

En la literatura se destacan 7 patrones identificables [18] con el análisis de información:

- **Hot Spot:** Zona de alto riesgo.
- **Hot Place:** Lugar de alto riesgo.
- **Hot Setting:** Entorno de alto riesgo.
- **Crime Series:** Delitos similares que se cree que son cometidos por el mismo sujeto o grupo de sujetos que actúan concertadamente o en conjunto.
- **Crime Spree:** Delitos ocurridos con una alta frecuencia en un período de tiempo acotado.
- **Hot Victim o Víctima Preferida:** Grupo de delitos cometidos sobre víctimas que comparten características similares y/o que tienen un comportamiento parecido.
- **Hot Product:** Delitos donde se tiene por objetivo robar un solo tipo de objeto.

Estos patrones entregan información valiosa sobre las tres entidades que componen oportunidades delictivas: Lugar, víctima u objeto deseado y delincuente.

2.3. Redes Sociales

Según Wasserman [19], una red social es la 'estructura relacional de un grupo o sistema social más amplio que consiste en el patrón de las relaciones entre la colección de actores'. Una red social está compuesta de diversos elementos, entre los cuales se destacan:

- **Nodos:** Los nodos representan a los actores individuales en la red, como personas, organizaciones o entidades.
- **Aristas:** Las aristas, también conocidas como enlaces o conexiones, representan las relaciones o interacciones entre los nodos en la red. Estas relaciones pueden ser de diversos tipos, como amistad, colaboración o intercambio de información.
- **Camino:** Un camino es una secuencia de nodos y aristas que conecta dos nodos específicos en la red. Puede haber múltiples caminos posibles entre dos nodos.
- **Peso:** El peso se refiere a una medida numérica asociada a una arista, que indica la fuerza o importancia de la relación entre los nodos conectados. El peso puede representar la frecuencia de interacciones, la confianza o cualquier otra métrica relevante en el contexto de la red.
- **Subgrafo:** Un subgrafo es un subconjunto de nodos y aristas extraídos de la red completa. Los subgrafos pueden utilizarse para analizar grupos o comunidades específicas dentro de la red o para estudiar patrones particulares de relaciones.

En esta área, es relevante explorar y definir varios conceptos clave para comprender el análisis de redes sociales. Entre ellos, es fundamental definir los siguientes enfoques que se utilizan en este campo:

- **Cohesión:** Este enfoque se centra en medir el grado de conexión y cohesión dentro de una red social. La cohesión refleja la tendencia de los nodos a formar grupos o comunidades altamente interconectadas. Los métodos de análisis de cohesión permiten identificar subgrupos de nodos que comparten relaciones fuertes y pueden revelar la existencia de comunidades dentro de la red.
- **Intermediación:** La intermediación se refiere a la capacidad de ciertos nodos de actuar como intermediarios o puentes en la red. Estos nodos tienen un papel crucial en la comunicación y transferencia de información entre otros nodos que de otro modo estarían desconectados. El análisis de intermediación ayuda a identificar los nodos clave que facilitan el flujo de información y control en la red.
- **Ranking:** El enfoque de ranking se enfoca en clasificar o jerarquizar los nodos de la red según alguna medida de importancia o relevancia. Esto puede basarse en diferentes criterios, como el grado de conexión, la centralidad o la influencia de los nodos. El análisis de ranking ayuda a identificar los actores más influyentes o importantes dentro de la red, brindando una comprensión más profunda de la estructura y la dinámica de la red.
- **Rol:** El enfoque de rol se centra en la posición o función que desempeñan los nodos en la red. Cada nodo puede tener un rol distinto según su comportamiento, sus características

o su posición en la red. Identificar los roles de los nodos puede ayudar a comprender las dinámicas sociales y los patrones de interacción en la red, como líderes, seguidores, intermediarios o enlaces débiles.

En la etapa de análisis de una red social, es importante considerar diferentes métricas que proporcionan información significativa sobre la estructura y el funcionamiento de la red. Algunas de estas métricas relevantes incluyen:

- **Grado de un nodo:** El grado de un nodo se refiere al número de conexiones que tiene con otros nodos en la red. Representa la medida de conectividad de ese nodo en particular. Los nodos con un alto grado suelen ser considerados como actores clave o centrales en la red, ya que tienen más conexiones y, por lo tanto, mayor potencial de influencia o acceso a información.

- **Densidad:** La densidad de una red es una medida que indica qué tan conectados están los nodos en la red en comparación con el número total de conexiones posibles, por lo que toma valores entre 0 y 1. Se calcula dividiendo el número de conexiones reales entre los nodos por el número total de conexiones posibles. Una densidad alta indica una red densamente interconectada, mientras que una densidad baja indica una red más dispersa o fragmentada.

- **Centralidad:** La centralidad es una medida que evalúa la importancia relativa de un nodo en la red. Hay diferentes tipos de centralidad, como la centralidad de grado, la centralidad de intermediación y la centralidad de cercanía. La centralidad de grado se basa en el número de conexiones que tiene un nodo y su posición de influencia en la red. La centralidad de intermediación mide la participación de un nodo en los caminos más cortos entre otros nodos. La centralidad de cercanía mide qué tan cerca está un nodo de otros nodos en términos de la distancia promedio.

- **Centralidad media:** La centralidad media es una medida que proporciona una visión general de la importancia promedio de los nodos en la red. Se calcula tomando la media de las medidas de centralidad de todos los nodos en la red. La centralidad media ayuda a comprender la importancia global de los nodos en relación con la red completa, y puede ser útil para identificar nodos influyentes o determinar la estructura general de la red.

- **Modularidad:** La modularidad es una medida que permite cuantificar la estructura de modularidad en una red y se utiliza para identificar comunidades o grupos dentro de una red. La medida de modularidad de una partición es un valor que va entre -1 y 1 y que mide la densidad de los vínculos dentro de una comunidad comparados a los vínculos entre comunidades.

2.4. Programación lineal

Otro campo relevante en el desarrollo de este trabajo es el de la optimización, en particular, la Programación Lineal. La Programación Lineal es una técnica matemática que permite resolver problemas de asignación eficiente de recursos considerando restricciones de recursos limitados. En esta técnica se minimiza o maximiza una función lineal con múltiples restricciones lineales. Esta disciplina se utiliza en numerosas organizaciones para resolver problemas de distinta índole, tales como en la gestión de inventarios y logística, asignación eficiente de

recursos o planificación de rutas de transporte.

En una organización como el Ministerio Público, la técnica de Programación Lineal ofrece numerosas posibilidades de aplicación para aumentar la eficiencia de procesos tales como el análisis de información, ante los cuales existen recursos humanos limitados para procesar grandes cantidades de información. De esta forma, podría ayudar a obtener mejores resultados en menores tiempos.

Dentro de los procesos clave de la programación lineal se encuentra la fase de modelado matemático, donde se pretende representar el problema a resolver mediante un lenguaje matemático. Esta formulación matemática incluye una función a optimizar, variables de decisión y restricciones lineales. Una vez planteado el modelo, se utilizan algoritmos para la resolución de estos problemas ejecutados mediante algún software de optimización. Algunos de los algoritmos utilizados para resolver los problemas de programación lineal son:

- **Simplex Primal:** El método Simplex Primal es un enfoque utilizado para resolver problemas de programación lineal. Se basa en un algoritmo iterativo que busca mejorar continuamente la solución actual moviéndose de una solución factible a otra a lo largo de las restricciones del problema. El método Simplex Primal se enfoca en maximizar o minimizar una función lineal sujeta a un conjunto de restricciones lineales.
- **Simplex Dual:** El método Simplex Dual es una variante del método Simplex que se utiliza para resolver problemas de programación lineal en su forma dual. El enfoque dual se basa en una formulación alternativa del problema original, donde las variables y las restricciones se intercambian. El método Simplex Dual busca mejorar la solución a través de iteraciones y optimizar una función dual asociada al problema original.
- **Método de barrera:** Método de barrera: El método de barrera es otra técnica común utilizada en la resolución de problemas de programación lineal. El método de barrera se basa en una aproximación de la solución óptima al problema mediante la inclusión de barreras en las restricciones del problema. A medida que las barreras se hacen más grandes, la solución se acerca a la solución óptima.

2.5. Procesamiento de lenguaje natural

En el ámbito de las herramientas para el procesamiento del lenguaje humano, se encuentran diversas técnicas que abarcan desde enfoques más básicos, como la tokenización o lematización, hasta el empleo de modelos más avanzados conocidos como *Large Language Models* (LLMs) o Grandes Modelos de Lenguaje.

La arquitectura de *transformers* es un tipo de modelo de gran capacidad en el procesamiento del lenguaje natural. Los transformers son modelos de aprendizaje automático que han demostrado ser altamente eficientes en tareas relacionadas con el lenguaje, gracias a su capacidad para capturar patrones y relaciones complejas en los datos.

Otro concepto que emerge en esta área es el de *token*. En el contexto del procesamiento del lenguaje natural, un *token* se refiere a una unidad indivisible de texto, que puede ser una palabra, subpalabra o incluso un carácter, dependiendo de la granularidad. La tokenización

es el proceso de dividir un texto en tokens, lo cual es esencial para que los modelos de lenguaje comprendan y procesen la información de manera adecuada.

Entre los LLMs, destacan modelos como gpt-4. Este modelo se basa en la arquitectura de transformers y se destaca por su capacidad para predecir el próximo *token* en un documento. Su entrenamiento se realiza mediante la utilización de datos públicos disponibles en internet, así como datos proporcionados por terceros. Gracias al extenso conjunto de datos de entrenamiento, gpt-4 logra resultados notables al generar respuestas coherentes en una variedad de temas planteados por el usuario.

Otro modelo relevante es ada-embedding-002. Este modelo permite generar *embeddings* a partir de un texto. Los *embeddings* son representaciones vectoriales que capturan la información semántica de un texto. La utilidad de estos *embeddings* radica en la capacidad para comparar el sentido de dos párrafos y evaluar su similitud en relación al sentido y al contenido del texto.

Capítulo 3

Metodología

Los modelos de redes sociales que se han desarrollado para el Ministerio Público buscan focalizar la investigación de los y las Fiscales en delitos contra la propiedad privada, donde no se tiene la totalidad de los imputados que participaron del hecho delictual. Para aplicar estos modelos es necesario tener al menos un imputado conocido, con el fin de obtener la agrupación criminal más probable que haya participado en el delito.

Tal como se mencionó en el Capítulo 1, al menos un 90 % de los delitos de robo, no cuentan con imputados conocidos al momento de ingresar al Ministerio Público. Esto motiva la construcción de un nuevo modelo de redes sociales que permita enriquecer la investigación, con un enfoque alternativo: estudiar la relación entre causas.

Si bien existen metodologías enfocadas en el descubrimiento de patrones en distintos conjuntos de datos tales como KDD o CRISP-DM, en este trabajo se propone una metodología más específica dadas las disciplinas que se escogen para abordar el problema, tal como el Análisis de Redes Sociales (SNA) y Modelos de Optimización Lineal. Además, es importante destacar que la mayor parte del trabajo se realiza en base a relatos de hechos delictuales, datos no estructurados que para su procesamiento necesitan un tratamiento distinto a las etapas que proponen las metodologías anteriormente mencionadas. Este enfoque y metodología ofrece ventajas tanto en la visualización como en la interpretabilidad de los resultados, tal como se evidenciará en el Capítulo 5. Las principales etapas de la metodología propuesta se resumen en el esquema de la Figura 3.1.

Figura 3.1: Esquema que resume las principales etapas de la metodología propuesta. [Fuente: Elaboración propia]

Las actividades específicas que contempla cada etapa se detallan a continuación:

Entendimiento del problema

1. Realizar un estado del arte de los modelos de redes sociales aplicados actualmente en el Ministerio Público.
2. Elaborar un listado de variables y características relevantes para la identificación de patrones criminales, en base a la revisión de literatura, así como entrevistas con profesionales del Ministerio Público.

Recopilación de los datos

3. Procesar los datos necesarios para construir las variables levantadas en el entendimiento del problema.

Análisis exploratorio de los datos

4. Realizar un análisis descriptivo de los datos.

Construcción de la red social

6. Construir una red social que relacione causas delictivas.

Planteamiento del modelo de optimización

7. Plantear un modelo de optimización lineal que permita proponer sujetos que podrían haber participado en un delito con Imputado Desconocido (ID).

Validación del modelo

8. Interpretar los resultados.
9. Evaluar la efectividad y las posibles áreas de aplicación del modelo propuesto.
10. Realizar una comparación y análisis con otros enfoques existentes en la detección y resolución de causas delictivas, evaluando ventajas y limitaciones.

Capítulo 4

Desarrollo metodológico

4.1. Entendimiento del problema

La persecución penal se refiere al conjunto de actividades que se llevan a cabo para investigar, recopilar pruebas, procesar y enjuiciar a quienes hayan cometido un delito. En Chile, esta actividad se asigna de forma exclusiva al Ministerio Público.

Dentro del Ministerio Público, se conforma el Sistema de Análisis Criminal y Focos Investigativos que busca contribuir en la persecución penal, proponiendo distintas estrategias que permitan abordar fenómenos complejos, destacándose por salir de la lógica de investigación del caso a caso. Un estudio de evaluación realizado por la Universidad de Chile [9], concluye que la lógica de trabajo que propone SACFI, ha permitido obtener mejores resultados en las actividades de la persecución penal, logrando mejores penas asignadas y también identificando más imputados para causas que se investigan como 'foco investigativo' versus aquellas que se analizan de forma separada.

Entre las tecnologías que utiliza SACFI, destacan herramientas que, en base a grafos o redes sociales y modelos de optimización, logran proponer la agrupación criminal más probable a partir de un imputado conocido. Estas ayudan a focalizar la investigación en casos donde se desconoce el total de imputados que cometieron un delito. Sin embargo, para delitos de robo, cerca de un 90 % de los casos no cuentan con un Imputado Conocido (IC) y, por lo tanto, no se puede recurrir a la consulta con estos modelos.

De esta forma, para delitos con Imputado Desconocido (ID), las estrategias de investigación son más bien manuales. Para esto, los analistas criminales deben realizar primeramente un 'Diagnóstico de Criminalidad Regional (DCR)', para luego proponer focos investigativos donde se identifican causas que debiesen incorporarse, considerando aquellas que den cuenta de un problema relevante y que revistan información de interés para la investigación.

El enfoque de trabajo manual para analizar grandes cantidades de información presenta desventajas en la identificación de patrones complejos, en la eficiencia de los procesos, y también en los sesgos personales que cada equipo pueda tener. Es por esto que se busca proponer un nuevo modelo que colabore en el trabajo de análisis e investigación de los equipos SACFI, identificación de estructuras de criminalidad reconocible, en particular, en base a similitud de modus operandi y sujetos o agrupaciones que perpetran el delito.

4.2. Recopilación de datos

Para realizar este trabajo se construyó un set de datos que permitiera la validación del modelo mediante la comparación de resultados obtenidos por un equipo SACFI de forma manual. Para esto se observaron los focos investigativos declarados dentro de los últimos cuatro años que se encontraran terminados a la fecha.

Para el tratamiento de estos datos se utiliza una medida de similitud entre delitos desarrollada por Pablo Pincheira, tesista del Magíster de Ciencia de Datos de la Universidad de Chile, cuyo trabajo, a la fecha de escritura de este informe, se encuentra en curso de publicación [20]. Esta medida de similitud presenta grandes ventajas ya que logra capturar distintos patrones entre delitos que guardan relación con el victimario, la víctima y el contexto del delito.

Sin embargo, tal como se discute más adelante en la sección 4.4, se evidencian algunas limitaciones ya que esta medida compara coincidencia exacta de lugares geográficos. En otras palabras, dos delitos cometidos bajo las mismas circunstancias de víctima, victimario y contexto serán distintos en igual medida si es que estos se cometen en regiones muy lejanas entre sí, o bien si estos suceden en comunas aledañas pero no coincidentes. Debido a esta limitación, se decide trabajar con delitos de tres meses en un espacio geográfico acotado a una sola comuna. De tal modo que en ese espacio existen causas que fueron consideradas en al menos dos focos investigativos, y permiten la validación o refutación del modelo, y además se evita la limitación identificada en el cálculo de la similitud geográfica.

Para la recopilación de datos se realizó un proceso de extracción y anonimización a partir de las bases de datos con las que cuenta el Ministerio Público. Posteriormente, se aplicó un procesamiento con modelos de lenguaje para estructurar datos a partir de relatos del hecho delictual. El conjunto de datos a estudiar contempla 637 causas delictuales, donde cada una contiene un solo delito[6] y una sola víctima[7]. Las causas cuentan con los siguientes datos:

- **RUC:** Rol Único de Causa, identificador único de la causa.
- **Fecha del hecho:** Fecha en que ocurre el hecho delictual.
- **Fecha de recepción:** Fecha en que el Ministerio Público recibe la denuncia.
- **Código estado del caso:** Código que da cuenta del estado de la causa: Ingresado, vigente, suspendido, terminado, no vigente o transferido.
- **DNI imputado:** Número de identificación del individuo al que se le imputa el delito.
- **IMPNN:** Variable que da cuenta si el imputado es conocido o no.
- **Edad delito:** Edad que tenía el imputado al momento de cometer el delito.
- **Código familia:** Código de la familia a la cual pertenece el delito. En términos generales, existen distintas categorías que agrupan delitos del mismo tipo, por ejemplo,

[6] La estructura de datos de las bases de datos del Ministerio Público permite que en una causa puedan existir uno o más delitos, según el criterio de agrupación de delitos que considere cada Fiscal.

[7] Este filtro se aplica para poder aplicar la medida de similitud señalada, ya que se trabajó en delitos con una sola víctima.

dentro de la familia 'robos' existen robos con intimidación, robos por sorpresa, robos a la propiedad privada, entre otros.

- **Código delito:** Código del delito cometido.
- **Comuna del delito:** Nombre de la comuna donde se cometió el delito.
- **DNI víctima:** Documento de identidad de la víctima del delito, si corresponde. En ocasiones hay delitos que no tienen una víctima directa o bien no se logra identificar la víctima.
- **Relato:** Relato del parte policial o del organismo que recibe la denuncia. El relato podría haberse elaborado por policías o bien haberse recibido directamente en la Fiscalía, por parte de la víctima.

La parametrización de variables se realiza mediante el uso de LLMs (*Large Language Models*). En particular, el modelo gpt-4 desarrollado por la empresa OpenAI permite, entre otras cosas, estructurar datos presentes en texto libre según la instrucción o *prompt*[8] que se le indique al modelo. Primero se selecciona un subconjunto de datos para testear un *prompt* efectivo que permita la obtención de datos con la estructura deseada. Una vez diseñado el *prompt*, se aplica sobre todos los relatos anonimizados, indicando además que si la información no está presente en el relato, el campo quede vacío. De esta forma, se obtienen las variables:

- **Modus operandi:** Se refiere a la manera en que se lleva a cabo el delito, mencionando métodos o técnicas utilizadas.
- **Transporte de llegada de delincuente(s):** Forma en que se acerca el delincuente a cometer el delito.
- **Transporte de huida de delincuente(s):** Forma en que el delincuente huye del lugar.
- **Tipo de arma:** Corresponde al tipo de arma utilizada para cometer el delito, ya sea para amenazar o atacar a la víctima. Puede ser un arma de fuego, arma blanca u objeto contundente.
- **Violencia:** Determina si hubo una agresión física a la víctima.
- **Lesión en la víctima:** Determina si la víctima resultó lesionada por el delincuente.
- **Características físicas de delincuente(s):** Descripción de vestimentas, y apariencia física de delincuente(s).
- **Evidencia del hecho:** Testigos, cámaras, rastreo GPS, pericias policiales, u otros elementos que permitan confirmar el hecho.

[8] Prompt: Instrucciones utilizadas para ejecutar Modelos de Lenguaje (LLMs).

4.3. Análisis Exploratorio de los Datos

Esta sección presenta un análisis de las primeras variables descritas, que se obtienen directamente de las bases de datos del Ministerio Público. Se estudian las variables más relevantes identificadas en el Capítulo 1, y otras que puedan ser de interés para las siguientes secciones.

Imputados desconocidos

Para el set de 637 delitos analizados, se observa en la figura 4.1 la cantidad de causas en las que existe un imputado conocido. Tal como se observa en las estadísticas mencionadas en el Capítulo 1, más de un 90 % de los delitos no cuentan con imputado conocido.

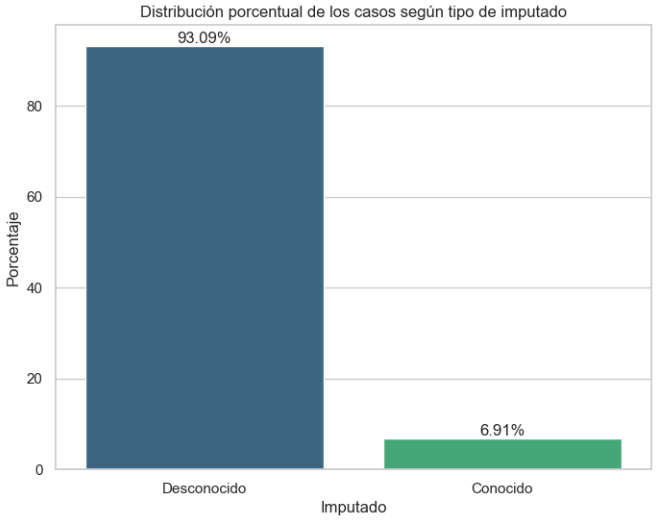


Figura 4.1: Distribución porcentual de casos según tipo de imputado. [Fuente: Elaboración propia]

Espacio temporal

La gran parte de los delitos se concentran en tres meses del año 2019, aunque existen otros 8 delitos adicionales que no se encuentran en este intervalo. Esto se debe a que en el primer segmento de tiempo mencionado (delitos 2019), no había suficientes causas con repetición de imputados que permitieran la validación posterior de los experimentos. Por esta razón, se identificaron los imputados conocidos de esas causas, para luego buscar otros delitos que hayan cometido con posterioridad y se integraron estas causas al set de datos. La distribución de las fechas de los delitos sucedidos en 2019 se evidencia en la figura 4.2.

Si bien la fecha en la que ocurrieron los delitos coinciden con el estallido social vivido en octubre de 2019, las causas se han filtrado con el fin de que delitos asociados a ese fenómeno no alteren el set de datos. Esto se logra aplicando un filtro según la cantidad de víctimas afectadas, ya que en muchas ocasiones cuando se vulneraron tiendas o empresas en el contexto del estallido social, aparecía más de una víctima, y tales casos fueron excluidos.

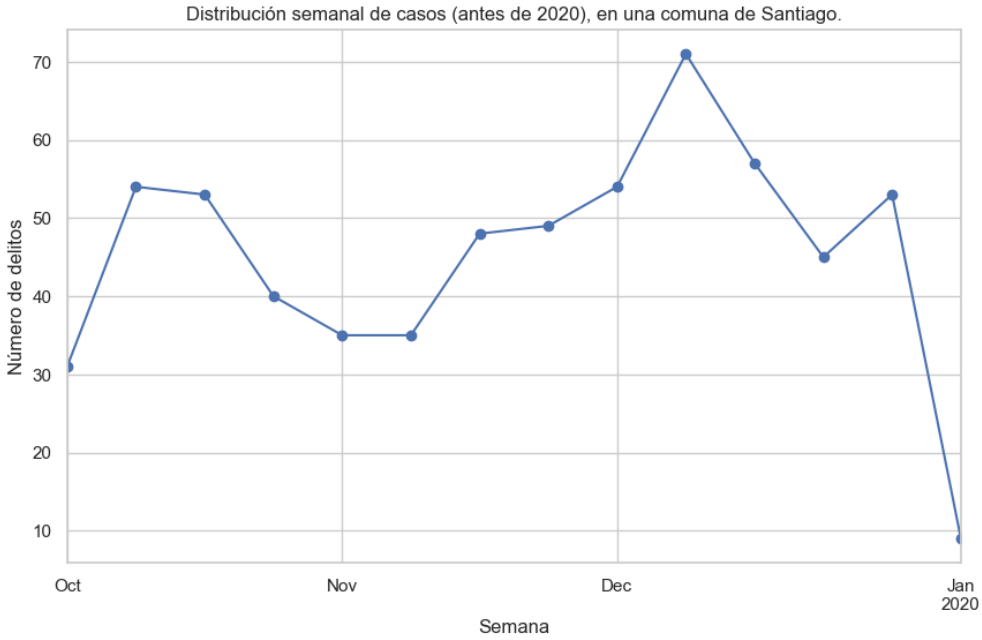


Figura 4.2: Distribución semanal de casos (antes de 2020). [Fuente: Elaboración propia]

Número de imputados

Por otra parte, el análisis de imputados por delito se muestra en la imagen 4.3. Se puede apreciar que cerca de un 97 % de los delitos tienen solo un imputado asociado. Sin embargo, estos datos son extraídos directamente desde la base de datos del Ministerio Público, por lo que los datos podrían tener errores de registro, y ser menos fiables que la cifra que se extrae a partir del análisis del relato del hecho.

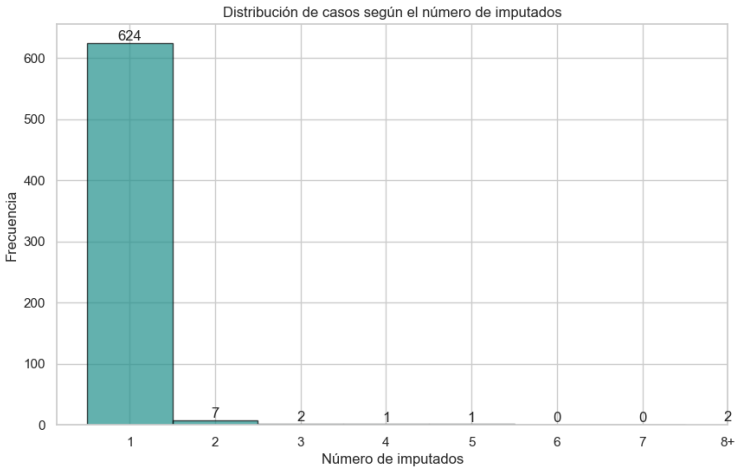
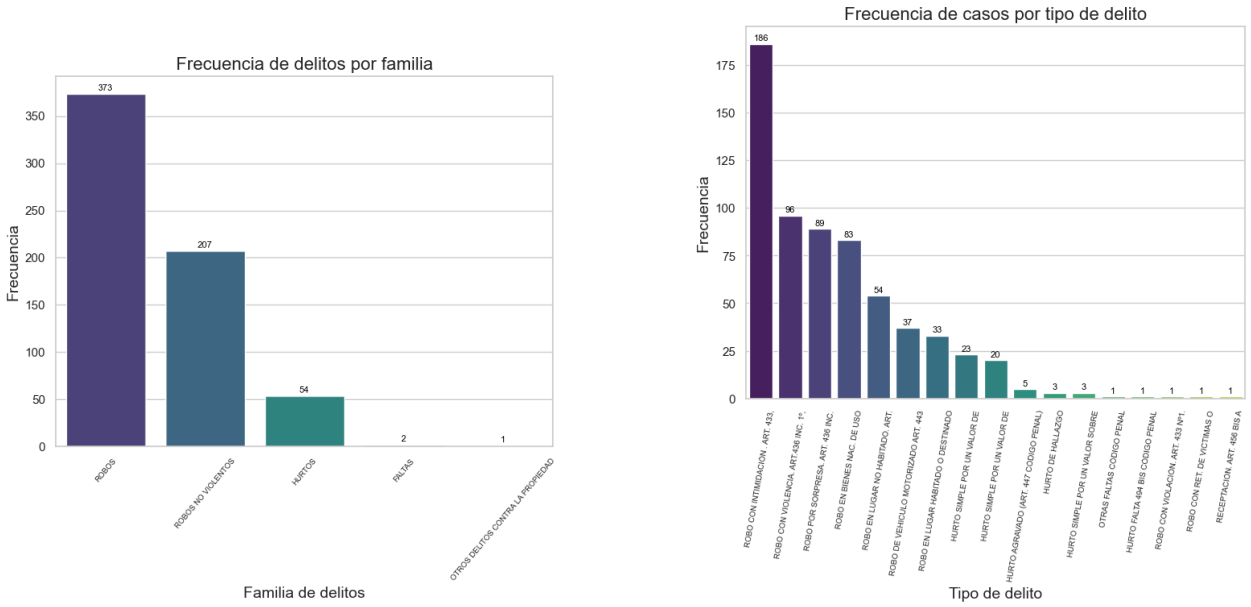


Figura 4.3: Distribución de casos según el número de imputados. [Fuente: Elaboración propia]

Código del delito

El análisis de las variables de 'familia' y 'código del delito' se observa en los gráficos de la figura 4.4. Tal como se comenta anteriormente, se trabaja con delitos contra la propiedad. De la figura, se observa que la familia de delitos que tiene mayor frecuencia es la de robos, seguida por robos no violentos y hurtos. En cuanto al tipo de delito, destaca el robo con intimidación. Según el set de datos, estas cifras podrían variar pero es relevante el análisis previo para interpretar de mejor manera los resultados del modelo que se propone.



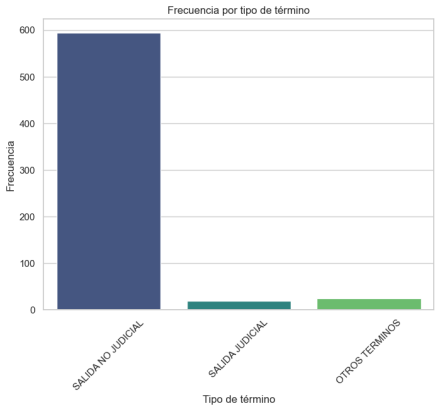
(a) Distribución frecuencia

(b) Distribución porcentual

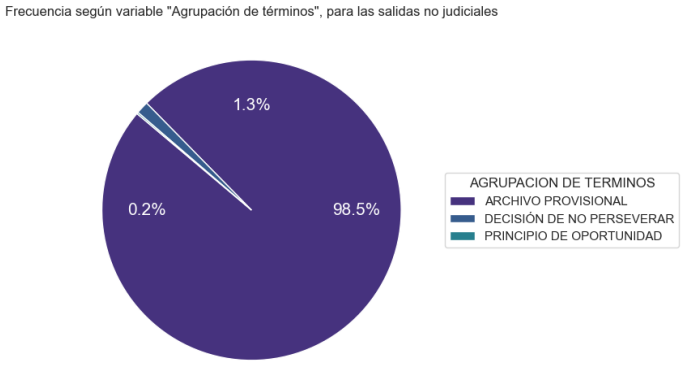
Figura 4.4: Distribución de casos por familia de delito, y por tipo de delito. [Fuente: Elaboración propia]

Término de las causas

Otra situación que se evidenció en el capítulo Capítulo 1, es la gran cantidad de causas que terminan con término 'archivo provisional'. En la figura 4.5, se evidencia que casi 600 causas de 637 terminaron con salida no judicial, es decir que la causa terminó sin la intervención de un tribunal. Al indagar en esta categoría, el 98.5 % de las causas que tuvieron una salida no judicial corresponden a términos de archivo provisional. Si bien esto se ha identificado como un problema por diferentes autores, reviste de igual manera una oportunidad para extraer información relevante que permita hacer el cruce de información con otras causas y así aumentar las probabilidades de éxito en la persecución penal.



(a) Distribución frecuencia



(b) Distribución porcentual

Figura 4.5: Distribución de casos por familia de delito, y por tipo de delito. [Fuente: Elaboración propia]

4.4. Construcción de la red social

La información se modela en una red social donde los nodos corresponden a causas delictivas, que están conectadas por una arista que refleja la medida de similitud entre estas. La presencia de este arco se establece para los valores de similitud mayores a un valor umbral de similitud $\bar{s}_{i,j}$. De esta forma, la red social modelada es no dirigida, debido a la simetría en la relación de causas, y los arcos son ponderados según la magnitud de la similitud.

Dicho de otra forma, la red social se construye con nodos que representan causas o delitos, y existe un vínculo entre estos nodos si la medida de similitud calculada entre ambos es superior o igual a un valor[9]. Dado que la medida de similitud no tiene un sentido de dirección desde un delito hacia otro, se dice que la red social es 'no dirigida'. Luego, cuando existe un vínculo, se diferencian aquellas causas con una medida de similitud más fuerte entre sí, por tanto los arcos son ponderados.

Figura 4.6: Representación de la red social propuesta.
[Fuente: Elaboración propia]

La figura 4.6 representa la propuesta de red social donde cada nodo representa un delito y los arcos se construyen según la medida de similitud que existe entre dos nodos. En esta figura, la ponderación de los arcos guarda relación con la intensidad del color. Aquellos arcos en tonalidades celestes implican un valor de similitud menor que aquellos que están más concentrados.

[9] Este valor se determina en la etapa siguiente.

4.4.1. Nodos

Caracterización de Nodos

Se propone un atributo denominado 'valor investigativo' de cada nodo, que asigna un valor entre 0 a 1, según la cantidad y calidad de información que contenga una causa, en relación a la posibilidad de continuar la investigación.

La motivación de la construcción de esta variable, radica en la metodología de trabajo de los Sistemas de Análisis Criminal y Focos Investigativos. SACFI basa su análisis en fenómenos criminales, definidos como conjuntos de delitos que podrían tener una o más estructuras de criminalidad reconocible. A su vez, una estructura de criminalidad reconocible se entiende como la forma en que podrían interactuar los delitos que se consideran dentro del fenómeno criminal. Finalmente, la relación entre hechos delictuales se da por la interacción entre los elementos que caracterizan cada uno de estos, como por ejemplo el tipo de delito, comportamiento de delincuentes o víctimas, bienes afectados o lugar de ocurrencia. [8] [18]

De esta forma, el 'valor investigativo' de una causa, mide la cantidad y calidad de la información presente en el relato. A medida que este valor sea más alto, la causa permitiría ampliar las posibilidades de investigación de ese delito en particular, así como también de otros delitos que estén vinculados por la medida de similitud. Además de aportar elementos para continuar la investigación, podría ser un buen filtro para identificar causas que puedan tener elementos probatorios en la persecución penal y así obtener mejores resultados.

Las variables descritas a continuación buscan dar cuenta de los elementos que existen en el relato, que permiten conocer el comportamiento del delincuente, el bien sustraído, el lugar, y también el potencial comportamiento que tendrían las víctimas en el transcurso de la investigación.

Descripción de delincuente(s)

Se determinan niveles de 0 a 3 que dan cuenta del nivel de características del delincuente presente en el relato.

- **Nivel 0:** No existe información respecto a delincuente(s).
- **Nivel 1:** Solo existe información parcial respecto a las vestimentas de delincuente(s).
- **Nivel 2:** Existe información respecto a las vestimentas, como también características físicas de delincuente(s), tales como la edad aproximada, presencia de tatuajes, características morfológicas.
- **Nivel 3:** Existe información que permite identificar al individuo de forma particular. Esto se logra cuando las víctimas mencionan dónde viviría el sujeto (al ser vecinos), con la mención de un alias o apodo, o bien mediante la identificación del sujeto (nombre y apellido).

Descripción del transporte de llegada de delincuente(s)

Se determinan niveles de 0 a 3 que dan cuenta del nivel de características del modo de llegada de los delincuentes.

- **Nivel 0:** No existe información respecto a la forma de llegada de delincuente(s).
- **Nivel 1:** Solo existe información parcial respecto al medio de transporte utilizado: a pie, en vehículo, transporte público, en bicicleta, etc.
- **Nivel 2:** Además de indicar el medio de transporte, existen características de este. Por ejemplo: vehículo de color rojo, motocicleta sin placa patente, bicicleta negra.
- **Nivel 3:** Existe información que permite identificar el medio de transporte de forma única. Esto se logra, por ejemplo, cuando el relato contiene información de placas patentes.

Evidencia del hecho

Se determinan niveles de 0 a 3 que dan cuenta de la cantidad de medios que permiten verificar la ocurrencia del hecho, o bien, la posibilidad de rastrear el bien robado.

- **Nivel 0:** No se menciona la existencia de registro del incidente, ni medios que permitan rastrear la especie robada.
- **Nivel 1:** Se menciona la potencial existencia de un medio de verificación, como la presencia de cámaras en el lugar. Sin embargo, no se afirma que efectivamente estén operativas o que sean de fácil acceso para recuperar el material visual.
- **Nivel 2:** Se mencionan al menos dos potenciales medios de verificación distintos, por ejemplo la eventual presencia de cámaras, además de la presencia e identificación de testigos que estuvieron cuando ocurrió el hecho.
- **Nivel 3:** Se cuenta con algún medio de verificación de registro del hecho, mediante la declaración de uno o más testigos, mediante la entrega del medio audiovisual, o mediante el rastreo de la especie robada con el sistema de GPS.

Fecha del hecho

Se propone esta variable bajo la hipótesis de que quienes hayan sido víctimas de un delito de forma más reciente, tienen más disposición a colaborar en la investigación. Además, mientras más reciente haya sido el delito, hay más probabilidad de contar con información relevante como el rastreo actualizado de las especies.

Por ejemplo, si alguien es víctima de un robo hace un año, posiblemente tenía la posibilidad de rastrear la especie durante un par de semanas, pero luego de que la especie se introduce en un mercado, la posibilidad de tener aún este registro disminuye de manera drástica. Lo mismo ocurre con la posibilidad de acceder a cámaras del lugar que hayan registrado del hecho, dada la limitación de almacenamiento de estos sistemas.

Para asignar un valor, se divide el espacio temporal en 4 segmentos, donde el más reciente toma un nivel 3, y el más lejano toma un nivel 0.

Imputado Conocido

La incorporación de esta variable se hace luego del análisis de estrategias que ocupan los equipos SACFI, quienes, cuando identifican delitos relacionados por el mismo modus operandi, buscan sujetos que operen de forma similar en aquella zona, para luego identificar aquellas causas en las que han participado para contar con mayores antecedentes que puedan aportar a la investigación. De esta forma, la variable toma el valor 1 cuando el imputado es conocido, y 0 en caso contrario.

De forma general, estas variables se construyen en base a la observación de las tablas de información parametrizada que elaboran los analistas criminales de forma manual.

Una de las limitaciones que podría tener este enfoque es la correlación positiva entre la variable 'imputado conocido' y 'descripción de delincuente(s)', dado que posiblemente cuando existe imputado conocido, también existe una identificación de forma única de los delincuente. Sin embargo, se busca asignar un valor mayor a aquellas causas que además de tener imputado conocido, entregan una descripción de las características de los delincuentes, para eventualmente cruzar esta información con otras causas de similares características que hayan ocurrido en un horizonte cercano de tiempo.

Aplicación sobre el conjunto de datos

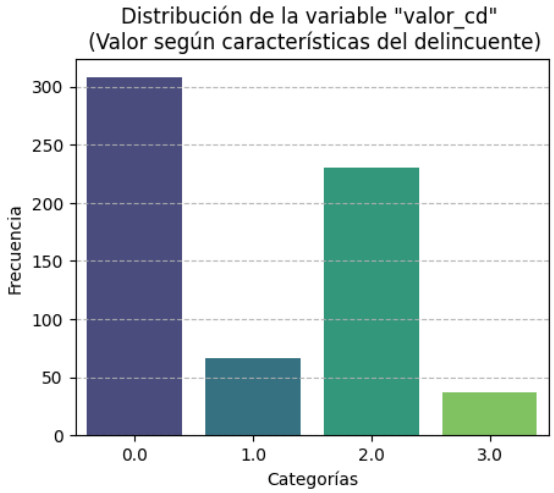
Para la evaluación de la información presente en cada variable, se utiliza el modelo gpt-4, por medio de un *prompt* que describe brevemente el razonamiento para asignar el nivel a cada descripción.

La Figura 4.7 muestra la distribución de las cinco variables mencionadas anteriormente. En cada gráfico se logra ver que en general las víctimas no recuerdan demasiada información respecto al modo de llegada de los delincuentes, probablemente debido al efecto sorpresa que estos buscan. Lo mismo ocurre con la evidencia del hecho, donde en la mayoría de los casos no se tienen medios de verificación, o simplemente se menciona la posibilidad de contar con un registro.

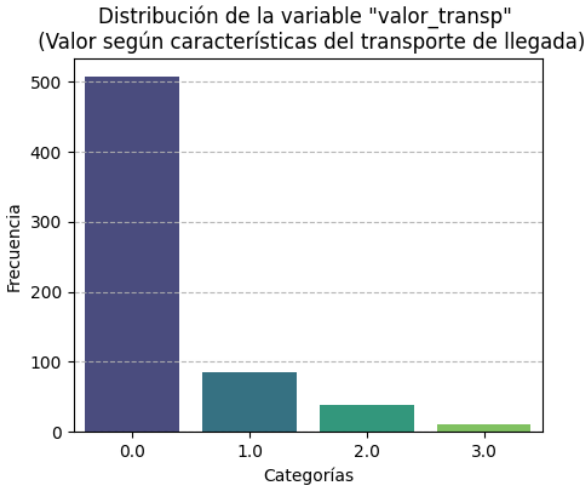
El valor de la descripción de los delincuentes ofrece una posibilidad ya que en al menos la mitad de los casos se cuenta con información que permitiría caracterizar individuos que operen en la zona.

La distribución de la variable imputado conocido es coherente con los antecedentes levantados en el Capítulo 1, dada la alta cantidad de casos donde no se identifican los individuos que cometieron el delito.

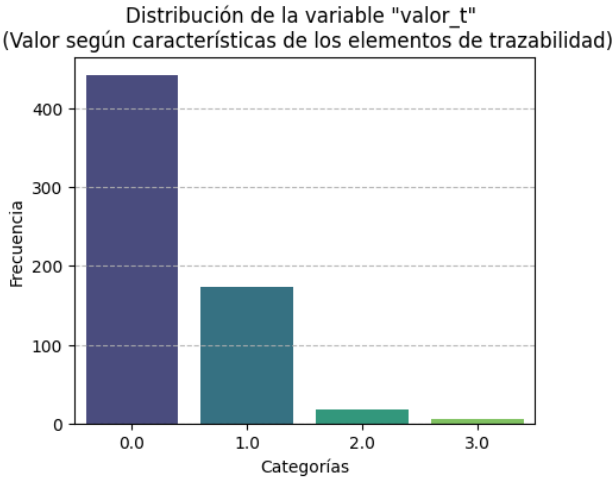
La distribución del valor de la fecha en que ocurre el delito tiene una limitación, dado que por simplicidad, se escogieron 3 meses, y se determinó una división temporal por mes. De tal forma que no existen valores menores a 1. Sin embargo, si se aplicara la metodología sobre un horizonte de tiempo mayor, por ejemplo 1 año, el tamaño de cada segmento tendría otra escala y luego de un cierto período de tiempo, las causas más lejanas tendrían un valor de 0.



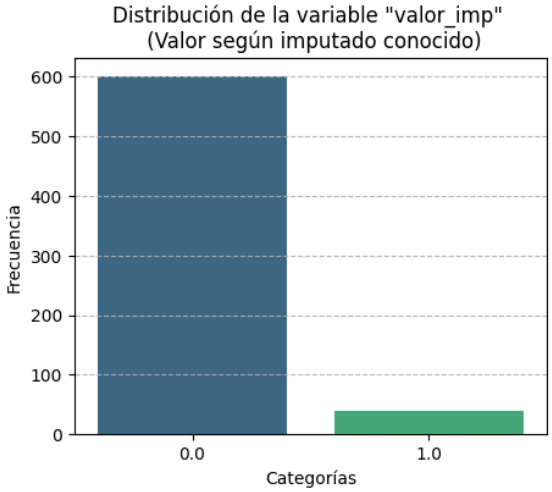
(a) Valor según características del delincuente



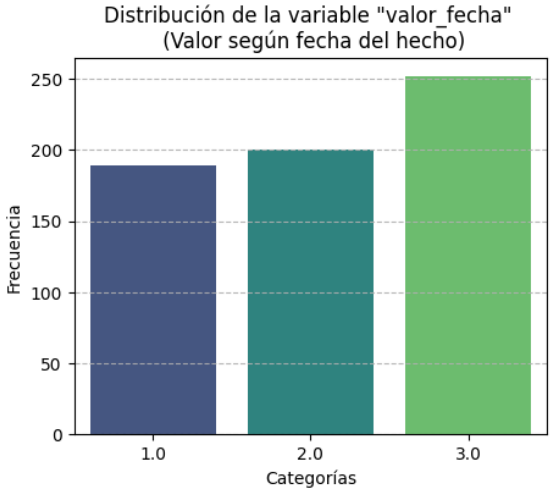
(b) Valor según características del transporte



(c) Valor según características de los elementos de trazabilidad



(d) Valor según imputado conocido



(e) Valor según fecha del hecho

Figura 4.7: Distribución de las variables

Ponderación de variables

Las variables que cuantifican la calidad y cantidad de información de las distintas dimensiones del delito, se ponderan para construir el 'valor investigativo' de cada delito.
De esta forma, la ecuación 4.1 muestra la composición del valor investigativo en cada una de las dimensiones descritas. Donde v_i es el valor investigativo de cada causa, compuesto por vi_{fecha} que corresponde al valor investigativo que aporta la fecha, vi_{imp} es el valor investigativo que aportan los imputados conocidos de una causa, $vi_{valorcd}$ es el valor investigativo que aportan las características de los delincuentes, vi_{transp} es el valor que aportan las características del transporte de llegada y vi_{traz} es el valor que aportan las características de los elementos de registro del delito.

$$v_i = \frac{(1\frac{vi_{fecha}}{3} + 4\frac{vi_{imp}}{1} + 1.5\frac{vi_{valorcd}}{3} + 1.5\frac{vi_{transp}}{3} + 2\frac{vi_{traz}}{3})}{10} \tag{4.1}$$

En esta ecuación se observa que el valor investigativo de cada una de las dimensiones está ponderado por los números 1, 4, 1.5, 1.5 y 2 respectivamente. Cada uno de estos números representa el peso que tiene cada dimensión. De esta forma, se le asigna más valor si es que los imputados de una causa son conocidos, luego si es que existen elementos de trazabilidad o registro, en tercer lugar el valor que aportan las características de los delincuentes y el transporte de llegada y finalmente el valor de la fecha del hecho. Cabe destacar que cada una de las variables se divide por el valor máximo que puede tomar cada una de estas. Además, se aplica el método '$MinMaxScaler'$, con el objetivo de que esta variable esté entre 0 y 1. La distribución que toma esta variable aplicada a los datos se observa en la figura 4.8.

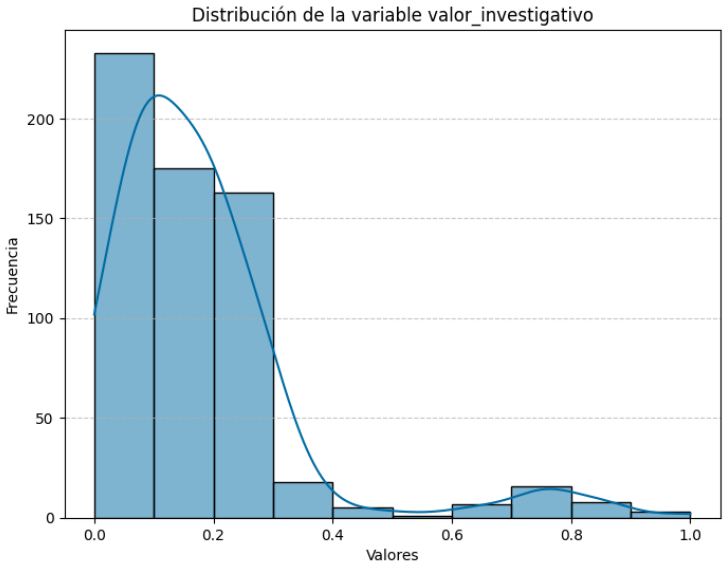


Figura 4.8: Distribución de la variable 'valor investigativo'. [Fuente: Elaboración propia]

Si bien la gran mayoría de los registros no tienen un 'valor investigativo' mayor a 0.5, se puede ver que alrededor de un tercio de los registros cuentan con un nivel 0.2 que permitiría cruzar al menos una de las variables descritas anteriormente con otras informaciones.

4.4.2. Vínculos

Medida de similitud entre causas

Durante el segundo semestre del año 2023, se desarrolló una medida de similitud entre causas [20]. Esta medida de similitud corresponde a un valor decimal que va entre el 0 y el 1 y busca dar cuenta de qué tanto se parecen dos causas. Esta métrica se construye en base a variables categóricas, binarias, numéricas y de texto, que describen un hecho delictual.

Variables categóricas

- **Categoría horaria:** Corresponde a un valor categórico que da cuenta del momento del día en que sucedió el delito.
- **Comuna:** Corresponde a la comuna en donde se cometió el delito.
- **Sexo de la víctima:** Variable binaria que da cuenta del sexo de la víctima.

Variables binarias

- **Uso de arma blanca:** Variable binaria que indica la presencia o ausencia de arma blanca, que se refiere a aquellas armas que se utilizan para infligir daño a través del corte o la perforación.
- **Uso de arma de fuego:** Variable binaria que indica la presencia o ausencia de arma de fuego.
- **Uso de objeto contundente:** Corresponde al uso de objetos contundentes como método de violencia en el hecho delictual.
- **Delito violento:** Variable que captura si el delito fue violento.
- **Lesión en la víctima:** Variable que captura si la víctima resultó lesionada.

Variables numéricas

- **Cantidad de delincuentes:** A partir del análisis del relato, se extrae la cantidad de delincuentes que observa la víctima al momento del incidente.
- **Avalúo:** Da cuenta del valor monetario de la(s) especie(s) sustraída(s).

Variables de texto

- **Resumen del relato delictual:** Corresponde a una síntesis del relato contenido en la denuncia del hecho. De forma general, el resumen del relato contiene el modus operandi, es decir, el patrón o método característico que usan el o los individuos que cometen el hecho delictual.

La siguiente tabla, extraída desde la tesis “Diseño de una medida de similitud entre delitos” [20], da cuenta de las variables contenidas en dos delitos de alta similitud.

ID Delito	1717208	1076660
Q_hora	Q1	Q3
comuna_delito	CERRILLOS	CERRILLOS
victip_sexo	hombre	hombre
avaluo_num	0	0
n_delicuentes_num	3	4
arma_fuego_dummy	1	1
arma_blanca_dummy	1	1
objeto_contundente_dummy	1	1
violencia_dummy	1	1
lesion_dummy	1	1
resumen_delito	El delito descrito involucra violencia y agresión excesiva. Según el relato, los involucrados se acercan a las presuntas víctimas en la vía pública y las agreden con armas de fuego y armas corto punzantes para sustraer sus pertenencias.	El relato menciona que la víctima fue agredida con un arma de fuego, golpes de puño y cortes con un arma punzante. Además, se menciona que hubo una discusión previa en la que la víctima también fue agredida por su conviviente.

Tabla 4.1: Ejemplo de dos delitos de alta similitud. [Fuente: 'Desarrollo de una medida de similitud entre delitos', Santander P. [20]]

Para el cálculo de la medida de similitud, se parametrizan las variables propuestas a partir del procesamiento de relatos con LLMs, tal como se indica en la etapa 'Recopilación de datos'. Una vez extraídas estas variables, se aplica una transformación de los datos para construir la similitud entre cada tipo de variable. Finalmente, se ponderan en una medida de similitud general entre todas las variables.

A continuación, se describen los resultados obtenidos en cada una de las dimensiones (categóricas, numéricas, binarias y de texto), así como el valor global de similitud. En cada uno de estos resultados se discute acerca de las ventajas, limitaciones y adaptaciones que se realizaron al cálculo de la medida de similitud.

Similitud entre variables categóricas

Para la aplicación de la similitud entre variables categóricas se consideraron las variables:

- **Q_hora:** Categoría horaria en la que sucedió el delito. Se definen 4 categorías Q1, Q2, Q3 y Q4, según los rangos $[00:00-06:00)$, $[06:00-12:00)$, $[12:00-18:00)$, $[18:00-00:00)$ respectivamente.
- **Victip_sex:** Sexo de la víctima (Hombre / Mujer).

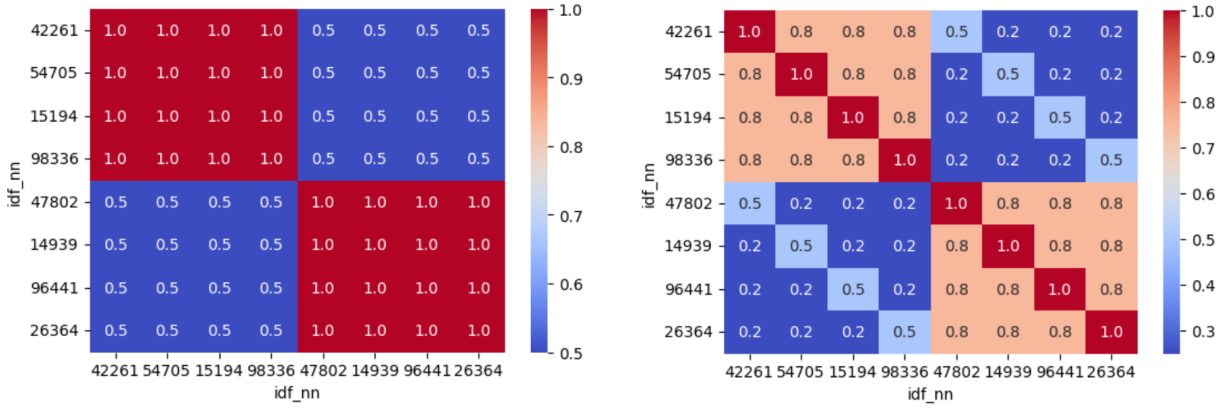
Cabe destacar que en el cálculo de similitud original, se incluía la variable 'Comuna'. Sin embargo, esta no se incluyó en este set de datos. La principal limitante al momento de incorporarla, es que aquellos delitos que no coinciden de manera exacta en la comuna, se definen como distintos, sin existir una mayor ponderación para aquellos que suceden en comunas aledañas versus otros delitos que ocurren en otras regiones. Para efectos de este trabajo se decidió trabajar con un set de delitos que ocurrieron en una sola comuna.

Al momento de aplicar la similitud de variables categóricas se adaptó ligeramente el código original. Para entender el comportamiento de esta medida de similitud se consideran 8 causas, que representan las distintas combinaciones posibles entre las variables categóricas. Estas causas contemplan los siguientes valores:

idf_nn	Q_horas	victip_sexo
42261	Q1	hombre
54705	Q2	hombre
15194	Q3	hombre
98336	Q4	hombre
47802	Q1	mujer
14939	Q2	mujer
96441	Q3	mujer
26364	Q4	mujer

Tabla 4.2: Set de causas con distintas configuraciones para variables categóricas

A partir de la matriz de similitud generada para estas causas, se elabora un mapa de calor que sintetiza los resultados.



(a) Matriz de similitud original entre variables categóricas.

(b) Matriz bajo la nueva propuesta de similitud entre variables categóricas.

Figura 4.9: Comparación de matrices de similitud bajo el enfoque original y la nueva propuesta. [Fuente: Elaboración propia]

En el cuadrante superior derecho de la figura 4.9.a, se observa la similitud entre causas donde la variable *victip_sexo* no es coincidente. En contraste, el cuadrante superior izquierdo, muestra la similitud entre causas con el mismo valor en *victip_sexo*. Lo primero que llama la atención es que no hay un valor de similitud mayor cuando hay coincidencia en la variable *Q_hora*. Es por esto que se plantea un cambio en el código donde se calcula esta similitud, para llegar a los nuevos resultados que se observan en la figura 4.9.b.

En este caso, se asigna un valor mayor cuando ambas variables coinciden, y un valor menor de similitud cuando ninguna de estas coinciden. Sin embargo, a futuro podrían incorporarse otras mejoras que logren determinar un peso a la coincidencia entre variables. Por ejemplo,

podría valorarse más la coincidencia horaria entre dos hechos delictuales que el tipo de víctima. Estos cambios deben ser fundados siguiendo las teorías de comportamiento criminal, y patrones de análisis criminal.

Además, otra observación de esta forma de calcular similitud entre variables categóricas es que hay mayor sensibilidad al número de categorías posibles. De esta forma, se penaliza más la diferencia de sexo que la diferencia entre categorías horarias, dado el número de clases posibles por cada variable.

Similitud entre variables binarias

Para este cálculo se consideran las variables originales: uso de arma blanca, uso de arma de fuego, uso de objeto contundente y lesión en la víctima.

Para un trabajo futuro, se puede estudiar la posibilidad de incorporar el objeto contundente utilizado como una variable categórica o bien de texto, de tal forma de otorgar una similitud mayor para hechos delictuales en los que se ocupó el mismo objeto contundente.

Similitud entre variables numéricas

Al aplicar la similitud entre variables numéricas se identifican algunos puntos de mejora. Para ilustrarlas de mejor manera, se toma un conjunto de 20 causas con sus variables *avaluo_num* y *num_delincuentes*. Los datos se aprecian en la tabla 4.3.

idf_nn	**avaluo_num ($)**	**n_deli_num**
19762	0	1
54705	0	1
20621	0	1
59246	0	1
26364	130000	1
19565	320000	1
87147	700000	1
14939	700000	1
98336	7000000	1
56604	0	2
53852	0	2
36233	55000	2
42261	2500000	2
15194	2800000	3
48276	300000	4

(a) 20 Causas con configuraciones distintas de variables numéricas.

idf_nn	**19762**
26364	1.00
54705	1.00
20621	1.00
59246	1.00
19565	0.99
87147	0.95
14939	0.95
56604	0.47
53852	0.47
36233	0.46
98336	0.28
48276	0.15
42261	0.04
15194	0.00

(b) Similitud de la causa 19792 con el resto de causas.

Tabla 4.3: Análisis de la similitud entre variables numéricas

La tabla 4.3.a, muestra los atributos de 20 causas con configuraciones distintas. En la tabla 4.3.b se resume la similitud de la causa de identificador 19762 con el resto, y se observa que existen 4 causas con similitud 1, lo que hace sentido ya que son causas que coinciden de forma exacta en sus variables numéricas.

El valor de similitud 0.99 llama la atención, pues corresponde a otra causa que tiene, al igual que la primera, *n_delincuentes_num* igual a 1, pero existe una diferencia de $320.000 respecto a la primera en el avalúo de la especie robada. Luego, aquella que difiere solo en el número de delincuentes, de 1 a 2, tiene un valor de similitud de 0.47. En conclusión, la variación en la similitud numérica que aporta la variable *num_delincuentes_sum* es mucho más importante que aquella que aporta la variable *avaluo_num*. Esta limitación podría resolverse abordando el cálculo de estas variables de forma separada.

Más allá de las limitaciones en el sentido de los resultados numéricos, se propone abordar la variable *avaluo_num* como una variable categórica que indique el bien que se pretendía robar, por sobre el avalúo de este. Al mirar nuevamente la tabla 4.3.a, se observan 6 causas donde el avalúo es $0, esto puede deberse a que fueron robos frustrados. Sin embargo, el patrón que se busca captar es el bien preferido que buscaba el delincuente, sin importar el éxito o fracaso de su objetivo. De esta forma, se evita tener una similitud muy baja en casos donde el bien preferido esté en un rango de precios muy amplio, como por ejemplo, un celular. Actualmente, se pueden encontrar celulares en el mercado entre $100.000 a $1.500.000, lo que incorporaría ruido a la medida de similitud.

Similitud entre variables de texto

Para el cálculo de la variable de texto, se aplica la función de similitud coseno sobre el *embedding* generado a partir del modus operandi del delito. De forma matemática, la similitud coseno implica calcular la distancia angular entre dos vectores en el espacio.

Este enfoque ofrece ventajas a la hora de comparar textos similares mediante la técnica de *embeddings*, sin tener problemas de consistencia debido a la longitud de distintos textos. A futuro, se plantea la posibilidad de incorporar esta técnica en otro tipo de datos de texto, como por ejemplo, la descripción de los delincuentes.

Cabe destacar que se calcula el *embedding* sobre el modus operandi, y no sobre el relato completo, dado que se captura de mejor forma el sentido del texto, sin tener ruido debido a marcas textuales repetitivas, tales como *'Relación de los hechos, doy cuenta (...)'*, *'Se presenta ante (...)'* que varían según la institución que da cuenta del delito. El modus operandi es generado por el modelo gpt-4, bajo la indicación de tener un lenguaje neutro, sin hacer referencia a lugares concretos o individuos en particular, sino que a la descripción de técnicas o circunstancias en las que se desarrolló el delito.

Ponderación final

Una vez calculadas las similitudes por tipo de variable, se consolidan en la similitud final entre causas, definiendo ponderadores para cada variable. Para este caso, luego de las limitaciones levantadas anteriormente, se elige la configuración siguiente:

- Ponderador Variables Categóricas: 1
- Ponderador Variables Binarias: 2
- Ponderador Variables Numéricas: 1

- Ponderador Variables de Texto: 6

Con estas ponderaciones se busca establecer relaciones entre causas, centrándose principalmente en el modus operandi. Luego, el tipo de armas utilizadas tiene mayor valor que las categóricas y numéricas dada la interpretación sobre el set de datos que se realizó.

Umbral de similitud

Una vez construida la matriz de similitud para el set de datos, se obtiene una similitud entre causas que va de 0.87 a 1. La justificación del rango en que se encuentran estos valores, está en la ponderación asignada a las variables de texto. Dado que las variables de texto son aquellas que tienen más peso en la medida de similitud final, es esperable que los valores tiendan a ser más altos. Esto se explica ya que el enfoque de *embeddings* permite calcular la similitud coseno de dos textos en base a la naturaleza de estos. En principio, cualquier tipo de delito estará cercano entre sí, dado que comparten el tema de hechos ilícitos, independiente del nivel de violencia, arma utilizada, o bien sustraído. Esto haría que incluso delitos que son muy distintos, como un homicidio y un hurto, estén relativamente cercanos.

Luego, se define un umbral a partir del cual se considera que dos nodos se conectan entre sí, en caso contrario, la red social generada tendría densidad 1 dado que todos los nodos estarían conectados entre sí. Esto haría que el análisis de la red se vuelva computacionalmente más complejo y podría entregar información menos relevante.

Para la determinación del umbral, se estudia la variación de densidad, modularidad y cantidad de nodos que comprende el mayor grafo que los contiene, a medida que aumenta el umbral. Los resultados se ven en la figura 4.10.

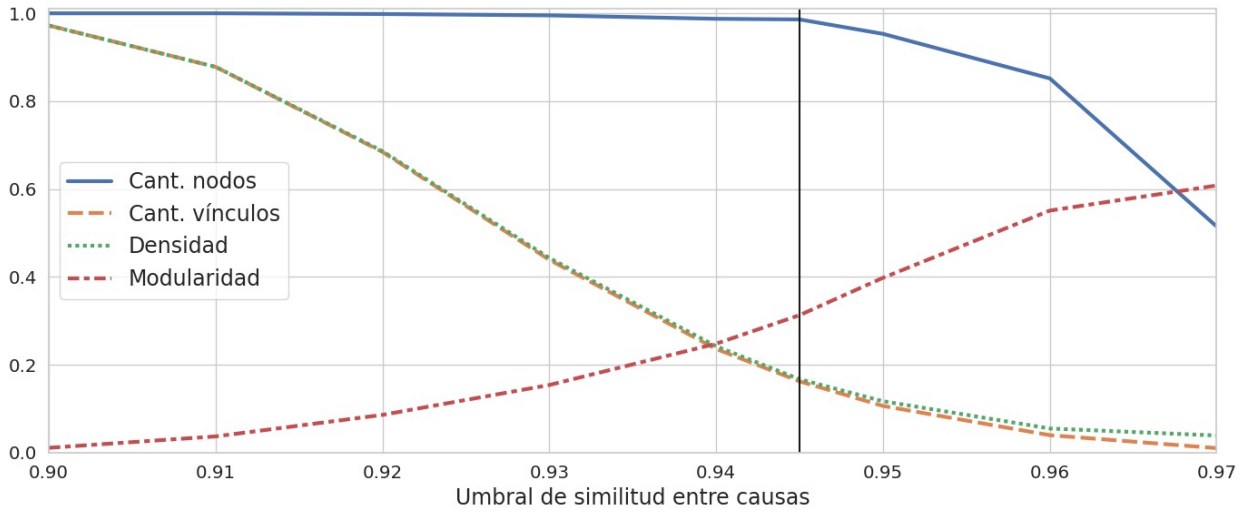


Figura 4.10: Tamaño, densidad y modularidad de la red social en función del umbral de similitud. [Fuente: Elaboración propia]

Como es de esperarse, la cantidad de vínculos o densidad son conceptos análogos que dan cuenta de qué tan conectada está la red, y por ello siguen curvas idénticas. Al mirar la cantidad de nodos, esta comienza a decrecer con una tasa mayor a partir del nivel 0.945, por

otra parte, a medida que la densidad de la red disminuye, existe mayor modularidad, dado que existen comunidades o grupos más definidos.

Considerando que se busca analizar las causas ocurridas en una comuna en un rango de tiempo, es importante tener la capacidad de analizar todas o un gran porcentaje. Es por esto que se determina como umbral el valor 0.945, que permite analizar un 98,6 % de las causas en un grafo conexo, teniendo el mejor valor posible de modularidad, que permita una mejor identificación de comunidades o *clusters*.

4.5. Planteamiento del Modelo de Optimización

Motivación

Para entender el modelo matemático, es relevante comprender el objetivo del modelo. De forma general, el modelo busca proponer causas dentro de un foco delictivo, que sean valiosas para las primeras diligencias investigativas. De esta forma, si se tiene un universo de 40 causas que guardan relación con un tipo de fenómeno delictual, se escogerá un subconjunto de causas que maximice el 'valor investigativo' en su conjunto, minimizando la distancia entre estas causas.

Para reforzar el concepto veamos un ejemplo más práctico. Suponemos que una Fiscalía Regional declara un foco ligado al robo de especies al interior de autos estacionados en la vía pública. Bajo el enfoque manual que los equipos siguen actualmente, el proceso que se sigue es:

1. Consolidar un universo acotado de causas que cumplan con filtros en la base de datos. Estos filtros pueden incluir espacios temporales, espacios geográficos, y presencia de palabras claves ligadas al fenómeno.
2. Leer manualmente relatos de causas que guarden relación con el fenómeno estudiado,
3. Parametrizar la información presente en los relatos de las causas.
4. Analizar si hay *match* entre las variables de la tabla parametrizada.
5. Proponer un número de causas que posibiliten las primeras diligencias investigativas.

Bajo el nuevo enfoque del modelo de redes sociales, el proceso que se podría aplicar es:

1. Consultar al modelo en base a un RUC que tiene un relato. (Para este ejemplo, un relato relacionado al fenómeno).
2. El modelo construye la red de causas más cercanas que podrían constituir el universo acotado descrito en el punto 1 del proceso anterior.
3. El modelo selecciona un número de causas similares que optimizan el 'valor investigativo' global, y que, por lo tanto, permitirían focalizar la lectura de causas pertinentes.

Construcción del Modelo Matemático

Elección y definición del modelo

El modelo matemático propuesto se basa en el Modelo StRAM [8], para orientar la investigación criminal y facilitar la consecución del material probatorio.

Parámetros

$$vi_i = \text{valor investigativo del nodo } i$$

$$s_{ij} = \text{similitud entre las causas } ij$$

Variables de decisión

$$y_i = \begin{cases} 1, & \text{si el nodo i pertenece al conjunto solución} \\ 0, & \text{en caso contrario} \end{cases} \tag{4.2}$$

$$x_{ij} = \begin{cases} 1, & \text{si el arco i,j pertenece al conjunto solución} \\ 0, & \text{en caso contrario} \end{cases} \tag{4.3}$$

$$f_{ij} = \text{flujo a través del arco (i,j) en el conjunto solución} \tag{4.4}$$

Función objetivo

$$\max P = \frac{1}{vi^{max}} \sum_{i=1}^{N} v_i y_i - s_{ij}^{max} \sum_{n=1}^{N} \frac{1}{s_{ij}} x_{ij} \tag{4.5}$$

Restricciones

1. Restricción de antecesor
$$\sum_{i \in N} x_{ij} = y_j \qquad \forall j \in N \tag{4.6}$$

2. Conservación de flujo
$$\sum_{i \in N} f_{ij} - \sum_{i \in N} f_{ji} = y_j \qquad \forall j \in N \tag{4.7}$$

3. Link entre variables
$$f_{ij} \leq (|N| - 1) x_{ij} \qquad \forall (i,j) \in A \tag{4.8}$$

4. Valor investigativo mínimo
$$\sum_{i \in N} y_i \geq vi^{min} \tag{4.9}$$

5. Dominio de las variables
$$x_{ij} \in \{0,1\} \qquad \forall (i,j) \in A \tag{4.10}$$
$$y_j \in \{0,1\} \qquad \forall j \in N \tag{4.11}$$
$$f_{ij} \geq 0 \qquad \forall (i,j) \in A \tag{4.12}$$

Este modelo permite ajustar ciertos parámetros según el fenómeno que se esté analizando. En particular, se plantea la posibilidad de que los analistas puedan calibrar el valor vi^{min}, dado que dentro de un foco, podrían existir delitos con mucha más información que otros. Luego, es esperable que si las causas de un foco en general tienen más información, esta restricción sea más fuerte.

4.6. Validación del Modelo

Para interpretar los resultados del modelo, se realiza la visualización de la red construida y con Análisis de Redes Sociales se identifican comunidades, y se comentan algunas métricas relevantes explicadas en el Capítulo 1. Además, para la interpretación de los resultados del Modelo de Optimización, se toma un set de causas que permitan interpretar las sugerencias del modelo.

En el capítulo Capítulo 5 de Resultados y discusión, se analizan con más detalle las métricas declaradas en el Capítulo 1 con el objetivo de evaluar la efectividad del modelo propuesto junto con las posibles áreas de aplicación, ventajas y limitaciones.

Interpretación de resultados

Se modelan las 637 causas en la red social propuesta. El rango de valores de los vínculos va de 0.872 a 1.0. Se fija el umbral de similitud en el nivel 0.945, valor que permite que la red sea conexa, con un nivel de modularidad 0.313 y densidad 0.167. Tal como se definió en el Capítulo 2, estos conceptos dan cuenta del nivel de separación de la red en grupos identificables de nodos, y de la cantidad de conexiones que existen entre los nodos de la red.

Utilizando el Software Gephi[10], se genera una visualización de la red social que se observa en la figura 4.11, utilizando el Layout[11] Force Atlas 2, que permite una distribución espacial que considera las fuerzas de atracción entre nodos dadas por el peso o similitud entre estos, además de entregar una visualización intuitiva de las comunidades que se forman.

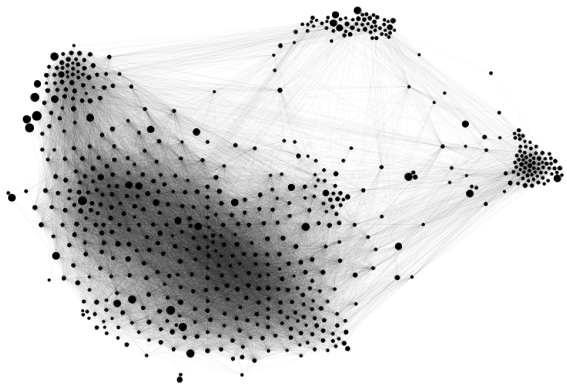


Figura 4.11: Modelamiento de 637 causas delictuales mediante redes sociales.
[Fuente: Elaboración propia]

[10] Herramienta de software de código abierto utilizada para el análisis y la visualización de datos complejos, especialmente en el ámbito de las redes y grafos.
[11] Layout: Algoritmo de diseño que ayuda a organizar los nodos según el objetivo deseado.

Cabe mencionar que existe una diferencia en la tonalidad de fondo de cada agrupación de nodos, según la cantidad de vínculos que existan entre ellos. A partir de esta primera visualización, ya es posible identificar clusters o comunidades bien delimitadas.

A partir de esta red, se calcula la modularidad utilizando el parámetro de resolución por defecto igual a 1. Esto hace sentido ya que se busca tener un panorama general de las comunidades que pueden detectarse, sin tener una sensibilidad por las comunidades muy pequeñas o muy grandes. Además, en el cálculo de la modularidad se considera el peso de los arcos, ya que interesa agrupar causas que además de estar muy conectadas entre sí, estén 'fuertemente' conectadas, en función del peso del arco (similitud entre causas).

Una vez calculada la modularidad, se particiona la red en función de la clase de modularidad asignada a cada nodo. La figura 4.12 ilustra como se identifican las comunidades dentro de la red original.

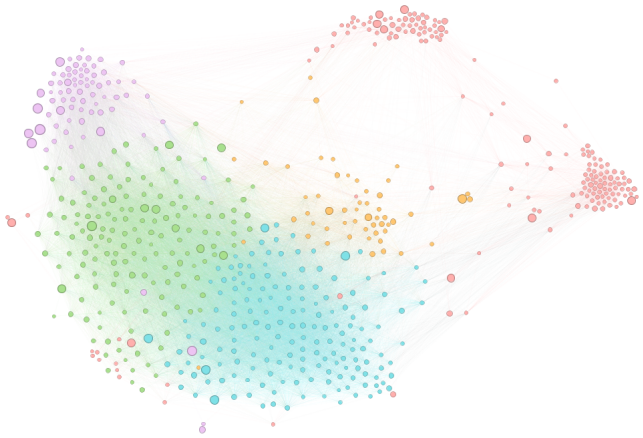
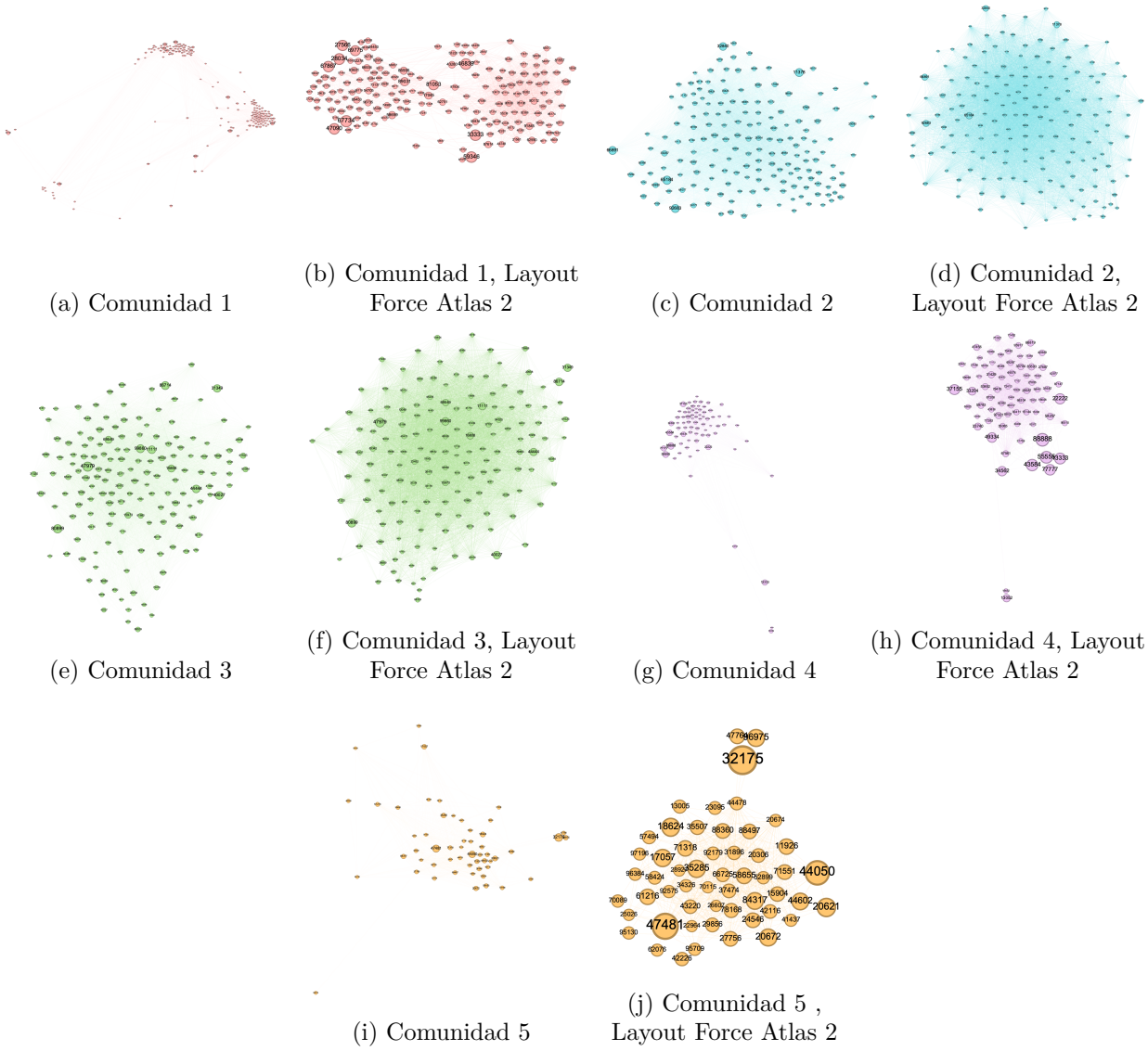


Figura 4.12: Detección de comunidades mediante el cálculo de modularidad. [Fuente: Elaboración propia]

Análisis de Comunidades

A partir del análisis de métricas de cada una de las comunidades, es posible entender la estructura de esta y así interpretar los resultados en el análisis criminal. La figura 4.13 ilustra las comunidades encontradas en la visualización original, y su distribución luego de aplicar el Layout Force Atlas 2, que permite ver considerar la fuerza de atracción entre nodos similares.



(a) Comunidad 1

(b) Comunidad 1, Layout Force Atlas 2

(c) Comunidad 2

(d) Comunidad 2, Layout Force Atlas 2

(e) Comunidad 3

(f) Comunidad 3, Layout Force Atlas 2

(g) Comunidad 4

(h) Comunidad 4, Layout Force Atlas 2

(i) Comunidad 5

(j) Comunidad 5 , Layout Force Atlas 2

Figura 4.13: Comunidades de la red original, y su nueva visualización con el Layout Force Atlas 2

La tabla 4.4 resume algunas métricas que dan cuenta del tamaño y densidad de cada una de las comunidades.

Causas	**C1**	**C2**	**C3**	**C4**	**C5**
Cantidad de Nodos	175	168	159	76	54
Cantidad de Vínculos	2042	9080	7341	1793	595
Densidad	0.134	0.647	0.584	0.629	0.416
Long. de Ruta Promedio (APL)	2.537	1.355	1.419	1.42	1.685

Tabla 4.4: Métricas asociadas a las comunidades encontradas

Las comunidades están ordenadas según la cantidad de nodos que contiene cada una respecto de la red original. La más numerosa corresponde a la Comunidad 1, pero al mismo tiempo, es la red menos densa. Esto podría significar que las causas que contiene no se relacionan de forma tan fuerte entre todas. Por otra parte, la Comunidad 2 presenta el nivel de densidad más alto, lo que podría significar que las causas contenidas en esta comunidad son más homogéneas entre sí, permitiendo una alta conexión entre todas.

Para corroborar la correcta interpretación de los datos, se toman 5 causas de la Comunidad 1 y 2, dado que contienen mayor cantidad de nodos y al mismo tienen los valores mínimo y máximo de densidad respectivamente, al compararlas con el resto de comunidades.

idf_nn	**resumen_delito**
71321	Intimidación con arma blanca para robo de guitarra.
95917	Sustracción de pertenencias mediante amenaza con arma blanca.
62977	Sujeto se acercó y extrajo un arma blanca tipo cuchillo con la cual intimidó a la víctima para robar su teléfono celular.
75825	Sustracción de teléfono celular con intimidación mediante arma blanca.
33642	Hurto de armamento desde el domicilio particular en el closet del dormitorio.

Tabla 4.5: Muestra de 5 causas en la Comunidad 1

idf_nn	**resumen_delito**
97806	Sustracción de teléfono celular desde el interior del vehículo aprovechando que el vidrio de la puerta delantera estaba abierto, en un semáforo en luz roja.
39160	Lanzamiento de piedra para quebrar vidrio y robo de teléfono celular.
78349	Fractura de vidrio del vehículo y sustracción de celular.
58714	El delincuente quebró el vidrio lateral delantero derecho del vehículo con una piedra y sustrajo un teléfono celular.
36900	Sujeto lanza piedra a vehículo detenido en tráfico para quebrar vidrio y sustraer especies, no logra su cometido y huye.

Tabla 4.6: Muestra de 5 causas en la Comunidad 2

El resumen de los delitos confirma la interpretación inicial, donde a mayor densidad mayor homogeneidad en el tipo de delitos que se incluyen. Esta medida de densidad podría ser un buen parámetro para determinar qué tan útil es una comunidad para declarar un Foco Investigativo.

En casos donde la densidad de la comunidad no sea 'suficientemente buena', en base a los criterios que se definan con los profesionales expertos, se podría volver a iterar en la detección de comunidades en grupos con delitos sin muchas características en común. Cabe destacar que para este caso de aplicación, se calculó la modularidad con el software Gephi, utilizando el parámetro de resolución igual a 1. Sin embargo, si se quisieran detectar más comunidades, bastaría con disminuir este valor de manera que sea cercano a 0. Para efectos de la decisión de cuántas comunidades detectar a priori, se podría trabajar con un gráfico que muestre cómo mejora o empeora la modularidad (que da cuenta de la calidad de las comunidades identificadas) en función del número de comunidades.

Por otra parte, la longitud de ruta promedio o *Average Path Length*, en inglés, da cuenta de que en promedio, es necesario recorrer x veces la longitud máxima para pasar de un nodo a otro. Nuevamente, la Comunidad 1 destaca por tener el valor más grande en APL, esto significa que para conectar una causa con otra deben pasar por más causas intermedias.

La tabla 4.7 muestra la causa intermedia que conecta el par de causas 59246 - 24465.

idf_nn	**resumen_delito**
59246	Sustracción de placas patentes de un minibús estacionado en la vía pública utilizando llaves falsas y/o ganzúas.
24465	Sustracción de vehículo estacionado.
71067	**Uso de llaves falsas y/o ganzúas para sustraer vehículo.**

Tabla 4.7: Causa intermedia entre dos causas de la Comunidad 1.

La causa que los conecta (71067), comparte características con ambos. Con la causa 59245 comparten el objeto que utilizan para cometer el delito, y con la causa 24456 comparten el bien sustraído. Esto sustenta la idea desarrollada en el modelo de optimización, donde se impone que las causas pertenecientes al conjunto solución estén conectadas entre sí.

Distribución de la variable 'valor investigativo'

Una de las variables propuestas en este trabajo corresponde al valor investigativo. En la figura 4.14 se observa la distribución y dispersión de esta medida en cada comunidad.

A partir de esta visualización, es posible ver que la Comunidad 1 tiene una mediana más alta que el resto, junto con el rango de valores que abarca. Al tener un abanico más grande de causas con mucha información, los resultados del modelo de optimización podrían ser mejores que los que se encuentren en la Comunidad 3, ya que se observan solo 3 causas con valor investigativo mayor a 0.5, y además, la mediana alcanza uno de los valores más bajos en comparación al resto de comunidades.

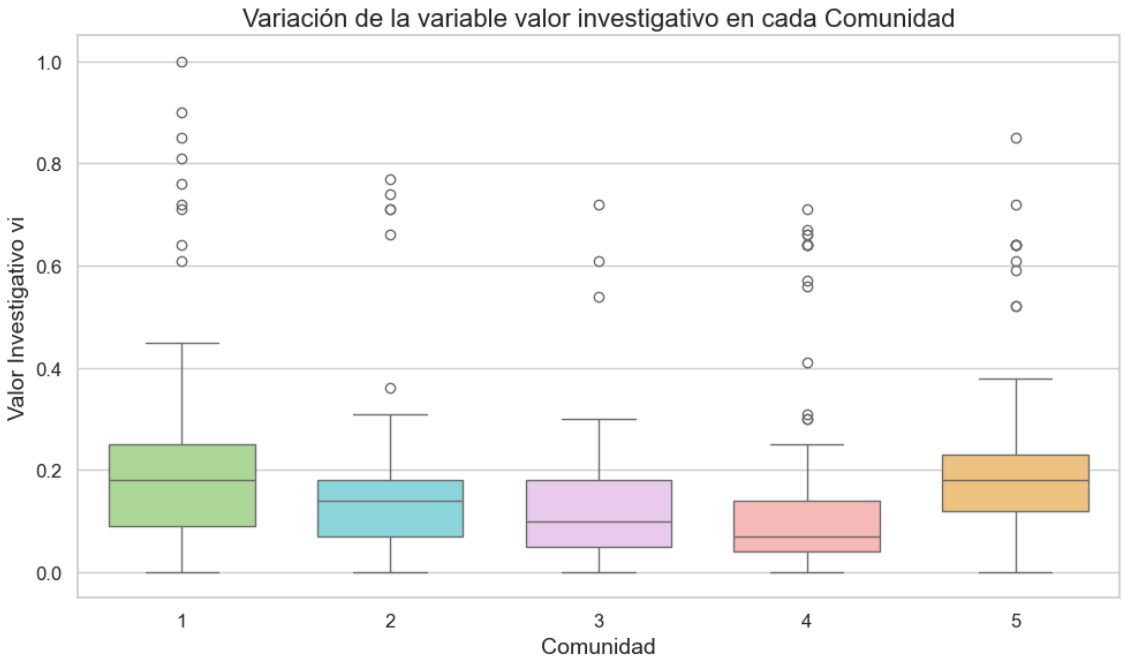


Figura 4.14: Distribución del valor investigativo en cada comunidad.
[Fuente: Elaboración propia]

Analizar esta variable, podría ser un buen criterio para dar una prioridad sobre el foco o Comunidad a investigar, dado que este representa también las posibilidades que tendrán las causas de cruzar información con otras que ya hayan sido resueltas.

Aplicación del Modelo de Optimización

Tal como se menciona en la sección 4.5, el propósito del modelo de optimización es la selección de causas relevantes para una investigación donde no se cuenta con imputados conocidos[12]. De tal forma que mediante la red social propuesta se visualizan aquellas causas similares, y el modelo de optimización sugiere el subconjunto de causas más pertinentes, considerando similitud y valor investigativo, definido como la cantidad y calidad de la información presente en el relato delictual.

Para aplicar el modelo de optimización sobre el conjunto de datos, se elige una comunidad, de tal manera que el grafo no sea tan grande y el modelo opere de forma más ágil. Esto se desarrolla con el lenguaje de programación Python y el solver de optimización Gurobi, que ofrece una licencia académica.

Para interpretar los resultados que ofrece el modelo, se elige la comunidad 4. Luego, se elige la causa 99860, y se consulta al modelo. Los atributos de la causa consultada y la respuesta se ilustra en la tabla 4.8.

[12] Sin perjuicio de esto, se evidencia que el modelo tiene otros potenciales usos, por ejemplo a partir de causas con imputado conocido se pueden buscar otras que sean 'candidatas' a resolverse con la información de esta causa con Imputado Conocido. Estas observaciones se discuten con mayor detalle en el Capítulo 5.

idf_nn	resumen_delito	vi	imp_vi	caracteristicas_imp	transporte_imp	evidencia_hec
13089	Los delincuentes detuvieron su furgón al lado de la víctima, descendieron y mediante intimidación con arma blanca y revisión de bolsillos sustrajeron pertenencias y dinero.	0.28	0.0	Estatura: 1.70 metros; Contextura: delgada; Color_pelo: negro; Tez: blanca; Vestimenta: polera burdeo, pantalón jeans azul; Estatura: 1.60 metros; Contextura: gruesa; Color_pelo: rubio; Tez: blanca; Vestimenta: blusa burdeo, falda roja; Estatura: 1.70 metros; Contextura: media; Color_pelo: negro; Tez: blanca; Vestimenta: polera blanca con rayas negras, pantalón claro; Rol: Conductor, Vestimenta: No informado	Furgón Fiat Fiorito color blanco	/
99860	Los delincuentes descendieron de un vehículo, insultaron y amenazaron a la víctima, y tras un forcejeo le sustrajeron su billetera con dinero en efectivo.	0.64	1.0	/	Vehículo color gris, placa patente XXXX11[a], marca Toyota, modelo RAV 4, año 2009	Reconocimiento de la propietaria del vehículo involucrado mediante dispositivo SIMCAR.

Tabla 4.8: Causa sugerida a partir de la aplicación del Modelo de Optimización Lineal

[a] La patente fue adulterada con el fin de proteger los datos personales, sin embargo, en su registro original se encuentra completa.

A partir del relato, es posible identificar que existen similitudes entre el modus operandi: en ambos casos los delincuentes descienden de un vehículo e intimidan a la víctima para sustraer la misma especie, pertenencias y/o dinero. La ventaja de realizar esta asociación es que la causa consultada tiene imputados conocidos que eventualmente podrían ser los mismos que participaron en el delito 13089. Una de las estrategias que se utilizan para identificar los imputados es el contacto con las víctimas, con el objetivo que puedan reconocer o no si fueron los mismos imputados.

Para probar el modelo se realizaron otros experimentos que se resumen y discuten en el capítulo 5. Para mayor claridad y facilitar la interpretación de resultados, se expone otra aplicación. En este caso se tomó la causa con identificador 54257, y se ejecutó dos veces el modelo de optimización variando el parámetro vi^{min}, expuesto en la restricción 4 de la sección 4.5. En el primer ejercicio se consideró el valor vi^{min}=1 y en el segundo vi^{min}=2. Las causas encontradas en cada aplicación se observan en las tablas 4.9 y 4.10.

idf_nn	resumen_delito	vi	imp_vi	caracteristicas_imp	transporte_imp	evidencia_hec
54257	Intimidación con cuchillo y pistola, golpe en la cabeza, robo de pertenencias y vehículo.	0.31	0.0	Hombre, estatura baja, contextura delgada, cabello negro, tez trigueña, pantalón negro, polerón negro; Hombre, pantalón gris, zapatillas negras, chaqueta rojo azul, jockey gris marca Jordan; Hombre, jeans azul, polera gris, jockey azul, tez morena, contextura delgada, estatura mediana; Mujer, 18 años aproximadamente, estatura baja, contextura delgada, cabello negro, tez trigueña, pantalón negro, polerón negro.	-	/
88888	Intento de robo de vehículo, uso de arma de fuego para intimidar a las víctimas, fuga en vehículo.	1.0	1.0	Primer detenido: 25 años, chileno, soltero, vestía polera negra y pantalón jeans. Segundo detenido: 25 años, chileno, soltero, vestía polerón plomo y pantalón negro.	Vehículo Suzuki Aerio color gris plateado, año 2015, patente XXXX12[a]	Cámaras de seguridad, GPS, pruebas balísticas, cadena de custodia de la pistola y ganzúa, informe preliminar de balística.

Tabla 4.9: Causa sugerida a partir de la aplicación del Modelo de Optimización Lineal, con parámetro vi^{min}=1.

[a] La patente fue adulterada con el fin de proteger los datos personales, sin embargo, en su registro original se encuentre completa.

idf_nn	resumen_delito	vi	imp_vi	caracteristicas_imp	transporte_imp	evidencia_hec
54257	Intimidación con cuchillo y pistola, golpe en la cabeza, robo de pertenencias y vehículo.	0.31	0.0	Hombre, estatura baja, contextura delgada, cabello negro, tez trigueña, pantalón negro, polerón negro; Hombre, pantalón gris, zapatillas negras, chaqueta rojo azul, jockey gris marca Jordan; Hombre, jeans azul, polera gris, jockey azul, tez morena, contextura delgada, estatura mediana; Mujer, 18 años aproximadamente, estatura baja, contextura delgada, cabello negro, tez trigueña, pantalón negro, polerón negro.	-	/
88888	Intento de robo de vehículo, uso de arma de fuego para intimidar a las víctimas, fuga en vehículo.	1.0	1.0	Primer detenido: 25 años, chileno, soltero, vestía polera negra y pantalón jeans. Segundo detenido: 25 años, chileno, soltero, vestía polerón plomo y pantalón negro.	Vehículo Suzuki Aerio color gris plateado, año 2015, patente XXXX12	Cámaras de seguridad, GPS, pruebas balísticas, cadena de custodia de la pistola y ganzúa, informe preliminar de balística.
22222	Interceptación del vehículo en un cruce de calles, intimidación con arma de fuego y arma blanca, agresión, sustracción de teléfono, vehículo y dinero.	0.72	1.0	Grupo de 5 sujetos, 4 hombres y una mujer.	-	CGPS del vehículo sustraído, cámaras de seguridad no mencionadas.

Tabla 4.10: Causa sugerida a partir de la aplicación del Modelo de Optimización Lineal, con parámetro vi^{min}=2.

Del ejercicio realizado con ambas configuraciones, con $vi^{min} = 1$ y $vi^{min} = 2$, se observa que la primera configuración permite recuperar una sola causa, que al contener mucha información presenta un valor investigativo máximo. Sin embargo, la descripción de los delincuentes no coincide, por tanto es posible pensar que no se trata de los mismos sujetos.

En la segunda iteración, con $vi^{min} = 2$, el modelo sugiere una nueva causa donde el resumen del delito coincide, y además hay mayor grado de coincidencia en la descripción de los delincuentes ya que se identifica la participación de una mujer.

De este ejercicio, se destaca la importancia de una correcta calibración del parámetro vi^{min}, para que el modelo tenga la capacidad de sugerir una cantidad suficiente de causas. Sin embargo, resulta difícil establecer un valor de vi^{min} fijo para cualquier delito, dado que según las características de la red social en la que se encuentre, los resultados del modelo podrían diferir en gran medida.

Por ejemplo, para un delito que tiene un nivel de especificidad muy alto en el modus operandi, y donde no se han logrado levantar mucha información de los relatos, es posible que con un nivel $vi^{min} = 1$, se sugieran muchas causas que en su conjunto alcanzan el nivel de valor investigativo deseado, pero que individualmente no superen un valor investigativo de 0.5. Sin embargo, para otras causas donde en general el valor investigativo es más alto, el modelo sugerirá tan solo una o dos causas ya que se alcanzará el valor vi^{min} de forma más fácil.

Es importante destacar entonces que el parámetro vi^{min} no tiene una forma estándar para todos los casos, y en un futuro podría pensarse en una función auxiliar que lo calcule en función de las características de la red social asociada a la causa. Sin embargo, para este trabajo se realiza un análisis de sensibilidad variando este parámetro con los valores $vi^{min} = 1$ y $vi^{min} = 2$, tal como se expone en el Capítulo 5.

Capítulo 5

Resultados y discusión

El presente capítulo se estructura en dos secciones. En la primera sección solo se exponen los resultados de los experimentos realizados según las métricas definidas en la subsección 1.7. Además de las tres métricas propuestas, se evalúa una nueva arista relacionada con los modelos StRAM y LiRAM que ya han sido desarrollados, con el fin de ver la factibilidad de interactuar y enriquecer los resultados con estos modelos.

En la segunda sección se discute sobre los resultados obtenidos, identificando las ventajas y limitaciones en cada dimensión.

5.1. Resultados

5.1.1. Identificación de Focos Investigativos

Durante el año siguiente, al que se considera en el período de tiempo de los delitos estudiados, se declaró un foco investigativo a partir de un problema delictual en el cual se identificaban delitos con el mismo modus operandi, en una zona caliente, con un mismo bien preferido. Para ver la factibilidad de detectar o formar focos investigativos a partir de la detección de comunidades, se etiquetó de forma manual si el delito cumplía las características del foco declarado, encontrando un total de 179 causas. De esta forma, busca identificar cuántas de estas causas se encuentran en cada comunidad. Si existe una cantidad homogénea en cada uno, las comunidades no tendrían capacidad de identificar focos investigativos. Por el contrario, si una gran cantidad de estos se encuentran en una comunidad, esta comunidad sugeriría la posible declaración de un Foco Investigativo.

Los resultados del ejercicio se condensan en la tabla 5.1. La primera fila muestra la cantidad de causas del Foco 1 que fueron encontradas en cada comunidad. Luego, la segunda línea contiene la cantidad total de causas por comunidad. De tal forma que la última línea muestra el porcentaje de causas del foco 1 en cada comunidad.

Causas	C1	C2	C3	C4	C5
Causas Foco 1	4	153	14	3	5
Total de Causas por Comunidad	174	169	157	77	54
Porcentaje del Foco 1 por Comunidad	2,24 %	85,47 %	7,82 %	1,67 %	2,79 %

Tabla 5.1: Causas del Foco 1 en cada Comunidad.

Con este enfoque, es posible captar un 85,47 % de las causas del Foco 1 en la Comunidad 2. El resto de causas faltantes se distribuye de forma homogénea en el resto de las comunidades con una leve diferencia en la comunidad 3 donde se encuentra cerca del 8 %.

Se repite el mismo ejercicio para un segundo Foco 2, etiquetado manualmente. En esta ocasión, el Foco 2 se caracteriza por el medio que utilizan los delincuentes para cometer el hecho, repitiendo el mismo modus operandi. De forma manual, se identifica que 13 causas corresponden al fenómeno descrito. Los resultados se muestran en la tabla 5.2.

Causas	C1	C2	C3	C4	C5
Causas Foco 2	0	0	12	0	1
Total de Causas por Comunidad	174	169	157	77	54
Porcentaje del Foco 2 por Comunidad	0,00 %	0,00 %	92,30 %	0,00 %	7,69 %

Tabla 5.2: Causas encontradas en cada cluster para el Foco 2.

De forma preliminar, los resultados muestran que es posible encontrar más del 80 % de las causas en una comunidad. Estos resultados se comentan de forma más extensa en la sección 5.2.

5.1.2. Tiempo de análisis de las causas

El modelo que se propone en este trabajo de título, contempla la utilización de herramientas de Ciencia de Datos y Modelos 'grandes' de lenguaje (LLMs) para automatizar procesos que en la actualidad se realizan de forma manual. La aplicación del modelo consta de distintas etapas. A continuación se detalla cada una de estas junto con el tiempo que toma el procesamiento aproximado de 650 causas.

- **Construcción de variables con modelos de lenguaje:** Para procesar los relatos del hecho delictual y extraer las variables descritas en el capítulo anterior, el modelo gpt-4 tarda aproximadamente 6 horas, comprendiendo las tareas de generación de texto, generación de *embeddings* y categorización del nivel de información.
- **Aplicación de la medida de similitud:** Una vez generadas todas las variables necesarias para el cálculo de la similitud, la aplicación del flujo que calcula la similitud entre causas toma cerca de 5 minutos.
- **Aplicación del modelo de optimización:** Al realizar una consulta al modelo de optimización, este puede demorar de 1 a 30 minutos según la densidad del grafo inicial que se entrega.

Por otra parte, para realizar una comparación en cuanto al tiempo que toma realizar este proceso de forma manual, se declaran algunos supuestos:

- Para el cálculo se considera el tiempo que necesita una persona dedicada en forma exclusiva a leer, parametrizar e identificar patrones en un conjunto de 650 relatos.
- La extensión de los relatos varía entre 20 hasta 12360 palabras, con un promedio de 650 palabras por relato.
- Se estima un ritmo de trabajo regular, entendiendo que la concentración por largos períodos de tiempo podría variar.

Bajo el supuesto de que una persona se dedicara en forma exclusiva a la lectura, parametrización y análisis de detección de patrones de un conjunto de 650 causas, este proceso tomaría alrededor de 52 horas. Considerando una capacidad de procesamiento de 100 causas diarias a un ritmo regular.

5.1.3. Desempeño del modelo de optimización

Para probar el modelo de optimización se toma un set de datos de validación, que corresponden a pares o tríos de delitos donde se repite al menos uno de los imputados. En ellos interesa ver el resultado de la consulta al modelo con 6 causas. Se escogieron pares con imputados coincidentes, y luego, según la estructura, se decidió dejar una de ellas con Imputado Desconocido, en caso de que en el relato no hiciera mención explícita al delincuente detenido. El set de datos de validación se describe en la tabla 5.3.

Causas ID	**Resumen**	**Causas IC**	**Resumen**
65487	Sujeto conocido en el sector por robos, abre vehículo y sustrae batería y radio.	11111	El delincuente se acercó a la víctima y le exigió la entrega de sus aros con amenazas verbales y físicas.
92954	Ingreso al domicilio y sustracción de bienes	44444	El delincuente desprendió el portón corredero del cierre perimetral y la protección de la ventana de la cocina para ingresar al inmueble y sustraer pertenencias.
14444	Pasajero de Uber intimida al conductor con arma de fuego y cuchillo para sustraer teléfono y dinero.	43584	Robo con intimidación, fuga en vehículo, colisión durante persecución, escape a pie.
54257	Intimidación con cuchillo y pistola, golpe en la cabeza, robo de pertenencias y vehículo	22222	Interceptación del vehículo en un cruce de calles, intimidación con arma de fuego y arma blanca, agresión, sustracción de teléfono, vehículo y dinero.
98044	Sujeto aprovechó congestión vehicular para sustraer teléfono celular a través de la ventana del vehículo de la víctima.	11378	Arrebato de teléfono celular desde las manos de la víctima aprovechando que el vidrio del vehículo estaba abierto
12222	Lanzamiento de piedra para romper ventanilla y sustracción de teléfono celular		

Tabla 5.3: Set de datos de validación.

En cada aplicación, se escogieron los subgrafos conformados por las comunidades en las cuales se encontraba cada causa.

Cabe mencionar que si bien el set de datos de validación es pequeño, dados los datos iniciales con los que se contaba, el modelo permite reconocer en 4 de 6 casos la causa que involucraba al mismo imputado. Esto se logra considerando un nivel de vi_{min} igual a 2, tal como lo muestra la tabla 5.4[13].

[13] El relato de las otras causas sugeridas por el modelo se detallan en el Anexo, Tabla A.1.

ID	Resultados ($vi_{min} = 1$)	Resultados ($vi_{min} = 2$)	Coincidencia Imp
65487	99999, 80899, 65487	99999, 80899, 47979, 65487	No
92954	83631, 92954, 69775	92954, **44444**, 69775, 47090	Sí
14444	99860, 36233, 14444	60551, 99860, 36233, 47979, 14444	No
54257	54257, 88888	54257, **22222**, 88888	Sí
98044	**11378**, 98044, 95437	33207, **11378**, 98044, 65184	Sí
12222	25917, **11378**, 12222	25917, **11378**, 30063, 66891, 12222	Sí

Tabla 5.4: Causas sugeridas por el modelo a partir de un ID de causa.

Análisis de relaciones entre causas e imputados de una misma comunidad

De forma adicional, se estudia la relación entre Imputados Conocidos de causas que pertenecen a una misma Comunidad, con el objetivo de ver si este modelo puede, en el futuro, identificar bandas criminales que se dediquen a cometer el mismo tipo de delitos. Para esto, se define la relación entre imputados si es que existe un camino que los conecte en el enfoque tradicional de la red social que proponen los Modelos StRAM y LiRAM.

Para este análisis, se identifican por cada comunidad las causas donde los imputados están relacionados (de forma directa o a través de otros imputados intermediarios).

	C1	C2	C3	C4	C5
Conjunto de causas relacionadas	-	98044-11376-12222	64361-99999 64361-47979	37155-54257-22222 43584 - 88888 34562-13333-22222 88888-55555-13333-77777	-
Conjuntos identificados	0	1	2	4	0

Tabla 5.5: Análisis de posibles vinculaciones entre imputados, para causas de una misma comunidad

De la tabla se observa que en la comunidad 4 se encuentra la mayor parte de relaciones entre imputados para ese conjunto de causas. Al observar los datos de esa comunidad, estos resultados pueden darse al ser la comunidad de tamaño más grande, lo que aumenta la probabilidad de encontrar relaciones. Otro punto a destacar es que los imputados conocidos de estas causas suelen actuar en conjunto. De las 11 causas con IC de esta comunidad, 8 tienen al menos 2 imputados conocidos.

5.2. Discusión

Análisis de Resultados

Causas consideradas en un foco investigativo

Los resultados muestran que la detección de comunidades podría permitir la identificación de delitos altamente conectados entre sí, que podrían dar cuenta de un problema delictual a investigar bajo la lógica de focos investigativos propuesta por SACFI.

En el caso del Foco 1, se identifica al menos un 85 % de las causas dentro de una Comunidad. Para el caso del Foco 2, la cifra se eleva a un 92 %. Esta diferencia en el rendimiento puede estar dada por la cantidad de delitos que se considera en cada uno. Para el Foco 1, se consideran 179 delitos que pueden tener mayor variedad en la redacción de la denuncia, y, por tanto, en el modus operandi que identifica el modelo. Por otra parte, el Foco 2 solo contiene 13 causas donde la sintaxis del modus operandi puede dar cuenta de la obtención de mejores resultados.

Otra característica importante que puede justificar estos resultados es la densidad de cada comunidad, detallada en la tabla 4.4, donde la Comunidad 2 tiene el nivel de densidad más alto con un 0.647. Entendiendo que esta se compone en su gran mayoría por delitos que comparten un modus operandi, se explica la alta densidad de la red. Por otra parte, la comunidad 3 tiene un valor de 0.584 que también da cuenta de que los delitos que componen esa comunidad están altamente conectados.

En ocasiones, una comunidad podría dar cuenta de más de un problema delictivo. En tal caso, se puede volver a iterar en la detección de comunidades hasta obtener parámetros que sean coherentes con los criterios definidos por los tomadores de decisiones, observando por ejemplo la modularidad de la nueva comunidad detectada.

La identificación de focos investigativos con esta herramienta, podría servir tanto para los equipos de Analistas Criminales como para la Unidad Coordinadora SACFI, a modo de contrastar los focos investigativos que han sido declarados en cada Región, versus el panorama global de los delitos que ahí suceden.

Tiempo de Análisis de las causas

El modelo propuesto permitiría procesar, analizar e identificar patrones delictuales en alrededor de 7 horas, mientras que el proceso actual (bajo el supuesto de dedicación exclusiva a cada actividad) toma 52 horas aproximadamente. De esta forma, se lograría un ahorro de más del 80 % del tiempo. Esta cifra podría aumentar a medida que las variables que se requieren para la aplicación del modelo se guarden en una base de datos, y así se evite repetir el proceso de generación de texto.

Actualmente, la etapa de construcción de variables considera el uso de Modelos de Lenguaje de pago, que aseguran una mayor precisión en las respuestas entregadas, en comparación a otros modelos de código abierto. El cálculo del costo depende del modelo utilizado y de

los *tokens*[14] de entrada y salida que genera el modelo. El costo total asociado a esta etapa se estima en 40 USD. Sin embargo, con el auge de los LLMs, se estima que cada vez el costo asociado al procesamiento de información disminuya.

Desempeño del Modelo de optimización

Los resultados relativos al desempeño del modelo de optimización permiten observar que en 6 de los 8 casos se logra identificar causas que fueron cometidas por los mismos imputados. Al analizar los casos en los que el modelo no sugiere las causas declaradas en la tabla 5.3, se observa que es más difícil establecer la presencia de un patrón delictual en el par 65487-11111, dado que tanto el bien sustraído como el nivel de violencia empleado y el enfrentamiento con la víctima son distintos, el nivel de violencia, y la presencia de la víctima en el relato. Algo similar ocurre con el par 14444-43584, donde si bien el elemento en común es la intimidación, existen diferencias respecto al detalle de las armas utilizadas, y otros hechos que podrían afectar la similitud, como lo es la persecución en el caso 43584, que no se observa en la causa 14444.

Por otra parte, en los casos donde el modelo sí sugiere las causas esperadas, los patrones delictuales son más evidentes. En el par 92954-44444 se comparte el lugar donde ocurre el delito y la ausencia de violencia. En el par 54257-22222 existe similitud entre las armas utilizadas, el nivel de agresión, y los bienes sustraídos. Finalmente, el conjunto 98044/12222-11378 comparten el lugar del hecho, el bien sustraído y tienen una gran similitud en el nivel de violencia empleado.

Calidad de los datos

Se observa que la generación de vectores a través de la técnica de *word-embedding* es muy sensible a la redacción del texto, de tal manera que la elección de las palabras que se utilicen serán de gran relevancia a la hora de calcular similitud con otros delitos.

Con el objetivo de quitar el sesgo de la redacción de los diferentes agentes que toman las denuncias, se le pide al mismo modelo de lenguaje que genere el texto, con una estructura similar entre sí. Sin embargo, existen espacios de mejora en la redacción del *prompt* de manera tal que sea más riguroso en las palabras que se escogen. Por ejemplo, en qué circunstancias se utiliza la palabra sustracción, robo, o pérdida de un objeto.

Parámetros del modelo

El modelamiento de datos a través de la red social y el Modelo de Optimización Lineal requiere de la calibración de varios parámetros. Como una aproximación inicial, se escogen valores que hagan sentido en la identificación de patrones delictuales, sin embargo, se necesita realizar un estudio más exhaustivo de los criterios que tienen más peso en la investigación de un delito. Esto incluye el consenso de criterios para la toma de decisiones de los equipos a nivel nacional, como también un análisis de sensibilidad del modelo para entender cuáles son los parámetros que más influyen en los resultados.

Para la construcción de la variable 'valor investigativo', según el enfoque deseado, podrían valorarse más aquellas características que guarden relación con la consecución de material

[14] Grupos de caracteres que represan la unidad fundamental del texto.

probatorio de un delito (cámaras, pericias policiales realizadas, evidencia del registro GPS del bien sustraído, entre otros, ...). O bien, utilizar técnicas que recalculen similitud entre otras variables como lo son las características de delincuentes, o características de la forma en que llegan al lugar del hecho.

Limitaciones

Una de las limitaciones que puede presentar este modelo es la complejidad computacional del algoritmo. El tiempo de resolución podría aumentar considerablemente a medida que se aumente el universo de delitos a analizar, sobre todo, dependiendo de la densidad de los grafos iniciales que se entregan al solver.

Otra limitación guarda relación con una de las técnicas utilizadas para la comparación de texto. La técnica de *word-embedding* constituye un buen acercamiento, sin embargo la falta de interpretabilidad del modelo que actúa para generar los vectores podría ser complicada. Además, los modelos de lenguaje utilizados pueden tener cuestionamientos éticos, que se explican por el set de datos utilizados para el entrenamiento de estos.

Escalabilidad del modelo

El modelo propuesto podría adaptarse a otros delitos, bajo la selección de nuevas variables que sean de interés para el análisis criminal. Una vez extraídas las variables, se replica el proceso de similitud entre causas y se puede recalibrar los parámetros del modelo de optimización para que se ajuste a las necesidades de los equipos.

También, se abre una oportunidad de implementación de un sistema que ayude a la toma de decisiones en la Tramitación de Causas Menos Complejas (TCMC). Mediante la asociación temprana con otras causas similares que permitan la investigación, las tasas de Imputado Conocido podrían aumentar y de igual manera, disminuiría la cantidad de términos de causas en 'archivo provisional'.

Riesgos asociados

La utilización de técnicas basadas en modelos de lenguaje presenta riesgos y consideraciones importantes a tener en cuenta en caso de implementar el modelo. Algunos de los problemas más importantes en esta área son:

- **Generación de contenido falso o 'alucinaciones':** Los modelos de lenguaje son susceptibles de generar contenido falso en base al texto de entrada entregado. Para evitar estos problemas, existen parámetros que se pueden calibrar dentro del modelo junto con definiciones precisas en el *prompt*, a fin de evitar que responda cuando no encuentra una información.
- **Sesgo en los datos:** Es importante recalcar que los modelos utilizados han sido entrenados con grandes cantidades de datos que presentan sesgos humanos, y por tanto replican y masifican estos mismos. En este sentido, es importante evaluar constantemente los resultados entregados por los modelos y someter estos a evaluaciones éticas.
- **Privacidad de los datos:** Una gran preocupación al seno de diferentes instituciones, y aún mayor al interior del Ministerio Público, es la privacidad de los datos, con el fin

de resguardar a los sujetos involucrados en una causa, tanto víctimas como imputados. En este sentido, la aplicación de los modelos se realiza por medio de una API, que al contrario del popular 'ChatGPT', protege la privacidad de los datos en cuanto no se utilizan para reentrenar modelos y se establecen canales seguros para la transmisión de datos.

Recomendaciones

Ante una eventual implementación del modelo, se hace necesario el diseño de una metodología adecuada que permita controlar las diferentes etapas que se incluyen para la obtención de resultados, desde la selección inicial de datos hasta la interpretación de resultados.

Se destaca que para obtener mejores resultados, siempre es útil contar con buenos datos. El relato del hecho ofrece muchas posibilidades de extraer atributos y conocimientos que ayuden en la interpretación del hecho, y es por esto que debiesen existir metodologías estándar para la atención a víctimas, con el fin de recabar la mayor cantidad de detalles posible.

Es necesario tener una visión a largo plazo con el fin de definir una estructura de datos más conveniente y actualizada que permita la aplicación de las nuevas herramientas tecnológicas que surgen para el procesamiento de grandes cantidades de información.

Capítulo 6

Conclusiones

En el presente trabajo se ha construido satisfactoriamente un modelo basado en Redes Sociales y Optimización Lineal, que permite enriquecer la investigación de delitos contra la propiedad que en un inicio no cuentan con imputado conocido.

El modelo propuesto tiene una fuerte relación con otros ya desarrollados en el Ministerio Público bajo el ecosistema Fiscal Heredia, tales como el modelo StRAM y LiRAM. Estos últimos tienen como unidad de análisis los imputados de causas, e identifican la mejor asociación delictual en base a relaciones entre imputados y sus respectivos niveles de *pcg*, atributo que representa la propensión a participar en una banda criminal. El modelo que se presenta en este trabajo, en cambio, tiene como unidad de análisis la causa delictual, de forma que identifica las causas más relevantes para la investigación de un delito que no cuenta con imputado conocido, en base a la similitud entre causas y al *valor investigativo* que presenta cada una, en base a la información del relato del hecho.

El trabajo permite validar la hipótesis investigativa que sugiere la existencia de correlación entre causas similares y la participación de los mismos individuos. Esta conclusión surge al mirar los resultados que indican que, seleccionando causas que sean similares y que tengan suficientes antecedentes para investigar, se identifica la participación de los mismos individuos entre estos delitos similares.

Se concluye que este modelo permite identificar patrones delictivos a través de una red social que vincula causas del Ministerio Público a través de una medida de similitud. Si bien esta última característica tiene oportunidades de mejora, tanto en la construcción de variables como en la ponderación de cada una de estas en una medida de similitud, se observan ventajas a la hora de capturar varias dimensiones que caracterizan y vinculan delitos, con atributos que guardan relación con el autor del delito, el bien sustraído, lugar, y también con las características de la víctima.

Además, mediante las técnicas de Análisis de Redes Sociales (SNA), como lo son la descripción de Redes, cálculo de métricas como la modularidad, densidad, y la evaluación de atributos de los nodos, se extraen conocimientos que suponen interpretaciones interesantes en el ámbito del Análisis Criminal. Se concluye también que el Análisis de Redes Sociales facilita la visualización de grandes cantidades de datos. Los resultados obtenidos muestran que mediante la detección de comunidades podrían proponerse Focos Investigativos. Con es-

to, los tomadores de decisiones pueden definir los criterios necesarios para que un problema delictual se transforme efectivamente en un Foco Investigativo, además existe la posibilidad de calibrar el modelo para que las sugerencias sean más precisas.

Otra de las conclusiones que se extraen a partir del trabajo realizado, es que un Modelo de Programación Lineal resulta ser una herramienta poderosa al momento de abordar problemas complejos como lo son la selección de causas que sean similares y que al mismo tiempo aporten información relevante para continuar la investigación. La metodología necesaria para plantear un modelo adecuado, requiere también de la estandarización de criterios de selección que actualmente varían en los distintos SACFI regionales. El planteamiento de un primer modelo abre la discusión frente a los atributos que son más valiosos al momento de investigar un fenómeno delictual.

El enfoque aplicado en este trabajo utiliza como insumo principal el relato del hecho. Es por esto que la calidad de la información presente, influenciará en gran forma los resultados que este modelo pueda entregar. Si bien los relatos suelen seguir una estructura regular, existen muchas oportunidades de mejora en cuanto a las preguntas que los agentes pueden realizar a las víctimas para recopilar la mayor cantidad de información, entendiendo que muchas veces estas se acercan en estado de shock a realizar la denuncia, según el nivel de violencia empleado en el delito. Los resultados del modelo permiten observar que este tiene un mejor desempeño para casos en que los relatos se redactan de una forma similar, entendiendo esto como la estructura gramatical que se escoge para describir el delito.

Este modelo presenta ventajas a la hora de escalar su aplicación en otros fenómenos delictuales más allá de delitos contra la propiedad, por ejemplo, en fenómenos de estafas, donde el modus operandi es muy característico de la posible agrupación criminal que opera en un período de tiempo. Sin embargo, al utilizar técnicas de procesamiento masivo de los datos, es importante considerar el riesgo que presenta la herramienta en la generación de contenido falso, sesgo y privacidad de los datos. En este sentido, es importante que las personas que interactúen con el sistema comprendan de forma clara y transparente la forma en que este opera.

Finalmente, se concluye que el trabajo realizado propone una nueva forma de analizar y minar grandes cantidades de información contenidas en relatos de hechos delictuales, reduciendo los tiempos de análisis e identificando patrones no siempre reconocidos a través de un enfoque manual, mejorando así la eficiencia en las actividades que componen la persecución penal llevada a cabo por el Ministerio Público.

Capítulo 7

Trabajos futuros

Construcción de la red social

Selección inicial de causas a analizar

La aplicación de la red social se realizó sobre un conjunto de datos enfocados en un espacio de tiempo y geográfico acotado, con el objetivo de poder contrastar los resultados entregados por el Modelo desarrollado versus las herramientas de trabajo tradicionales. Sin embargo, la selección de causas a analizar puede ser crucial para la identificación de nuevos patrones que permitan continuar la investigación, por ejemplo, añadiendo causas que no necesariamente compartan la misma comuna o fechas del hecho relativas, sino que se levanten gracias a la coincidencia en otras entidades presentes en el relato del hecho.

Medida de Similitud

Se puede explorar la mejora de la medida de similitud mediante las sugerencias propuestas en este trabajo, además de la consolidación de una nueva medida de similitud que permita analizar otros fenómenos delictivos.

Análisis de Comunidades Difusas

Estudiar la factibilidad de aplicar algoritmos de clustering difuso para identificar comunidades difusas en el grafo, donde los nodos pueden pertenecer a múltiples comunidades con diferentes grados de pertenencia.

Modelo de Optimización Lineal

Calibración de parámetros

El modelo de optimización que se plantea en este trabajo resulta ser una buena aproximación a la selección de causas más relevantes para una investigación, sin embargo, existen varios parámetros que necesitan de una revisión más exhaustiva para afinar sus valores. En particular, este modelo propone la elección de un parámetro vi^{min}, que da cuenta del valor total que deben tener las sugerencias entregadas por el modelo. Sin embargo, la elección de este parámetro no es tan intuitiva, y podría estudiarse la calibración de forma automática según las características de la red, o de la causa inicial a partir de la cual se genera la solución.

Incorporar nuevas fuentes de información

El modelo propuesto presenta una oportunidad para experimentar cruces de información entre los resultados que otorgan los modelos de redes sociales desarrollados para el Ministerio Público (StRAM y LiRAM). Una forma podría ser la consulta de posibles sujetos en esta red social de causas, para luego, a partir de estos, consultar los modelos StRAM o LiRAM que puedan proponer otros individuos relacionados.

Además, la incorporación de nuevas variables como el aporte de antecedentes de víctimas posterior a la denuncia, podría enriquecer la construcción de la variable 'valor investigativo', ya que al construirse solo en base al relato del hecho, podría estar sesgado a la forma en que el agente a cargo redacta la denuncia.

Bibliografía

[1] Alvarado, P., "Por quinto mes consecutivo los chilenos se muestran como ...," 2023, https://www.ipsos.com/es-cl/por-quinto-mes-consecutivo-los-chilenos-se-muestran-como-los-mas-preocupados-por-el-crimen-y-la.

[2] González Guarda, C., "El análisis criminal en el Ministerio Público: ¿Modernización de la persecución penal o sustitución de las funciones policiales?," Ius et Praxis, vol. 28, no. 3, pp. 171–190, 2022.

[3] CEAD, Centro de Estudios y Análisis del Delito, "Estadísticas delictuales," 2023, https://cead.spd.gov.cl/estadisticas-delictuales/.

[4] Ministerio de Justicia, "Ley 19640, Ley orgánica constitucional del Ministerio Público," 1999, https://www.bcn.cl/leychile/navegar?idNorma=145437&idVersion=2010-10-08.

[5] Fiscalía de Chile, "Plan institucional anual 2023," 2023.

[6] Ministerio Público, "Organigrama del ministerio público," 2023, http://www.fiscaliadechile.cl/Fiscalia/quienes/organigrama.jsp.

[7] Ministerio Público, F. d. C., "Reglamento orgnánico de las divisiones de la fiscalía nacional," 2013.

[8] Alvear, C., González, L., y Rua, G., "Sistema de análisis criminal y focos investigativos: la experiencia del ministerio público de chile," Sistemas judiciales. Una perspectiva integral sobre la administración de justicia, pp. 71–82, 2020.

[9] Facultad de Economía y Negocios, Universidad de Chile, C. M., "Informe final: Primer estudio de evaluación del sistema de análisis criminal y focos investigativos del ministerio público, año 2018," 2019.

[10] Troncoso, F. y Weber, R., "Integrating relations and criminal background to identifying key individuals in crime networks," Decision Support Systems, vol. 139, p. 113405, 2020.

[11] Troncoso, F. y Weber, R., "A novel approach to detect associations in criminal networks," Decision Support Systems, vol. 128, p. 113159, 2020.

[12] Ministerio Público, F. d. C., "Estadísticas," 2023, http://www.fiscaliadechile.cl/Fiscalia/estadisticas/index.do.

[13] Ministerio de Justicia, "Ley 19696, establece Código Procesal Penal," 2000, https://www.bcn.cl/leychile/navegar?idNorma=176595&idParte=8646741&idVersion=.

[14] Zegarra, M., "El archivo provisional en chile: su aplicación y su problemática (2008-2015)," Nova criminis: visiones criminológicas de la justicia penal, no. 13, pp. 201–244, 2017.

[15] Troncoso, F., "Prediction of recidivism in thefts and burglaries using machine learning,"

Indian Journal of Science and Technology, vol. 13, no. 6, pp. 696–711, 2020.

[16] Zhang, X., Liu, L., Xiao, L., y Ji, J., "Comparison of machine learning algorithms for predicting crime hotspots," IEEE access, vol. 8, pp. 181302–181310, 2020.

[17] Fontalvo Herrera, T. J., Vega Hernández, M. A., y Mejía Zambrano, F., "Método de clustering e inteligencia artificial para clasificar y proyectar delitos violentos en colombia," Revista Científica General José María Córdova, vol. 21, no. 42, pp. 550–572, 2023.

[18] Sobrinho, D. S. M. y Bastos, Í. V. S., "1. información institucional," Buenas prácticas en análisis criminal, p. 57.

[19] Wasserman, S. y Faust, K., "Social network analysis: Methods and applications," 1994.

[20] Santander, P., "Desarrollo de una medida de similitud entre delitos [Tésis en curso de publicación]," 2023.

Anexo

Anexo A. Ejemplo de causas

Causas ID	Resumen
11378	Arrebato de teléfono celular desde las manos de la víctima aprovechando que el vidrio del vehículo estaba abierto.
30063	Sujeto lanzó piedra contra el vidrio del vehículo para robar teléfono celular.
33207	Sujeto se acerca a la ventana del vehículo y exige las llaves mientras la víctima bajaba a su hijo de la camioneta.
36233	Los delincuentes solicitaron un viaje mediante la aplicación 'XXXX', y al llegar a su destino, uno de ellos apuntó a la víctima con un objeto en el cuello simulando ser un arma de fuego y exigió la entrega del celular, mientras el otro sustraía dinero de la billetera.
47090	El sujeto escaló la reja perimetral de aproximadamente 50 centímetros de altura y fue sorprendido por vecinos del lugar.
47979	El delincuente identificado ofreció un servicio de traslado a través de la aplicación 'XXXX'. Tras llevar a las víctimas a una plaza, recibió un llamado y se dirigió con una de las víctimas a un lugar desconocido, donde sacó una pistola y obligó a la víctima a descender del vehículo, robando todas las especies y huyendo del lugar.
60551	Los delincuentes abordaron a la víctima mientras abría su vehículo, uno de ellos con arma de fuego lo intimidó exigiendo la entrega del dinero.
65184	Sujeto en estado de ebriedad se acerca a la ventanilla del conductor y sustrae lentes ópticos bifocales mientras la víctima esperaba al alcalde. Otro sujeto amenaza a la víctima de agresión si no retira la demanda.
66891	Los delincuentes sustrajeron herramientas y lanzaron piedras para quebrar vidrios de una retroexcavadora mientras estaba detenida en un semáforo.
69775	Ingreso al domicilio y hurto de dinero en efectivo desde un mueble.
80899	Sujetos desconocidos descienden de una camioneta y rodean a las víctimas, intimidándolas con armas de fuego y sustrayendo sus pertenencias.
83631	Ingreso al domicilio y sustracción de dinero en efectivo desde una caja metálica.
95437	Sujeto se acerca a vehículo detenido por congestión vehicular y sustrae teléfono celular de la mano de la víctima aprovechando que el vidrio estaba abajo.
98044	Sujeto aprovechó congestión vehicular para sustraer teléfono celular a través de la ventana del vehículo de la víctima.
99860	Los delincuentes descendieron de un vehículo, insultaron y amenazaron a la víctima, y tras un forcejeo le sustrajeron su billetera con dinero en efectivo.

Tabla A.1: Resumen de relatos de un set de causas.