UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

POTENTIEL DES CONCORDANCES ADN DANS LES AFFAIRES CRIMINELLES POUR GÉNÉRER DU RENSEIGNEMENT SUR LES RÉSEAUX DE CO-DÉLINQUANCE ET LES INCONNUS

THÈSE PRÉSENTÉE
COMME EXIGENCE PARTIELLE
DU DOCTORAT EN BIOLOGIE CELLULAIRE ET MOLÉCULAIRE

PAR
LÉO LAVERGNE

SEPTEMBRE 2024

Université du Québec à Trois-Rivières

Service de la bibliothèque

Avertissement

L'auteur de ce mémoire, de cette thèse ou de cet essai a autorisé l'Université du Québec à Trois-Rivières à diffuser, à des fins non lucratives, une copie de son mémoire, de sa thèse ou de son essai.

Cette diffusion n'entraîne pas une renonciation de la part de l'auteur à ses droits de propriété intellectuelle, incluant le droit d'auteur, sur ce mémoire, cette thèse ou cet essai. Notamment, la reproduction ou la publication de la totalité ou d'une partie importante de ce mémoire, de cette thèse et de son essai requiert son autorisation.

UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

BIOLOGIE CELLULAIRE ET MOLÉCULAIRE (DOCTORAT)

Direction de recherche :

Emmanuel Milot	
Prénom et nom	directeur de recherche

Jury d'évaluation

Cyril Muehlethaler	président du jury
Prénom et nom	Fonction du membre de jury
Quentin Rossy	évaluateur externe
Prénom et nom	Fonction du membre de jury
Rémi Boivin (co-direction)	membre évaluateur UdeM
Prénom et nom	Fonction du membre de jury
Maxime Bérubé	évaluateur interne UQTR
Prénom et nom	Fonction du membre de jury

À mes défunts parents qui m'ont toujours
encouragé à la curiosité.

À mes collègues du LSJML qui partage avec moi
une passion pour l'avancement des sciences
judiciaires.

REMERCIEMENTS

Quelques mots ici pour remercier les nombreuses personnes qui m'ont épaulé dans ce projet. En tout premier lieu, Emmanuel Milot, mon directeur à l'UQTR, qui a su donner au départ l'inspiration et par la suite le support nécessaire au projet. La confiance et le respect que je te portais ont été pour moi déterminants pour accéder à l'élaboration d'un projet de recherche, une aventure, alors que j'étais au seuil de la retraite. Merci aussi de ta confiance dans ce contexte particulier qui pouvait facilement être hasardeux. À feu Carlo Morselli qui lui aussi, à titre de co-directeur avait cru, aux premières heures, à la pertinence de cette étude. C'était clair au départ qu'il n'était pas des plus enthousiastes à voir débarquer un retraité d'expérience en science judiciaire, déguisé en étudiant. Il a su prendre le temps d'écouter mes arguments, jauger ma passion, pour au final accepter d'épauler le projet. Par la suite, son encouragement et son expérience des réseaux criminels ont été mis à contribution dans la mise en place du projet et tout au long de ces deux premières années avant son départ hâtif. Nul doute qu'il aurait aimé voir l'aboutissement de ce projet. À Rémi Boivin, avec qui j'ai eu de nombreuses rencontres et conversations, au fil de mes séjours précovid au CICC, et qui ont été des plus enrichissantes pour mon cheminement vers la criminologie. Rémi a aimé dès le départ l'essence du projet et moi ses connaissances de criminologue. Par la suite, comme second co-directeur, tu as su respecter le cap et poursuivre avec tes conseils comme tu l'avais fait dès mes premiers jours à l'Université de Montréal. Les encouragements, les conseils et le support sont parfois aussi venus d'autres professeurs et collaborateurs. Je pense particulièrement ici à Franck Crispino, Simon Baechler, Cyril Muehlethaler, Liv Cadola, Amy Gignac au laboratoire et Céline Van Themsche à la direction pour ce qui est de l'UQTR et à David Décary-Hétu, Frédérick Ouellet, Samuel Tanner et Maurice Cusson ainsi qu'à Geneviève Riou au secrétariat au CICC.

Les nombreux étudiants et stagiaires postdoctoraux de l'UQTR et du CICC ont aussi été par leurs conseils et leur présence des collaborateurs inspirants. Je pense en autres à Tommy Harding, Carla Aimé, Roxane Landry, Jessie Beauchemin, Audré Gareau-Léonard dans l'équipe d'Emmanuel Milot et à Anne-Marie Nolet, Silas

Nogueira de Milo, Carlotta Carbone et Amalia Campo-Delgado au CICC. Leurs conseils ont toujours été des plus utiles.

À mes anciens collègues du LSJML et plus particulièrement à Diane Séguin, Karine Fiola et Jean-François Lefebvre qui ont participé au partage et à l'organisation des données ainsi qu'à la publication de deux chapitres de cette thèse. Merci à Josée Noël et France Mailly pour leur encouragement indéfectible dès les premières heures jusqu'à ce jour.

À Patrick Jeuniaux de l'institut national de criminalistique et de criminologie de Bruxelles, pour la mise en réseau des données du LSJML ainsi qu'à Éric Chartrand de la Sûreté du Québec pour son apport des données policières, essentielles au dernier chapitre. Toujours à pied d'œuvre, souvent il me téléphonait directement à la maison pour discuter des informations qu'il devait me transmettre.

Aux membres du comité d'évaluation qui sont venus épauler mes directeurs, Quentin Rossy et Cyril Muehlethaler, merci d'avoir pris de votre temps pour apprécier cette thèse et de vos compétences pour l'enrichir de votre évaluation.

Un grand merci aussi pour l'aide financière qui m'a été octroyée par mon directeur de recherche, par l'UQTR pour la diffusion, par le CICC pour les déplacements et par l'équipe de recherche sur la délinquance en réseau (ERDR) pour la rédaction.

Et au final, toutes ces personnes, je ne les aurais jamais rencontrées si ce n'était de l'acceptation du projet par ma conjointe qui m'a permis de plonger dans cette extension de carrière pour l'avancement des sciences judiciaires. Déjà à la retraite, depuis cinq ans en 2018, elle a fait preuve d'une patience inouïe au travers de ces années qui se sont étirées bien malgré nous. Merci, ma chérie, pour ton soutien, malgré les moments difficiles pour nous deux, je suis bien heureux de t'avoir encore à mes côtés.

AVANT-PROPOS

La genèse d'un projet

La thèse que vous vous apprêtez à lire ou à consulter est le fruit d'une réflexion particulière et bien personnelle qui s'est imposée à son auteur arrivé en fin de carrière après vingt-huit années en tant que spécialiste en biologie judiciaire au Laboratoire de sciences judiciaires et de médecine légale du Québec (LSJML). Durant la deuxième moitié de ma carrière, j'ai eu le privilège, en tant qu'administrateur principal du fichier québécois de la Banque nationale de données génétiques (BNDG) pour l'identification des criminels, de pouvoir organiser la gestion des concordances ADN pour le Québec, avec bien entendu le support de plusieurs collègues qui m'ont épaulé dans cette tâche. Tous les efforts de cette équipe ont été mis à contribution pour bien gérer annuellement les milliers de concordances que le système pouvait mettre au jour, provenant principalement du LSJML, et dans une moindre mesure, des autres laboratoires judiciaires du pays. Globalement, il fallait colliger les concordances entre dossiers dévoilant du coup des crimes en séries ainsi que gérer la réception des identifications provenant des concordances avec le fichier des condamnés de la BNDG (GRC, Ottawa). Afin de bien gérer toutes ces informations, tout en tenant compte des cas complexes et du transfert de ces identifications aux administrations policières locales, il nous est venu à l'idée d'intégrer, pour chaque individu, une fiche de concordances, et ce qu'il soit connu ou inconnu, ces derniers n'étant détectés que par leurs ADN. Cette fiche individuelle, liée au système informatique de la gestion des dossiers, nous permettait de colliger un ensemble d'informations pertinentes, nous permettant d'aller plus loin que la simple gestion de ces concordances. En effet, en obtenant l'historique complet des dossiers ADN d'un individu, il devenait facile d'utiliser cette fiche pour obtenir éventuellement des statistiques générales sur nos données. Mon souvenir un peu vague me dit que nous sommes quelque part fin 2001 début 2002, la banque ayant été mise en fonction en juillet 2000, et à cette époque, j'étais bien loin de me douter que la visualisation des dossiers ordonnés dans ces fiches était pour me mettre en contact avec la science des réseaux. En effet, conséquence de la co-délinquance, certains dossiers présents dans la liste d'un individu sont aussi présents chez un second individu et un ou

d'autres dossiers de ce dernier peuvent se retrouver chez un troisième et ainsi de suite. C'est ainsi que vers le début 2014, l'idée avait fait son chemin de penser à développer une nouvelle approche de gestion des concordances pour faire plus que de l'identification au cas par cas. En effet, utiliser les données de concordances en mode réseau pour gérer ces dernières de manière globale pouvait générer davantage d'informations du type renseignement forensique.

En 2016, avec l'aval de la direction, il a été décidé que je consacrerais ma dernière année avant la retraite à la mise en place d'un système pouvant nous permettre de visualiser clairement toutes les informations des fiches en mode réseau. Pour ce faire nous pouvions compter sur l'aimable collaboration de Patrick Jeuniaux de l'Institut national de criminalistique et de criminologie de Bruxelles qui avait touché à l'aspect réseau des concordances ADN quelque temps auparavant (Jeuniaux et al. 2016). M. Jeuniaux était en fait un collègue de ma collègue administratrice de la banque de données génétique belge gérée par ce même institut que j'ai connue dans nos rencontres internationales. Comme quoi les criminels ne sont pas les seuls à avoir des réseaux! Et c'est ainsi qu'avec cette collaboration, la dernière année de ma carrière fut consacrée à un projet qui me tenait bien à cœur, bonifier la gestion routinière des concordances ADN afin d'en obtenir plus d'information. Les douze mois de travail prévu ont permis d'aboutir à un système complet permettant de bien visualiser les fiches de concordances en réseau.

Deux semaines avant de quitter, à la mi-avril 2017, un peu après avoir soufflé mes soixante bougies, j'exposais à mes collègues les résultats obtenus et le fonctionnement du système. Mais ce n'était que le début. Déjà vers novembre 2016 je voyais bien l'énorme potentiel de cette approche, et que de structurer les concordances ADN en réseau n'était pas comprendre ce que les réseaux pouvaient cacher ou apporter. Je me suis renseigné un peu et j'ai vite découvert que le monde des réseaux était évidemment toute une science en soi dont j'étais bien ignorant. Ma curiosité l'a emporté, sachant aussi que mes collègues n'auraient pas le temps d'approfondir le sujet avec un volet recherche étant constamment submergé de travail et d'urgences. Encouragé par le

professeur Emmanuel Milot de l'UQTR j'ai plongé dans ce projet avec la collaboration du professeur Carlo Morselli, à l'époque directeur du centre international de criminologie comparée (CICC) à l'Université de Montréal, malheureusement aujourd'hui décédé, emporté au sommet de sa carrière par une grave maladie. Il avait vu lui aussi le potentiel de développement du renseignement autour de ces concordances ADN incluant des individus toujours absents des études traditionnelles en criminologie.

Voilà, quelques mots sur la genèse de cette thèse qui allait m'occuper pour les cinq années à venir, en me plongeant dans la terra incognita de ces individus inconnus des services de police, intégrés dans les données ADN en mode réseau. Durant ces années j'ai parfois même eu l'impression d'être comme un archéologue travaillant sur un terrain inexploré. J'espère que vous trouverez autant de plaisir et de curiosité à lire cette thèse, que moi j'en ai eu à creuser ce monde des réseaux et de la criminologie en faisant le pont avec les données de la criminalistique.

J'écris ces mots au moment où le projet arrive bientôt à terme et bizarrement j'ai récemment reçu des documents du gouvernement fédéral concernant une certaine pension… Je ne sais trop… Il va falloir que je tire ma révérence, mais heureusement la recherche garde jeune!

RÉSUMÉ

Retracer les criminels non identifiés pour étudier leur réseau de complices ou pour effectuer des études criminologiques n'est pas une tâche courante puisque ces derniers sont absents des dossiers policiers et judiciaires. Des études récentes ont démontré que les résultats de concordances ADN conservées dans les banques de données judiciaires ont un potentiel adéquat pour développer ce champ de recherche autour des individus connus uniquement par la présence de leur ADN sur une scène de crime. Cette étude utilise les données de concordances accumulées sur 18 ans d'analyse génétique judiciaire récoltées au Laboratoire de sciences judiciaires et de médecine légale (LSJML) du Québec et comporte trois volets.

Dans la première partie, nous nous attardons à l'étude du positionnement des inconnus dans les réseaux. Par l'utilisation de quatre paramètres d'analyse des réseaux sociaux, nous démontrons que les quelque 1400 inconnus n'occupent pas de positions plus marginales que les plus de 13,000 connus présents dans les données. En fait, les inconnus supportent à eux seuls jusqu'à 18% du poids de l'intermédiarité et 46% du coefficient d'agglomération. Ainsi, les individus inconnus sont intégrés de façon assez centrale avec suffisamment de co-délinquance et cette connaissance du positionnement des inconnus dans les réseaux est une première condition à remplir pour permettre la production du renseignement dans la suite de cette étude.

Au deuxième volet, nous avons porté notre attention sur les activités criminelles de ces délinquants absents des dossiers de police qui imposent une limite à la compréhension des comportements criminels dans les études criminologiques et pour les opérations policières. En utilisant les informations des dossiers criminels associées aux concordances ADN, nous démontrons que les individus inconnus présentent des divergences comparativement à leurs complices connus. Nous avons pu évaluer l'activité criminelle des individus connus et inconnus en utilisant leurs comportements de récidive et de co-délinquance ainsi que la diversification, le niveau de gravité, les types de délits et l'intermédiarité des individus. Nous avons découvert que les l448 inconnus étudiés étaient en quelque sorte marginaux lorsque comparés aux connus. Ils

sont plutôt récidivistes solitaires, actifs dans moins de cas, avec moins de violence et plus spécialisés lorsque plus actifs. Nos résultats sont en accord avec d'autres études qui démontrent que l'activité criminelle des inconnus est conforme à l'hypothèse d'exposition stipulant que les individus demeurant inconnus des services de police sont ceux qui s'exposent le moins ou attirent le moins l'attention.

Supporté par les connaissances acquises, le troisième volet de cette étude propose une approche de production de renseignement. Les concordances ADN ont l'avantage indiscutable de nous proposer des inconnus à l'étude, mais ces individus que l'on voudrait bien identifier sont présents dans une très petite portion des délits criminels rapportés aux services de police. La production de renseignement procèdera donc par de l'ajout d'information en provenance des services de police. Au travers de quatre exemples de réseaux ADN incluant des inconnus et par l'ajout d'informations obtenues par une collaboration avec le bureau des enquêtes de la Sûreté du Québec, nous avons pu démontrer que l'utilisation d'un modèle relationnel de réseau enrichi d'un schéma de série serait adéquate pour structurer du renseignement. Notre hypothèse prend appui sur la possibilité que les inconnus ont assurément plus de délits à leur actif que ce que laissent entrevoir les résultats de concordances ADN. Ces autres délits, provenant de l'ajout d'informations policières, pourraient tout aussi bien avoir été perpétrés en co-délinquance avec le ou les mêmes complices connus et inconnus, vus dans le ou les dossiers ADN. En procédant par une analyse dynamique c.-à-d. en portant une attention particulière aux endroits et aux périodes d'activités, nous avons été en mesure de mettre en évidence des groupes de délits ayant un fort potentiel de recoupement avec l'inconnu à identifier.

Le modèle mis à l'essai dans nos exemples démontre son potentiel, mais pour cerner avec précision l'environnement de vie d'un inconnu il faut évidemment compter sur de nombreuses informations obtenues des enquêtes, informations qui n'ont pas la garantie d'être toujours au rendez-vous. Cette découverte associée à une approche d'analyse en réseau est novatrice et pourrait avoir un impact plus important que prévu sur les enquêtes et les politiques avec des implications pour le renseignement.

Mots-clés : contrevenants inconnus, données absentes, réseaux criminels, concordances ADN judiciaire, co-délinquance, hypothèse d'exposition, renseignement forensique

TABLE DES MATIÈRES

CHAPITRE II

LISTE DES TABLEAUX

Tableau **Page**

Tableau Page

Chapitre IV

Annexe A

LISTE DES FIGURES

Figure **Page**

Chapitre III

Chapitre IV

Annexe A

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

AA : Agression armée ou Agravated assault

AD : Autre désigné (délit criminel)

ADN : Acide désoxiribonucléique

APL : Average path length

AS : Agression sexuelle

BNDG : Banque nationale de données génétique

BUR : Burglary

CA : Cambriolage

CC : Canadian case

CODIS : Combined DNA Index System

COI : Convicted offender index

CSI : Crime scene index

D1, D2 : Estérase D, polymorphisme 1 et 2

d : Component density

Di : Diversification index

Di max : Maximum possible of the diversification index

Di nor : Normalized diversification index

DNA : Déoxyribonucleic acid

FBI : Federal Bureau of Investigation

FSC : Forward specialization coefficient

HO : Homicide (français et Anglais)

ID : Identification

IN : Introduction par effraction

ITR : Item theory-based response

k : Number of crime categories

K : Known

LCA : Latent class analysis

LSJML : Laboratoire de sciences judiciaires et de médecine légale

N : Number of nodes in a component

NAD : Not admissible (crime)

ND : Non désigné (délit)

NDDB : National DNA databank

Nu : Count of unknown

PCR : Réaction de polymérisation en cascade (polymerase chain reaction)

PGM1, 2 et 3 : Phospho-gluco mutase, polymorphisme 1,2 et 3

pi : Proportion of crime

PO : Dossier police

r : coefficient de corrélation

RCMP : Royal Canadian Mounted Police

RGB : Red, Green, Blue; définition de tonalité de couleur

ROB : Robbery

RS : Rare secondary (crime)

S : Seriousness score of a crime

SA : Sexual assault

Sc : Component size

SD : Standard deviation

SE : Standard error

SEC : Secondary crime

CHAPITRE I

INTRODUCTION

> *"Criminal networks are not simply social networks operating in a criminal context. The covert settings that surround them call for specific interactions and relational features within and beyond the network"* (Carlo Morselli 2009B p. 8)

1.1 Description de la problématique

1.1.1 L'analyse de substances biologiques au laboratoire judiciaire

Dans mon avant-propos j'ai expliqué les motivations bien personnelles qui m'ont conduit à l'élaboration de cette thèse. À cela, j'ajouterais que bien connaître les bases de fonctionnement d'un laboratoire de sciences judiciaires est un autre élément à bien saisir pour comprendre le développement de cette thèse. Dans ce milieu scientifique, et plus spécifiquement au Laboratoire de sciences judiciaires et de médecine légale (LSJML), où j'ai passé l'essentiel de ma carrière, des spécialistes, de divers horizons, assistent les enquêteurs et les procureurs[1] dans leur recherche de vérité au service de la justice, et ce depuis plus de cent ans au Québec.

Comme dans tout domaine scientifique, l'évolution des connaissances et leur application ont fait du chemin au fil des nombreuses années. C'est ainsi qu'en 1989 une nouvelle technologie pour l'identification des criminels fut implantée au LSJML. En effet, ce que l'on appelle aujourd'hui « le profil génétique », est en usage pour tous les

[1] L'utilisation du masculin dans le texte n'est que pour simplifier la lecture et d'ailleurs on remarquera au passage que cet usage est parfois incongru puisque dans certains milieux comme chez les procureurs de la couronne, les femmes sont majoritaires depuis plusieurs années à l'instar des femmes en sciences judiciaires au Québec.

dossiers criminels de la province pour lesquels une expertise est demandée, et ce depuis plusieurs décennies déjà. Quatre grandes raisons sont à l'origine du succès de l'utilisation de la génétique dans les délits criminels. En premier lieu, la valeur probante ou de rareté extrêmement élevée d'un profil génétique obtenue par l'utilisation de 13 loci génétiques ou plus. D'ailleurs à l'époque de l'implantation de cette nouvelle technologie, l'utilisation particulière du nom « d'empreinte génétique », intégré au titre de la première publication ayant rapport à ce sujet (Jeffreys 1985), avait frappé l'imaginaire collectif, véhiculant l'impression d'unicité des profils génétiques. Deuxièmement, l'universalité de l'ADN dans les tissus permettant de comparer du sang, du sperme, des objets touchés (ADN abandonné par desquamation) ou toute autre source de tissus humains. Troisièmement, l'évolution des technologies qui permettent d'analyser des quantités infimes de substances biologiques, comme on en retrouve dans les traces sur les scènes de crime. Finalement, la facilité d'identification de contrevenants par l'utilisation d'une banque contenant les profils génétiques d'individus condamnés, qui permet, dans de nombreux cas, de « mettre un nom » sur l'ADN d'une scène de crime. Toutefois, il reste dans les données ADN colligées des scènes de crime un ensemble non négligeable de profils génétiques d'individus qui demeurent inconnus malgré les centaines de milliers de condamnés déjà mis en banque au Canada.

Ainsi, l'introduction de l'analyse de l'ADN en sciences forensique a été en soit une assez grande révolution imposant, par son pourvoir de discrimination, un virage complet vers son utilisation et l'abandon quasi total des techniques de sérologie traditionnelle[2]. Il faut comprendre qu'avant les travaux de Jeffreys (1985), les substances biologiques retrouvées sur les scènes de crime étaient analysées en procédant à la détection des polymorphismes de diverses protéines sériques que l'on pouvait retrouver dans les traces de sang ou de sperme récoltés par les enquêteurs ou leurs services d'identité judiciaire. Depuis leur implantation dans les années 70, les différentes substances antigéniques des groupes sanguins (A, B, AB, O) ainsi que les

[2] Même si l'ADN est devenu le moteur de l'identification dans un laboratoire de biologie judiciaire, quelques techniques de détections immunologiques sont encore utilisées pour orienter les décisions à prendre dans l'approche analytique d'un délit.

polymorphismes des protéines estérase D (D1 et D2) (Hopkinson *et al*. 1973) et phospho-glucomutase (PGM1, PGM2 et PGM3) (Quick *et al*. 1974), représentaient la presque totalité des analyses disponibles, applicables à aux traces biologiques provenant des scènes de crimes. Sachant que, par exemple, le groupe sanguin A est présent à 42 % dans la population Canadienne, que l'estérase D1 peut se retrouver chez 88 % des individus et que la PGM1 chez 76 %, on comprendra que le pouvoir de discrimination est bien faible, même en regroupant ces polymorphismes (ainsi la fréquence attendue dans la population d'un individu A/D1/PGM1 est de 0,42 x 0,88 x 0,76 = 0,28). À tout le moins et, disons au mieux, quand les résultats des échantillons en provenance de la scène de crime étaient différents de ceux établis pour le suspect sous enquête, on pouvait dès lors, exclure cet individu comme étant la source des substances biologiques recueillies.

De nos jours, c'est par l'analyse directe de l'ADN en utilisant de treize à quinze loci génétiques, chacun ayant en moyenne une douzaine d'allèles différents, dont les fréquences dans une population varient entre moins de 1 % à 10-12 % et n'ayant que quelques rares cas qui approchent le 30 %, qu'on en vient à des valeurs probantes avoisinant les 10^{-10} (Weir 2007; Budowle *et al*. 2011). Ainsi, l'implantation de l'ADN dans les laboratoires de criminalistique et l'utilisation de nombreux loci ont permis; 1) d'augmenter substantiellement le pouvoir d'exclusion et 2) de doter la justice d'un moyen pour prendre, dans certains cas, une décision quant à l'identité de l'individu à la source de la trace (individualisation). D'autant plus que les techniques modernes d'analyse par amplification (PCR) ont permis l'utilisation d'une infime quantité d'ADN comme celle que l'on peut retrouver à l'état de trace sur des objets manipulés.

Toutefois, il n'en demeure pas moins que l'apport d'une nouvelle technologie, si utile soit celle-ci par son universalité ou sa capacité à distinguer des individus, n'est pas en soi un facteur de changement dans le paradigme de fonctionnement d'un laboratoire vis-à-vis des scènes de crimes sur lesquelles il lui faudra effectuer des analyses. En effet, à la base, le processus de détection et de collecte des traces n'a pas changé et se déroule comme suit : l'activité criminelle génère des traces biologiques; ces traces, lorsque

détectées et prélevées par l'analyste de la scène, sont acheminées au laboratoire pour analyse. Dans le cas d'une analyse d'ADN, si la qualité et la quantité (de l'ordre du nanogramme d'ADN ou plus) du spécimen sont au rendez-vous, le spécialiste du laboratoire pourra sans doute en extraire un profil génétique complet.

1.1.2 L'utilisation du profil génétique

Cette nouvelle et révolutionnaire capacité d'identification des individus vient avec des changements de pratiques et les conséquences qui s'ensuivirent n'ont pas été des plus évidentes au début de l'implantation de la technologie. Évidemment, dans les débuts, on n'avait d'attention que pour les résultats spectaculaires apportés aux causes criminelles; tel violeur ou tel assassin ne pouvant démentir la présence de son sperme ou de son sang sur le lieu du crime. Les enquêteurs, impressionnés, en oubliaient parfois que la partie la plus importante d'une accusation criminelle trouve sa source dans la qualité et la structure de leur enquête; les analyses, du reste, fournissent des indices parmi d'autres lorsque vient le temps de présenter une preuve à la cour.

Un des premiers effets de l'utilisation du profil génétique fut l'augmentation très importante du nombre de dossiers à traiter par le LSJML. Avant l'ère ADN, les spécialistes en biologie judiciaire du LSJML faisaient des expertises sur environ 1 500 délits criminels par an depuis le milieu des années 70, en procédant par ds analyses sérologiques et immunologiques. La plupart des cas étaient des crimes graves où des prélèvements en quantité suffisante de sang ou de sperme devaient être disponibles pour analyse.

Avec l'avènement de la génétique, toute détection d'une substance biologique dans un dossier de nature criminelle, peut susciter l'espoir d'une analyse ADN ouvrant sur une piste d'enquête, et ce même pour les cas avec de très faibles quantités d'ADN. Au LSJML, j'ai travaillé à l'implantation de l'analyse ADN et l'engouement pour celle-ci a été très grand. Ce sont, en provenance de partout au Québec, des milliers de dossiers de cambriolages et de vols qualifiés qui aboutissaient au laboratoire alors qu'auparavant,

en l'absence de suspects, ces dossiers conservaient leur statut de non résolus au poste de police. Au moment de mon départ à la retraite, le LSJML traitait un volume de dossiers avoisinant les 6 500 par année.

1.1.3 La Banque nationale de données génétiques

À l'introduction de l'analyse de profils génétiques, au début des années 90, s'ajoute le rêve, rapidement abouti dans de très nombreux pays dont le Canada, d'utiliser une banque de données génétiques pour relier les crimes en série et obtenir des identifications. Ces identifications sont obtenues par l'utilisation d'un fichier des condamnés[3] qui est alimenté par les profils génétiques des contrevenants condamnés en justice pour une infraction criminelle. Les profils génétiques retrouvés sur les scènes de crime alimentent quant à eux le fichier de criminalistique et sont comparés entre eux ainsi qu'au fichier des condamnés. Malgré tous les efforts déployés, environ 10 % des profils provenant des scènes de crime restent sans identification.

Ainsi, à partir de juillet 2000, la BNDG est en fonction et l'utilisation d'une telle banque apporte une somme considérable de concordances entre dossiers. La détection de crimes en série, survenant lorsqu'un profil génétique bien précis se retrouve dans plusieurs dossiers, sera couplée à une somme presque aussi importante d'identification obtenue par comparaison au fichier des condamnés. Dans ce dernier cas, la concordance est exprimée par le lien d'un même profil génétique entre un dossier et un individu fiché. Des telles concordances se comptaient déjà au nombre de plus de mille par an durant les années où j'étais en fonction au LSJML.

Que dire de tout ce succès ? L'utilisation de l'ADN s'est avérée si efficace au fil des années, que la technique est maintenant inscrite dans les processus juridiques et d'enquêtes comme un incontournable (Gendarmerie royale du Canada, 2018-2019).

[3] Au Canada, l'analyse de l'ADN des condamnés, la mise en banque et l'administration de leurs profils génétiques est la responsabilité de la Banque nationale de données génétiques (BNDG), un organisme fédéral sous la responsabilité de la Gendarmerie royale du Canada (GRC).

Toutefois le processus d'analyse et de gestion de la banque génétique est encore organisé au cas par cas, un peu comme les demandes d'analyse qui entrent au laboratoire les unes à la suite des autres. Un spécimen est analysé, il fournit un profil génétique, la comparaison au fichier des condamnés fournit une identification qui est transmise au service de police responsable de ce dossier. Le dossier suit son cours vers son aboutissement judiciaire. Par contre deux conséquences majeures découlent d'emblée dans ce processus au cas par cas dans la gestion des dossiers d'enquêtes et des identifications. La première, plus évidente et très utile, découle de la raison d'être de l'utilisation d'une banque de données. Les profils génétiques qui demeurent en banque sont comparés avec tous les nouveaux résultats ajoutés au fil du temps. C'est ainsi que l'on peut découvrir des séries de crimes plusieurs années après qu'ils aient été commis. Dans cet aspect de son fonctionnement, la banque de profils génétiques peut apporter par l'identification du renseignement. Ainsi, de savoir qu'une série de crimes sont reliés par ADN peut ouvrir la porte à une stratégie de mise en commun des dossiers qui à son tour permet une redéfinition de l'approche stratégique de l'enquête.

Deuxièmement, et cette conséquence n'est pas apparue très évidente de prime abord, ce sont tous les délits pour lesquels l'ADN n'a fourni aucune concordance et qui demeurent non résolus. Ainsi, au lieu de se retrouver "fermé sans solution" dans les archives du poste de police, certains criminels « dorment » maintenant sous forme de profils ADN dans la banque de données génétiques comme des « inconnus », mais que l'on peut maintenant suivre à la trace (aux sens propre et figuré!) s'ils récidivent en laissant derrière eux une substance biologique. Pour avoir la capacité d'identifier à coup sûr tous les contrevenants, il faudrait que la BNDG contienne les profils génétiques de toute la population québécoise sans exception. Ce serait très efficace, à condition de pouvoir gérer la somme d'information, mais éthiquement intenable pour les droits et libertés d'une société démocratique comme le Canada. Ainsi, au moment d'écrire ces lignes, le fichier des condamnés du Canada est composé de 615 195 individus[4] et si un

[4] Pour une mise à jour, il s'agit de consulter la section statistique du site web de la BNDG : https://www.rcmp-grc.gc.ca/fr/sciences-judiciaires/statistiques-banque-nationale-donnees-genetiques

individu a pu continuer d'échapper aux enquêteurs durant ses derniers délits, et par conséquent à une condamnation l'obligeant à être fiché, il a encore la liberté d'agir en toute impunité pour développer sa carrière criminelle.

Devant cette situation où du renseignement tactique est obtenu par la détection de séries criminelles et que les identifications obtenues de la banque sont gérées au cas par cas, les enquêteurs, lorsque la seule piste solide de leur enquête qui s'offre à eux est l'ADN, finissent par adopter une approche d'attente d'identification par la BNDG.

Pourtant, d'une certaine façon, les inconnus se sont dévoilés, ils sont là dans la banque, parfois récidivistes, parfois co-délinquants, parfois les deux et on connaît leur(s) délit(s) et les lieux où ils ont été commis. Que pourrait-on faire avec ces inconnus, parfois nombreux, en attendant une identification à la faveur d'une enquête ultérieure ? L'analyse des données génétiques en mode réseau est une approche prometteuse pour mieux exploiter les informations que recèlent une banque génétique en général, et l'occurrence de profils ADN inconnus en particulier. Mais qu'entend-on au juste par une approche réseau ? Il s'agit ici de visualiser les liens de co-délinquance d'un ensemble d'individus partageant certains délits en commun, et pour lesquels ils ont laissé une trace d'ADN. L'approche permet donc d'avoir une image de la socialisation criminelle des individus. La Figure 1 (P.9) nous en montre un exemple où les liens entre un individu et un délit nous indiquent que l'on a retrouvé l'ADN de cet individu dans ce délit. Les situations de co-délinquances s'observent lorsque deux individus ou plus sont liés au même délit. Mais comment reconstituer un tel réseau avec des résultats génétiques qui sont gérés à la pièce, au cas par cas et versés dans une banque ? Pour mieux saisir la portée de cette nouvelle approche de la gestion des données de concordance, il faut bien comprendre la structure de ces données. Pour l'illustrer, je présente dans la section suivante l'organisation des concordances ADN au LSJML, qui est la source des données ayant servi pour la présente thèse.

1.1.4 La fiche de concordance ADN et les réseaux

Au LSJML, les concordances détectées par le logiciel de comparaison de profils génétiques avec la BNDG (plateforme CODIS du FBI) (Milot *et al.* 2013) sont colligées dans une fiche électronique appelée fiche de concordance. Cette fiche est en lien avec le système informatique de la gestion d'analyse des dossiers et a été créée principalement pour résumer l'information des concordances liées aux délits criminels afin de les transmettre au fur et à mesure aux services de police et en assurer le suivi. Une nouvelle fiche est créée chaque fois qu'un minimum de deux délits criminels présente le même profil génétique, associé à un individu précis, connu ou non. Si l'information à propos de l'identité de l'individu parvient au LSJML, par l'entremise d'une analyse de mandat ADN obtenu dans le cadre d'une enquête, ou par une concordance au fichier des condamnés via la BNDG, le nom et la date de naissance sont ajoutés à la fiche. Ainsi, un individu ayant un seul délit à son actif peut avoir sa fiche uniquement s'il a été identifié comme précédemment ou encore s'il y a concordance avec des délits criminels observés par la BDNG dans d'autres provinces canadiennes, ce qui lui conférerait le minimum de deux délits créant une concordance comme mentionné plus haut. Ainsi dans chacune des fiches on retrouve la liste de tous les délits criminels d'un individu, que ce dernier soit connu ou non. D'autres informations sont aussi accessibles, telles que la date des délits, la date d'identification obtenue de la BNDG, la date de naissance des individus connus et autres informations constituant une liste élaborée de données qui seront utiles à l'élaboration de cette thèse[5].

Les profils génétiques, associés aux délits criminels que l'on retrouve dans la banque font des concordances ADN lorsqu'un même profil génétique est observé sur des traces biologiques prélevées dans des délits différents. Pour reprendre le langage technique de l'analyse des réseaux sociaux, la concordance peut être représentée comme un *lien* entre deux *nœuds* : l'individu et le délit associés à un même profil génétique. Un

[5] L'ensemble des données présentes des formats variés, parfois complexes, nécessitant une compréhension approfondie des connaissances liées à la gestion de la banque au LSJML. On se référera à l'annexe A « gestion et vérification des données » pour saisir l'utilité de chacune d'elles et les corrections à y apporter.

lien peut aussi être établi entre deux individus dont les profils génétiques sont découverts sur la scène d'un même délit. Ce dernier se retrouve dans deux fiches, signalant du coup que ces individus ont tous deux été détectés par leur ADN dans ce délit. Lorsque des délits se retrouvent à répétition dans un ensemble de fiches, ils nous informent sur la structure de la co-délinquance. En d'autres mots, le composant[6] d'un réseau social criminel se dessine autour des individus actifs en co-délinquance. Dans l'exemple de la Figure 1, on retrouve le sociogramme[7] d'un composant observé dans le réseau[8] global de l'ensemble des données qui inclut cinq individus interconnectés, dont quatre sont connus (vert foncé) et un est inconnu (rouge). Ces derniers ont été actifs dans trois types de délits criminels : agression sexuelle (AS; bleu), vol qualifié (VQ; vert pâle) et cambriolage (CA; brun). On distingue les délits effectués en solo de ceux en co-délinquance, ces derniers étant reliés au minimum à deux individus. Les délits solos, eux, ne sont liés qu'à un individu, comme dans l'exemple de l'agression sexuelle attribuée à l'inconnu dans la Figure 1.

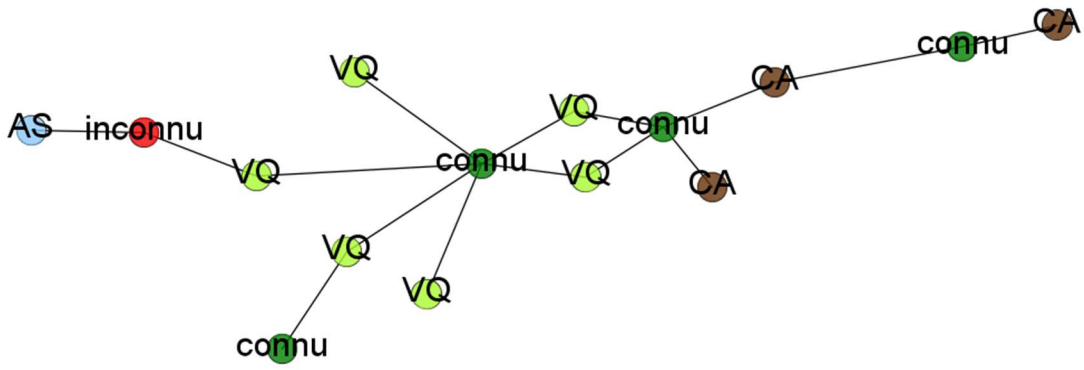


Figure 1 : Le sociogramme d'un composant ADN en structure bimodale c.-à-d. ayant deux types de nœuds, ici des individus et des délits criminels. On remarquera que deux individus connus ont été co-délinquants plus d'une fois sur autant de vols qualifiés.

---

[6] Un composant est un ensemble d'individus et de délits qui sont interreliés avec un minimum d'un lien formant un sous-groupe dans le réseau complet des données (Bichler 2019).

[7] Un sociogramme est la représentation en diagramme des liens sociaux d'une ou d'un groupe de personnes en lien avec leurs activités, ici des délits criminels.

[8] Regroupement des tous les composants obtenus à partir d'un ensemble de nœuds et des liens connus entre eux (grâce aux concordances ADN dans le cas d'un réseau ADN).

Pour étudier plusieurs aspects des réseaux sociaux criminels tels que reconstruits par l'ADN, il est possible de simplifier le sociogramme de la Figure 1 en portant l'attention directement sur les relations entre les individus. Pour ce faire, on transpose le composant bimodal en format unimodal ne retenant que les individus comme nœuds (Fig. 2), où les délits criminels sont pour ainsi dire implicites aux liens entre les individus.

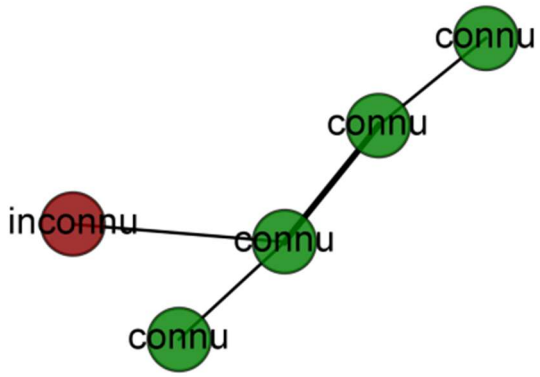


Figure 2 : Sociogramme unimodal du composant ADN de la Figure 1 ne montrant que les interrelations entre les individus. Noter l'épaisseur du lien au centre représentant les deux vols qualifiés communs aux deux individus du centre de la Fig. 1. Dans notre étude nous ne prendrons pas en compte cette force de lien qui apparaît à l'occasion dans les composants.

On comprendra à la lumière de cet exemple tout le potentiel d'information que cachent les concordances ADN analysées en réseau plutôt que dans la simple approche d'identification au cas par cas, avec laquelle les inconnus peuvent facilement tomber dans l'oubli lorsqu'on ne cherche que l'éventuelle identification avec la BNDG. Avec cette nouvelle approche, les inconnus sont intégrés dans le sociogramme avec les connus et l'étude de leurs interrelations de co-délinquances devient possible. Tout au long de notre étude, les données structurées en réseau seront notre matériau qui nous permettra de répondre à nos questions de recherche. Ces dernières visent à mieux comprendre divers aspects du comportement des inconnus afin de nous aider à les identifier (De Moor *et al.* 2017). En intégrant ces derniers dans un format réseau il nous sera possible, en premier lieu d'y appliquer diverses analyses de réseaux sociaux, pour comprendre

leurs interrelations de co-délinquances et dans un deuxième temps saisir, au travers de leurs délits, d'autres aspects de leur comportement, relevant plus de la criminologie.

L'objectif global de cette thèse était donc d'utiliser une approche d'analyse de réseau pour intégrer les individus délinquants demeurés inconnus, suite à des analyses génétiques en les intégrant à leur environnement social de co-délinquance. Des données de concordances, lorsqu'elles sont suffisamment complètes, peuvent être utilisées pour structurer cette approche analytique, mais nous verrons aussi plus loin que l'objectif final de production de renseignement autour des inconnus nécessite davantage de données qui seront obtenues des services de police pour bonifier, sous plusieurs aspects, le réseau autour des individus connus.

1.2 L'état des connaissances sur l'utilisation de l'ADN et les réseaux sociaux

1.2.1 L'approche sociologique des réseaux

Les liens sociaux entre des individus existent de toute évidence depuis la nuit des temps et ne sont pas uniques à l'espèce humaine. C'est au dix-neuvième siècle par Auguste Comte (1798-1857) que la sociologie moderne prend forme, quand ce dernier démontre que les interrelations entre individus sont à la base de la structure des sociétés, il donne à la sociologie son titre de science moderne (Freeman 2004). Une longue évolution s'en est suivie qui mène jusqu'à l'analyse des liens sociaux sous forme de sociogramme (Moreno, 1953). Moreno a introduit des outils mathématiques adaptés à la sociométrie, une approche analytique qui s'est largement répandue avec des travaux plus précis sur les divers types de calculs de centralités que l'on peut associer à des individus en réseau (Freeman 1978). À l'ère de l'informatique, l'ordinateur a permis une explosion des travaux en lien avec l'analyse des réseaux sociaux (ARS) pour laquelle de nombreux programmes dédiés ont vu le jour, tels UCINET (Borgatti *et al.* 2002), Gephi (Bastian *et al.* 2009) ou PAJEX (Batagelj *et al.* 2001), pour ne nommer que ceux-là.

La criminologie est en soi une branche de la sociologie, qui s'attarde à tous les aspects en lien aux activités illicites, aux criminels ainsi qu'aux intervenants judiciaires et juridiques qui y opèrent en y incluant aussi la victimologie. Elle emprunte à la sociologie son approche analytique axée sur l'activité et les motivations humaines, s'est enrichie de ses propres théories sociales (Sutherland 1947, Sutherland et Cressey 1960, Reiss 1951, Gottfredson et Hirschi 1990) et a évolué plus récemment en intégrant l'ARS dans sa boîte à outils, pour expliquer la formation des réseaux, leur structure et l'influence qu'ils peuvent avoir sur la criminalité et sa dynamique. L'utilisation des réseaux dans une perspective de développement de renseignement criminologique, qui a vu le jour avec Harper et Harris (1975) et qui s'est développé graduellement au fil des ans (Sparrow 1991; Papachristos 2017), a assurément été une inspiration pour le développement de l'analyse des réseaux criminels pour beaucoup de chercheurs, dont feu Carlo Morselli qui fut l'un des plus importants chercheurs de l'analyse du crime organisé dans son aspect réseau. Voici comment Klaus von Lampe termine son hommage à propos de Carlo et de sa carrière :

> *"Carlo's contributions to the study of organized crime are difficult to sum up, as extensive, diverse and profound as they are. He will be remembered as the doyen of criminal network analysis, as one of those who set new standards of academic excellence in the study of organized crime, and as one who sought to place the study of organized crime firmly in the fold of criminology and the social science."* (Von Lampe 2021)

En effet, les travaux de Morselli se trouvent dans une classe majeure, touchant avec précision de nombreux sujets, mais toujours avec cette approche des réseaux qui culmine avec son ouvrage principal, qui est aussi le plus cité : *Inside Criminal Network* (Morselli 2009B). Les réseaux ADN sont différents de ce que Carlo Morselli avait principalement étudié au fil de sa carrière, mais il n'en demeure pas moins qu'au moment de le consulter pour sonder son intérêt à participer, à titre de co-directeur à la présente thèse, il avait vu le potentiel de développement et d'avancement des connaissances que l'on pouvait obtenir de l'analyse de réseaux qui inclurait des individus inconnus. C'est ainsi que j'ai eu le privilège de côtoyer cet érudit professeur durant les deux premières années de mes travaux. Les plupart des réseaux qu'il avait étudiés étaient tous des rassemblements d'individus que l'on pourrait qualifier de

« contemporains » ces derniers étant tous actifs en même temps au sein d'une même organisation criminelle (Morselli 2009A, Morselli et Boivin 2016, Morselli et Ouellet 2018). Les réseaux ADN eux sont une représentation d'une suite de co-délinquance et d'ailleurs les études qui se concentrent sur cette dernière présentent souvent leurs résultats sous forme de réseau (McGloin et Piquero 2010). C'est un modèle de l'activité criminelle très étudié en criminologie auquel Morselli s'est aussi intéressé (Gründ et Morselli 2017), mais différent des réseaux de trafiquants de drogues ou d'armes pour ne citer que deux exemples (Morselli et Boivin 2016).

1.2.2 Les réseaux ADN

On comprendra mieux maintenant, à la lumière de l'évolution des études sur les réseaux criminels, que l'approche réseau social, utilisant la trace retrouvée sur la scène de crime, est des plus novatrices et se distingue, par une structure qui se construit au fil temps, contrairement aux réseaux d'individus œuvrant ensemble sur un projet commun en un temps déterminé, comme dans des réseaux de production et de distribution de substances illicites. La recension des écrits sur le sujet de la trace matérielle et des réseaux ne fournit que quelques publications assez récentes. Toutefois, les sujets d'analyses sont aussi diversifiés que la variété des traces qu'il est possible de trouver sur une scène de crime. Des analyses en réseau utilisant les informations obtenues de traces ont fait l'objet d'études sur les drogues illicites (Esseiva *et al.* 2007, Broséus *et al.* 2016), sur des images (Milliet 2014) et même sur des montres suisses de contrefaçon par l'analyse de la composition chimique des boîtiers et autres caractéristiques (Hochholdinger *et al.* 2019A, 2019B).

Toutefois la trace ADN n'est pas en reste et le premier montage en réseau de concordances ADN a été effectué avec des données de la banque génétique belge, qui inclut des inconnus (Jeuniaux *et al.* 2016). À partir des résultats de concordances structurées en réseau, Jeuniaux *et al.* posent trois questions de recherche. Premièrement, quelle est la prévalence des réseaux dans la Banque de Données Nationale Belge ?

Ensuite, quels sont les réseaux qui pourraient aider aux enquêtes et finalement, quels sont les crimes et les régions concernés par ces réseaux ?

L'approche est intéressante, car elle propose déjà une organisation de la pensée qui s'articule autour du renseignement forensique, un concept proposé il y a déjà quelques années par l'école de pensée de Lausanne (Ribaux et Margot 1999). Concernant leur première question de recherche, Jeuniaux *et al.* (2016) constatent que la structure des composants est très diversifiée; le nombre de composants observé décroît exponentiellement en fonction du nombre d'individus impliqués dans ceux-ci. Ces données sont tout à fait similaires à celles que nous observons dans les données du LSJML comme nous le présenterons un peu plus loin au chapitre II. Ce même genre d'observation se retrouve aussi abondamment dans d'autres études de criminologie (Morselli et Boivin 2016).

L'équipe belge aborde sa seconde question de recherche en portant une attention toute particulière aux composants incluant des individus inconnus. En effet, c'est la volonté d'identifier ces derniers qui a motivé le développement des banques de données génétiques destinées à soutenir les enquêtes criminelles. Toutefois, l'équipe de recherche belge s'en tient à démontrer la présence ou non des inconnus, ainsi que leur proportion, dans les composants. Ce faisant, elle en conclut que l'analyse des composants comportant plusieurs individus inconnus et connus en relation, peut aider au développement de stratégies d'enquêtes visant à identifier les inconnus. Autrement dit, les chercheurs n'ont pas eux-mêmes formellement utilisé l'analyse des réseaux sociaux, mais ils l'invoquent en discussion, comme une piste importante à développer afin de mieux comprendre la position des inconnus (p. ex.: sont-ils périphériques ou plutôt centraux dans la structure d'un composant ?). Ils soulèvent toutefois le risque que des contaminations ADN puissent augmenter artificiellement le nombre d'inconnus détectés et, en l'absence d'une banque d'élimination, il pourrait être hasardeux d'attribuer des valeurs de mesures d'ARS (p. ex une valeur élevée de centralité à un inconnu qui n'aurait rien à voir avec les individus interreliés qui l'entourent). Ces éléments de

réflexion au sujet de l'ARS et de l'utilisation d'une banque d'élimination sont considérés dans la présente étude, plus précisément au chapitre II.

Finalement, pour aborder leur troisième question de recherche, Jeuniaux *et al.* (2016) ont pu profiter des informations de géolocalisation des délits pour cartographier la répartition spatiale des composants. Ils ont mis au jour que de grands réseaux plutôt actifs en cambriolage étaient présents dans le nord de la Belgique tandis qu'au centre et au sud on retrouvait de grands réseaux actifs en vols qualifiés. Cette approche de renseignement est des plus intéressantes pour aider à organiser des interventions, policières ou sociales, dans la communauté.

De Moor *et al.* (2020) abordent aussi la question des inconnus dans la banque des données belge en utilisant une approche plus axée sur l'ARS. À ce jour, outre la présente étude, ce sont les seuls résultats que nous avons trouvés relatifs à la comparaison de mesures de centralité chez les individus connus et inconnus. De Moor *et al.* comparent les valeurs d'ARS obtenues sur des données dites complètes, c'est-à-dire incluant les identifications ADN couplées aux informations d'identification des dossiers de police, avec celles obtenues à partir des données de concordances ADN seules. Toutefois, ces dernières, contrairement aux premières, n'excluent pas les composants du réseau ne comptant qu'un seul individu. Une telle approche présente un défi d'interprétation majeur puisque les moyennes de centralité comparées entre elles incluent plus de 40% d'individus qui ne peuvent avoir une valeur de centralité étant des individus solo ou en duo ! Pour finir, leur analyse les amène à conclure que les individus connus présentent plus de centralité que les inconnus. Cette conclusion est à interpréter précisément dans le contexte de leur analyse globale. Dans notre approche, nous nous sommes attardés à l'analyse de composants comportant suffisamment d'individus pour obtenir des valeurs de centralité qui pouvaient être comparées en fonction des hypothèses que nous avancerons. D'ailleurs, dans leur conclusion, De Moor *et al.* (2020) proposent de n'utiliser que des composants plus grands où des valeurs de centralités pourraient être plus adéquatement comparées.

1.2.3 Les inconnus

L'analyse des inconnus dans les banques génétiques utilisées pour l'identification des criminels a attiré l'attention de certains chercheurs, et ce sans pour autant utiliser une approche basée sur les réseaux. Dans Lammers *et al.* (2012) on tente d'évaluer quels facteurs favorisent les individus ayant une propension à éviter leur arrestation et à demeurer inconnus. Pour ce faire, les auteurs utilisent le nombre de crimes effectué, leur niveau de gravité et la spécialisation obtenue par l'indice de diversification. Leur échelle de gravité des crimes est basée sur le nombre d'années qu'impose le système judiciaire hollandais pour chaque type de crime. En compilant les données en fonction des séries de crimes pour chaque individu, ils obtiennent un tableau descriptif des activités criminelles observées avec des données ADN, et ce, pour les individus connus et inconnus. On y démontre que plus les individus sont actifs, plus grandes sont les chances qu'ils finissent par être identifiés. Ceux qui se spécialisent dans un type de délit ont aussi plus de chance de passer inaperçus. À cela, ils ajoutent une analyse statistique de survie de Cox qui leur permet de prédire que 65% des contrevenants qui laissent leur ADN sur une scène de crime seront identifiés à l'intérieur d'une période de 8 ans suivant leur deuxième délit. Les 35% restants le seront plus tard ou peut-être jamais. Au sujet de la gravité des crimes, leur analyse prédit qu'elle n'a pas d'influence sur la probabilité d'arrestation. Cette conclusion semble contre-intuitive quand on sait que les crimes le plus graves attirent toute l'attention et les budgets d'enquêtes des services de police (Heller et McEwen 1973) et elle est, disons-le, contraire à la théorie de l'exposition (Ouellet et Bouchard 2017; Blumstein *et al*. 2010).

Au chapitre de la diversification, la conclusion de Lammers *et al.* (2012) est en harmonie avec les connaissances actuelles de la criminologie, à savoir que les individus spécialisés (dans leur type de délits) ont une probabilité plus faible d'être arrêtés comparativement aux individus plus versatiles (Clare 2010). Dans leur réflexion, les auteurs notent au passage que l'observation de crimes en série, détectés par l'ADN, permet aux policiers enquêteurs de modifier leur stratégie d'enquêtes. Ils ne mentionnent toutefois pas qu'une telle utilisation du renseignement pourrait être bonifiée par une analyse en réseau, comme proposé quelques années plus tard par Jeuniaux *et al.* (2016).

De plus, la présentation des résultats de la diversification est réduite à sa plus simple expression, à savoir deux moyennes générales comparées pour chacun des groupes d'individus (connus et inconnus). Pourtant, la diversification des activités criminelles d'un individu peut varier au cours de sa carrière et plusieurs facteurs pourraient en être responsables (McGloin *et al.* 2008; McGloin et Piquero 2010). De plus, le comportement des contrevenants peut être très différent selon que ceux-ci agissent en solitaires ou en co-délinquance (Haynie 2001).

Dans leur remarque finale, Lammers *et al.* (2012) abordent la question de l'incertitude autour de l'identification par ADN en spécifiant qu'il serait toujours possible qu'une concordance soit le fruit du hasard (fortuite) ou d'une contamination. À ce jour, la connaissance de la valeur probante de profil génétique complet sur de nombreux loci trouve difficilement sa place dans ces travaux et une critique concernant les éventuels profils non pertinents n'a toutefois pas été soulevée et méritait réflexion. Dans les aspects à développer pour le futur, ils abordent enfin la question de la géolocalisation, qui sera ensuite développée dans Lammers et Bernasco (2013).

Dans cette deuxième publication de Lammers, on se penche sur la question à savoir : est-ce que la dispersion géographique d'une série de délits influence la probabilité d'arrestation ? Si cette hypothèse était avérée, la dispersion des délits sur un plus grand territoire pourrait avoir pour effet de « cacher » notre individu aux efforts d'enquêtes, ce dernier bénéficierait alors d'une probabilité réduite d'arrestation. Ici, en utilisant les séries de délits observées dans la base de données génétique hollandaise, Lammers et Bernasco (2013) établissent, pour un individu et ses dossiers, un indice de dispersion qui tient compte de l'ordre temporel des délits. L'indice est évalué en mesurant la distance géographique qui sépare ces derniers selon que la séquence des délits reste dans une même zone administrative ou traverse la frontière entre diverses zones et en prenant aussi en compte s'il y a des zones intercalées, dans lesquelles aucun délit n'est enregistré pour les individus. Quoique cette approche n'implique pas un effet de causalité pour ce qui est de la distance séparant les délits, il n'en demeure pas moins que les contrevenants, qui ont des indices de dispersion plus grands, notamment

expliqués par un étalement des délits sur plusieurs juridictions, ont moins de chance d'être identifiés. Selon les auteurs, la qualité ou l'absence de coordination entre les corps policiers pourrait être une des raisons qui favoriseraient ces délinquants.

Pour apporter plus de précisions à ces conclusions, Lammers (2014) a fait une autre étude pour identifier à plus petite échelle les déplacements des contrevenants et en y ajoutant des résultats provenant d'individus inconnus (identifiés sur la seule base de leur ADN). L'analyse globale des délits confirme de nouveau que le nombre de délits commis par un individu a une influence significative sur sa probabilité d'arrestation. Puisqu'elle n'a pu établir de différences statistiquement significatives dans la corrélation entre la probabilité d'arrestation et l'indice de distance des déplacements, et ce, pour les délits associés aux individus connus et inconnus, cela leur permet de conclure qu'il n'y a pas de biais dans les données obtenues d'individus arrêtés dans l'étude précédente (Lammers et Bernasco 2013). De plus cela permet d'avancer l'hypothèse qu'un individu qui sera éventuellement identifié aurait eu tendance à se déplacer seulement ou surtout sur de courtes distances pour commettre des délits, qu'il aurait fait avec un minimum d'effort minimum. Mais il s'agit là d'un comportement que l'on retrouve aussi chez ceux qui évitent l'identification.

1.2.4 Le renseignement forensique

Cette vaste notion réfère à l'analyse stratégique, tactique et holistique d'informations recueillies par les services de police, dont les traces obtenues des scènes de crimes font partie. L'utilisation plus spécialisée des réseaux comme approche analytique est une des principales techniques pour synthétiser les informations et constituer du renseignement (Harper et Harris 1975; Sparrow 1991). Notre approche d'analyse par les réseaux utilisant la trace ADN s'inscrit donc dans une vision plus large qui consiste *« en un processus d'expansion graduelle du rôle des sciences forensiques, de son interprétation étroite à son intégration dans un style proactif de politiques et d'études de sécurité »* (Ribaux et Caneppele 2017).

Ce concept, qui rassemble des informations forensiques, associées à la trace, et policières dans un format organisé en réseau et interrelié, a été mis de l'avant à l'Université de Lausanne. Dans leurs travaux, Ribaux et Margot (1999) font la démonstration d'une production de renseignement en utilisant divers exemples de cambriolages. Leur conclusion globale est qu'il faille s'éloigner des analyses en silos (Ribaux et Caneppele 2017) et intégrer toutes les informations possibles entre les traces du laboratoire et les informations policières, tout en gardant à l'esprit le mode de fonctionnement et de raisonnement qui se développe au cours d'une enquête (Delémont *et al.* 2017).

Outre l'ADN, qui est en fait le petit dernier de la famille, les traces digitales, traces de pas et marques laissées sur les projectiles d'armes à feu, la composition et les caractéristiques des documents frauduleux sont au nombre des traces classiques pouvant être décelées sur les scènes de crime, et qui peuvent apporter une importante contribution aux enquêtes et mener à la détection de délits en séries dans un milieu donné (Ribaux et Caneppele 2017). En soi, une telle approche dans l'utilisation de la trace est présentée et argumentée comme un changement de paradigme dans l'approche des enquêtes (Ribaux *et al.* 2006; Mousseau *et al.* 2019).

En poursuivant avec les travaux d'Esseiva *et al.* (2007) sur les drogues illicites, l'école de Lausanne étend à un nouveau champ d'activité la notion de montages de réseaux et d'intégration de la trace. L'objectif ici n'est pas tant de montrer que les vendeurs et revendeurs de drogues opèrent en réseau, mais que la composition chimique des produits en circulation, incluant la concentration des principes actifs et des produits de coupage, présentant un profil chimique précis, permet d'associer d'innombrables lieux de ventes à quelques sites de production. Il faut ici toutefois saisir que dans ce contexte, les réseaux observés sont loin d'être stables puisque la composition des produits varie forcément dans le temps, et ce possiblement pour un même site de production.

Suite à de telles avancés, le concept d'utilisation de la trace judiciaire a davantage été approfondi en explorant les manières les plus efficaces de structurer l'échange et l'intégration des informations entre laboratoires et services de police, sans dénaturer l'approche des enquêtes et en prenant en compte l'évolution temporelle des activités criminelles (Ribaux *et al.* 2010A, 2010B; Kelty *et al.* 2013). Au fil du temps une approche de production de renseignement a été mise en pratique avec des informations associées au trafic de substances illégales et aux documents falsifiés (Morelato *et al.* 2013, 2014; Baechler *et al.* 2015). Ces avancées ont démontré qu'il fallait toutefois beaucoup d'information et une excellente collaboration entre les divers intervenants du milieu, des policiers enquêteurs au service d'identité judiciaire, en passant par les divers experts des laboratoires, pour instaurer une approche renseignement efficace (Rossy et Ribaux 2014; Ribaux et Wright 2014). La question du partage des données forensiques et policières a été abordée avec beaucoup de précision par Delémont *et al.* (2017), en ayant à l'esprit les difficultés auxquelles les divers intervenants pourraient faire face dans un objectif commun de production de renseignement. Ces auteurs soulignent la nécessité de former les administrateurs du milieu policier, ce qui serait un pas important à franchir pour amener les laboratoires de sciences judiciaires à un niveau de renseignement qui serait à la hauteur des capacités qu'offre la science forensique (Mousseau *et al.* 2019).

Dans la suite des exemples démontrant l'intégration possible de la trace dans le monde du renseignement, l'utilisation de L'ADN se retrouve dans les travaux de Rossy *et al.* (2013), qui intègrent ce type de données à de véritables dossiers de cambriolages en lien avec des marques de pas. Dans leurs observations et conclusions, les auteurs notent qu'un accès rapide aux données d'analyse (ADN entre autres) est un aspect important à considérer pour améliorer l'efficacité de l'approche intégrée d'enquête. De plus, ils mentionnent que l'approche par le renseignement ouvre la porte à des analyses spatio-temporelles de l'activité criminelle, qui permettraient d'évaluer des hypothèses sur les carrières criminelles, par l'incorporation de dates, lieux et autres informations.

En résumé, comme le constatent De Moor *et al.* (2017), les études utilisant les inconnus des banques de données ADN sont bien peu nombreuses, en particulier celles qui adoptent une approche axée sur l'analyse de ces inconnus intégrés en réseau. Aussi, l'évolution de l'utilisation de la trace forensique s'oriente vers la production de renseignement (Ribaux et Wright 2014), une tendance qui impose de réfléchir à comment former les intervenants participant aux enquêtes criminelles en vue de ce changement de paradigme (Kelty *et al.* 2013; Crispino *et al.* 2015; Mousseau *et al.* 2019).

1.3 Les objectifs

L'ensemble des travaux précédemment cités ont été une inspiration pour cette thèse qui veut apporter un éclairage plus approfondi sur certains des aspects énoncés plus haut. Globalement, l'objectif principal était d'utiliser les données de concordance ADN pour développer un modèle de production de renseignement afin de soutenir d'une manière concrète l'effort d'identification des contrevenants inconnus.

Plus précisément, nous devons au départ porter une attention toute particulière aux inconnus en les considérants comme le point central pour l'analyse en réseau. Ce faisant, nous avons décidé d'élaborer trois approches d'analyse permettant d'obtenir un tableau plus complet et précis de ce qui pourrait caractériser ces individus, en développant les connaissances autour de ces derniers. L'intégration de ces individus en réseau, conséquence de leurs co-délinquances, a pour but de créer un modèle novateur de production de renseignement pour les enquêtes. L'intégration des individus connus dans le réseau apporte les informations nécessaires à la création de renseignement pour tendre vers plus d'identification des inconnus. Pour ce faire, l'utilisation des informations policières est mise à contribution.

Finalement, précisons que ce projet fut développé dans l'esprit des enquêtes sur le terrain, qui se gèrent au cas par cas, selon des priorités qui ne sont pas nécessairement en lien avec des objectifs scientifiques complexes. En guise d'illustration, afin de mieux

exprimer cette orientation particulière de notre recherche, nous pourrions prendre l'exemple des travaux de Lammers *et al.* (2012), cités précédemment, dans lesquels l'analyse statistique a permis de conclure que 65% des contrevenants étaient identifiés dans les 8 ans suivant leur deuxième délit. Ce genre d'analyse statistique permet aux décideurs de porter un regard critique sur les résultats obtenus de l'utilisation des banques d'identification génétique et ils pourraient dans la foulée décider d'améliorer les enquêtes pour réduire ce temps de règlement des dossiers et notamment celui pour le 35% restant des individus non identifiés. L'étude statistique permet donc de bien évaluer la situation, mais n'apporte pas directement de solutions concrètes pour les enquêtes criminelles. Travailler à améliorer l'approche sur le terrain passe par une amélioration des connaissances relatives aux délits, aux contrevenants et aux caractéristiques propres à chaque situation. Ainsi, par une restructuration des concordances ADN en réseau, nous proposons un usage plus approfondi et complet de ces données qui, obtenues par comparaison routinière des profils génétiques, n'ont à ce jour été utilisées que pour l'identification judiciaire, selon une gestion au cas par cas. Cette nouvelle approche, que l'on propose ici, représente un avancement concret et en phase avec les concepts du renseignement forensique, développés depuis quelques années (Rossy *et al.* 2013; Ribaux et Wright 2014; Crispino *et al.* 2015; Mousseaux *et al.* 2019).

Pour notre premier objectif, au niveau des analyses à développer, nous posons l'hypothèse que pour qu'une approche réseau soit adéquate pour la production de renseignement, il faut que les individus inconnus soient aussi bien intégrés que les connus dans les composants du réseau social criminel. En effet, le potentiel de renseignement que représentent les données ADN sera vraisemblablement faible si les inconnus sont tous socialement en périphérie des composants, ou très peu interreliés aux individus connus les entourant.

Dans le deuxième objectif de cette étude nous proposons d'élargir, au-delà de l'analyse en réseau, la connaissance autour de ces individus inconnus, en utilisant les informations accessibles concernant les types de délits. Cela peut en effet aider à décrire les aspects de l'activité criminelle qui diffèrent entre les individus qui ont été arrêtés de ceux qui courent toujours. À notre connaissance, une telle étude concernant les activités

de délinquants totalement inconnus n'a pas encore été faite. Au mieux on a pu s'en approcher en examinant les délits antérieurs de délinquants fraîchement arrêtés (Ouellet et Bouchard 2017). Cependant, ce sous-groupe ne représente qu'une partie des délinquants, excluant ceux parmi les inconnus qui le demeureront toujours.

Finalement, pour le troisième objectif, nous utiliserons l'intégration de données policières sur quelques exemples de composants, en utilisant des données réelles, afin de vérifier la possibilité de mettre au jour une production de renseignement directement utile aux opérations policières. Pour ce faire, nous nous sommes adjoint la collaboration de la Sûreté du Québec pour avoir accès aux informations policières pertinentes.

1.4 De multiples questions sous-jacentes aux objectifs de recherche

À chacun des objectifs spécifiques énoncés précédemment, plusieurs questions de recherches peuvent être soulevées. Quels seraient les types de composants à utiliser pour développer les connaissances autour du positionnement de ces inconnus ? Quel est le niveau d'intégration des inconnus dans les composants ? Existe-t-il un ou des positionnements typiques ? Est-ce que les individus situés en périphéries sont obligatoirement isolés ou peuvent-ils présenter malgré tout un certain niveau d'interrelations avec leur environnement proche ? Si cette intégration locale existe, de quelle importance est-elle ? Quelles seraient les conséquences que pourrait avoir l'intégration ou non des individus en approche réseau sur une enquête ?

Par l'analyse des activités criminelles dans lesquelles les individus connus et inconnus s'engagent, nous abordons une portion de leur carrière criminelle, un sujet d'étude de prédilection en criminologie. À cet effet nous pouvons résumer la question synthèse de ce chapitre : considérant le nombre et type de délits, est-ce que les individus qui échappent aux pressions des enquêtes ont des comportements différents de ceux qui se font arrêter ? Plus spécifiquement, est-ce que la gravité des délits dans lesquels ils s'engagent est différente ? Est-ce que certains individus se spécialisent ? Est-ce que le nombre ou la gravité des délits changent selon que les individus agissent en solitaire ou

en co-délinquance ? Est-ce que ces mêmes comportements changent en fonction du nombre de complices impliqués dans les délits en co-délinquance ? Au final, devant cette multitude de paramètres à vérifier, serait-il possible de dégager une image globale qui permettrait de répondre à la question générale énoncée plus haut, à savoir si les individus connus et inconnus agissent de manière similaire ou non ?

En lien avec la production de renseignement en réseau, il faut prévoir mettre en place une collaboration avec des services de police pour l'ajout d'informations provenant des enquêtes criminelles, qui amène de nouveaux questionnements. Quelles sont les informations qu'il serait intéressant d'ajouter à la structure du réseau ADN de base pour le rendre plus complet et informatif ? Y en a-t-il à prioriser ? Jusqu'où devrait-on aller dans l'ajout d'informations supplémentaires sans alourdir le processus de production de renseignement et obtenir un bon niveau d'efficacité ? Quelle est l'influence de la dynamique du réseau, construit au travers du temps, sur l'approche du renseignement ?

1.5 La méthode

L'approche analytique choisie pour réaliser ce projet est présentée ci-dessous en trois sections correspondant aux trois objectifs de recherche.

1.5.1 L'analyse du positionnement des individus dans les composants

Pour débuter, le choix de l'ARS pour évaluer le positionnement des individus dans des composants est une approche de choix. Afin d'évaluer le niveau d'intégration des individus inconnus dans les composants par rapport aux connus, nous avons donc comparé quatre paramètres d'ARS pour deux versions des mêmes composants; l'une avec les inconnus et l'autre sans. Nous y reviendrons, mais avant, un mot sur le choix des données à utiliser pour être en cohérence avec cette approche de l'ARS.

Les paramètres ont été évalués sur les composants ayant un minimum de 5 individus, parmi lesquels au moins un inconnu. Ce faisant, nous avons obtenu, dans la

majorité des cas, des valeurs d'ARS mesurables même lorsque les inconnus étaient retranchés des composants. En utilisant 18 années de concordances ADN[9] collectées entre les années 2000 et 2019, au Laboratoire de sciences judiciaires et de médecine légale du Québec (LSJML), il s'agit de la plus grande étude de ce type à ce jour, basée sur 51 composants de ≥5 individus pour comparer les paramètres d'ARS entre individus connus et inconnus.

Les quatre paramètres d'ARS ont permis d'évaluer le positionnement relatif des inconnus, par rapport aux individus connus, dans les composants. La *centralité de degré*, une mesure qui évalue le nombre de liens qu'un individu possède avec les autres individus d'un composant. Ceux en périphérie du composant ont obligatoirement une valeur de un, et au-delà de cette valeur, la centralité de degré indique qu'un individu a plus de relations avec les autres individus. Le *coefficient d'agglomération* (en anglais "*clustering"*) nous renseigne quant à lui sur la forme de relations qu'un individu a avec deux autres individus, les trois formant un triangle fermé. Le *réseau personnel* (en anglais "*egonetwork*") nous renseigne sur le nombre de relations qui sont partagées entre les individus entourant un individu focal. Le quatrième paramètre d'ARS, celui d'*intermédiarité* (en anglais "*betweenness*"), indique quels sont les individus qui présentent une position intermédiaire, c'est-à-dire, qui sont placés sur le plus court chemin entre toutes les paires d'individus du composant. Ces quatre paramètres ont donc été mesurés pour les composants complets, c'est-à-dire incluant les inconnus, puis pour les mêmes composants sans les inconnus. La soustraction des valeurs obtenues pour les deux cas a servi à quantifier le positionnement des individus inconnus par rapport aux connus. Ainsi, plus la valeur obtenue est négative et plus les inconnus occupent, en moyenne, des positions importantes et sont bien intégrés dans la structure des composants.

[9] Les détails de la gestion de ces concordances ont été présentés précédemment (voir section 1.1.4, p.7).

1.5.2 L'analyse des activités criminelles

Les analyses ARS ayant démontré que les inconnus sont bien intégrés dans les composants du réseau social criminel, il devenait d'autant plus pertinent d'explorer leurs activités criminelles, tel que prévu en deuxième objectif de recherche. Une telle analyse des "comportements criminels" - au sens de *quel(s) type(s) de délit(s) a commis un délinquant, combien de fois, seul ou non, etc.* - et la comparaison entre les connus et les inconnus, impose ici d'opter pour une tout autre approche analytique.

Dans ce cas, pour les raisons détaillées au chapitre II, cela nous a aussi permis d'utiliser les données de concordances dans presque leur entièreté, notamment sans se limiter cette fois aux composants de plus de 5 individus. En effet, puisqu'il nous faut détailler des aspects de l'activité criminelle, tout individu agissant en solitaire et avec un minimum de deux délits, ou commettant au moins un crime avec complice(s), a pu être retenu dans l'analyse. Ainsi, seuls les individus associés à un seul délit commis en solitaire ont dû être exclus.

Sur ces données nous avons appliqué des analyses standard en criminologie pour décrire les activités criminelles (Morselli et Boivin 2016; Borgatti *et al.* 2018; Bichler 2019). Principalement, toutes les activités criminelles des individus ont été scrutées sous l'angle de la spécialisation/diversification, en utilisant l'indice de diversification (D_i)[10]. On peut en effet penser que des individus spécialisés développent une expérience avec un type de délit particulier qui les rend moins à risque d'être identifiés et interpellés par les forces policières. C'est à cette étape de l'analyse que les résultats seront confrontés à l'hypothèse d'exposition et de compétence. Par contre, il faut garder à l'esprit que d'autres facteurs peuvent aussi moduler ce risque, comme la cohésion du groupe social délinquant ou la gravité des délits (Clare 2011; McGloin et Piquero 2010). Ces éléments ne sont pas pris en compte par l'indice de diversification (Ouellet et Bouchard 2017). C'est pourquoi, dans la comparaison entre les connus et les inconnus, nous avons ajouté aux valeurs de D_i un critère de gravité des délits. Ainsi, une échelle de gravité exponentielle a permis de séparer des cas de figure assez opposés, par exemple un

[10] Pour la définition précise du calcul de cet indice vous référer chapitre 3 (3.2.3.2).

individu ayant à son actif un grand nombre de cambriolages (délit moins grave) d'un individu n'ayant qu'un seul délit, mais beaucoup plus grave, comme un homicide (Blumstein *et al.* 1988; Conrad *et al.* 2010). Cette approche a servi à estimer les niveaux d'exposition et de compétence (Ouellet et Bouchard 2017) entre ces deux groupes d'individus.

De plus, étant donné que la densité des composants est connue pour être corrélée positivement avec la spécialisation (McGloin et Piquero 2010), nous avons examiné comment les valeurs de diversification (Di) sont réparties en fonction de la densité de ces derniers. Pour cet aspect bien précis, nous reprenons les 51 composants de 5 individus ou plus identifiés au chapitre II. Il s'agit en effet de composants dont la densité est assez variable, et ce, avec et sans les inconnus, pour les besoins de l'analyse.

Toute cette approche relève d'un ensemble de questions touchant du domaine de la criminologie plutôt que de la criminalistique, mais qui peut être abordée en utilisant les informations présentes dans les données de concordances ADN du LSJML. Cela nous permet de faire le pont entre criminalistique et criminologie, en permettant à cette dernière de mieux cerner les individus qui sont absents des études criminologiques traditionnelles, basées généralement sur des données policières où tous les délinquants sont d'identité connue. Ce qui nous amène à présenter la méthode retenue pour répondre au troisième objectif, soit la production de renseignement.

1.5.3 La production de renseignement

Dans ce troisième et dernier volet de nos analyses, nous avons sélectionné quatre exemples de réseaux ADN qui ont servi de structure de base pour y ajouter des informations obtenues des dossiers de police. Ainsi, les concordances ADN, bien qu'elles ne représentent qu'une partie bien limitée des dossiers criminels connus de la police, deviennent ici le point de départ de la production de renseignement. En effet, la présence des individus inconnus, intégrés à ces données de base, étant la cible de notre intérêt d'identification. Toutefois, afin d'obtenir une vision plus complète des activités et

des relations autour de ces individus et des dossiers reliés à un composant ADN, il faut aller au-delà de cette structure de base. C'est en s'adjoignant la collaboration du service des enquêtes et des analystes de la Sûreté du Québec que nous avons eu la possibilité d'apporter les nouvelles informations associées aux individus connus présents dans les composants qui sont absentes des données de concordances du LSJML.

Ces informations ont une valeur ajoutée pour la production de renseignement. En premier lieu, il s'agit d'ajouter les individus suspectés ou complices dans les délits criminels présents dans le composant étudié, de même que tous les autres dossiers connus des services de police pour chacun des individus connus, présents à l'origine et ajoutés. Dans un deuxième temps, on ajoute les autres suspects et complices pour tous ces nouveaux dossiers de police (sans ADN). Cet assemblage beaucoup plus complet, comme on peut le voir dans les exemples proposés au chapitre IV, peut aider à déceler de nouvelles pistes d'enquêtes.

Une attention toute particulière est portée à la dynamique temporelle et spatiale de la création du composant entourant le ou les inconnus à identifier, nommément les dates et lieux des délits criminels. La détection d'une série de crimes dans un espace géographique et un temps restreints est l'un des éléments premiers qui devrait attirer l'attention de l'analyste criminel. La nature de l'information, présente dans une série de délits rapprochés, pourrait l'encourager à poursuivre, si nécessaire, le processus d'analyse par l'ajout d'informations ciblées associées aux dossiers clés entourant cette série. Ici, tout type d'information circonstancielle que l'on pourrait trouver dans un dossier d'enquête est susceptible de contribuer à établir des liens rapprochant l'analyste de l'identification d'inconnus. Il pourrait s'agir de numéro de téléphone, d'information obtenue d'une caméra de surveillance, de marques de pas ou de pneu ou d'une voiture, sa couleur, la marque, etc.

Afin de démontrer comment peut s'opérer cette production de renseignement, nous avons utilisé un modèle d'organisation schématique en sociogramme pour bien visualiser la chronologie des événements criminels (Rossy et Ribaux 2012; Rossy *et al.* (2013), grâce au logiciel libre Gephi (Bastian *et al.* 2009).

1.6 En bref…

L'ensemble de ces trois objectifs reflète la vision de ce que pourrait devenir la gestion des concordances ADN dans un modèle de renseignement de nouvelle génération. Disons en premier lieu que la gestion des concordances au cas par cas, comme approche de renseignement de "première ligne", sera toujours des plus efficaces pour rapidement transmettre des identifications et des liens de séries aux services de police. Toutefois, nos analyses démontrent qu'il existe un potentiel indéniable dans les données de concordances ADN pour développer du renseignement tactique, opérationnel, voire stratégique, autour des individus qui restent à identifier. La vision de l'école de Lausanne sur « le renseignement forensique » est donc tout à propos, s'agissant de soutenir les enquêtes policières. L'approche en réseau semble offrir un grand potentiel pour soutenir des enquêtes complexes, en apportant un angle différent et plus large, utilisant l'analyse des données liées aux inconnus. Ainsi, les concordances ADN, premier jalon de renseignement, offrent une fondation solide d'identification sur laquelle le montage d'informations supplémentaires plus tactique peut prendre forme, même si ces dernières sont parfois plus laborieuses.

Du laboratoire de criminalistique jusqu'à l'ajout de renseignement criminel au support d'enquêtes, les concordances ADN font le pont jusqu'à la criminologie par l'étude du contexte d'activités criminelles des inconnus.

CHAPITRE II

DETERMINING THE IMPACT OF UNKNOWN INDIVIDUALS IN CRIMINALITY USING NETWORK ANALYSIS OF DNA MATCHES

Léo Lavergne[a,b,*], Rémi Boivin[b,c], Simon Baechler[a,b,f,g], Patrick Jeuniaux[d], Karine Fiola[e], Diane Séguin[e], Jean-François Lefebvre[e], Emmanuel Milot[a,b,*]

a *Forensic Research Group and Département de Chimie, Biochimie et Physique, Université du Québec à Trois-Rivières, Trois-Rivières, Québec, Canada*
b *Centre International de Criminologie Comparée, Québec, Canada*
c *École de Criminologie, Université de Montréal, Montréal, Québec, Canada*
d *Institut National de Criminalistique et de Criminologie, Brussels, Belgium*
e *Laboratoire de Sciences Judiciaires et de Médecine Légale, Ministère de la Sécurité Publique, Montréal, Québec, Canada*
f *École des Sciences Criminelles, Université de Lausanne, Lausanne, Switzerland*
g *Domaine Traces et Analyse Criminelle, Police Neuchâteloise, Neuchâtel, Switzerland*

Le contenu de ce chapitre a fait l'objet d'une publication en anglais dans la revue Forensic Science International, Volume 331 février 2022; 111142. Cette revue utilise un processus d'examen par les pairs.

2.1 Contribution des auteurs

Dans cet article, l'ensemble de l'analyse des données, incluant plusieurs niveaux de vérification de ces dernières, ainsi que la rédaction des textes a été effectuée par Léo Lavergne. Seules les 1000 itérations aléatoires utilisées dans la section traitant de la distribution des inconnus ont été obtenues par un script R créé par Emmanuel Milot. Les graphiques tels que apparaissant à la Figure 4 sont le résultat d'une transposition avec R effectué par Emmanuel Milot à partir d'originaux fait par Léo Lavergne. Ce dernier ainsi que Rémi Boivin ont révisé le manuscrit jusque dans ses dernières versions. La première

version a aussi été révisée par Patrick Jeuniaux, Simon Baechler et Karine Fiola. Cette dernière ainsi que Jean-François Lefebvre ont procédé à l'anonymisation des données du LSJML. Patrick Jeuniaux a créé le script R utilisé par Karine Fiola pour organiser les données du LSJML en format réseau. Diane Séguin directrice de la section biologie au LSJML a rendu possibles tous ces travaux en acceptant de développer l'approche réseau au LSJML.

2.2 Résumé de l'article

Retracer les criminels inconnus pour étudier leur réseau de complices ou pour effectuer des études criminologiques n'est pas une tâche courante puisque ces derniers sont absents des dossiers policiers et judiciaires. Des études récentes ont démontré que les résultats de concordances ADN conservées dans les banques de données judiciaires ont un potentiel adéquat pour développer ce champ de recherche autour des individus connus uniquement par la présence de leur ADN sur une scène de crime. Cette étude utilise les données de concordances accumulées sur 18 ans d'analyse génétique judiciaire du Québec. Par l'utilisation de quatre paramètres d'analyse des réseaux sociaux, nous démontrons que les quelques 1400 inconnus n'occupent pas de positions plus marginales que les plus de 13,000 connus présents dans les données. En fait, les inconnus supportent à eux seuls jusqu'à 18% du poids de l'intermédiarité et 46% du coefficient d'agglomération. Ces résultats sont en opposition avec une étude récente et montrent clairement que les individus inconnus sont positionnés de façon assez centrale et que l'intégration de la connaissance du positionnement des inconnus dans les réseaux pourrait avoir un impact non négligeable sur les décisions à prendre lors de la gestion d'enquêtes, tout en étant aussi une première condition à remplir pour permettre la production du renseignement.

2.3 Article complet (Anglais) : Determining the impact of unknown individuals in criminality using network analysis of DNA matches.

Article history: Received 15 July 2021 Received in revised form 2 December 2021 Accepted 4 December 2021 Available online 6 December 2021

Keywords: Unknown offenders, DNA Databank matches, Criminal networks, Forensic DNA, Missing data, Forensic intelligence.

Abstract

Criminal offenders missing from police files limit the capacity to reconstruct criminal networks for criminological research and operational purposes. Recent studies show that forensic DNA databanks offer potential to address this problem, through large-scale analysis of DNA matches, many of which involve unidentified offenders. Applying social network analysis (SNA) to 18 years of DNA match data from Québec, Canada, we found that 1400 unknowns do not occupy more marginal positions in the network than 13,000 known offenders, and explain up to 18% of SNA values (e.g., betweenness centrality) for the latter while supporting 46% of their clustering values. Our results contrast with previous studies, showing moreover that unknown individuals who are positioned centrally in a network may have a larger impact than previously expected on investigation policing with implications for forensic intelligence.

> "*In essence, most studies of illicit networks focus on failed criminals, the ones for which current investigative practices are successful. The ones who manage to elude detection are the ones we need to know more about in order to disrupt their illegal activities*''. (Gallupe 2016)

1. Introduction

Three decades ago, the introduction of DNA typing for criminal identification was a revolution in forensic science. Using DNA data banking, two basic types of DNA profile comparisons became possible. First, DNA profiles uncovered from body fluids or biological material left at crime scenes could be compared to establish links between

caseworks. Investigators could thus detect authors of serial crimes and then group information about the cases involved to guide their investigations. Secondly, a comparison between DNA from crime scenes and an offender database provided investigators with ID matches, i.e., the names of potential suspects, as leads for criminal investigations. No fewer than 70 countries worldwide now maintain offender DNA databases, each one under the control of national legislation (Interpol).

Notwithstanding the legal specificities and bylaws in different countries that determine DNA database composition and use, the comparison of DNA profiles shows similarities across countries. Typically, DNA match information in the hands of an institution managing a databank (a forensic lab, a police department, a ministerial office, etc.) is communicated, after validation and in accordance with local regulations, to the police officer in charge of the specific investigation. Generally, no further analysis is done on DNA matches than those aiming at tactical intelligence, namely to perform the two aforementioned basic types of DNA profile comparisons. The main purpose of DNA databanks is to work out match data that way (Royal Canadian Mounted Police).

The large number of forensic identification reached every year attests to the success of the 'crime scene DNA – offender databank' comparative approach. Nevertheless, many individuals elude identification because the DNA they leave on crime scenes makes no match with the offender database. Hereafter we designate them as "unknowns," i.e., individuals whose genetic profile is available but not their identity. They belong to the 'dark figure' of criminality or, to paraphrase Gallupe's words, to this group of criminals who have not (yet) failed. But rather than keeping genetic information on hold until a match eventually reveals the identity of an individual, we should ask what else this information indeed tells us, which may translate into other forms of intelligence to be better able to get the hand on these unknowns (De Moor *et al.* 2017).

Actually, other uses of DNA match data have been recently proposed. For instance, Jeuniaux *et al*. (2016) used Belgian forensic DNA matches to reconstruct

criminal social networks from the association between cases and offenders. Rossy and Ribaux (2014) go further by suggesting using network analysis based on graph theory to incorporate DNA with other types of crime data. In Switzerland, for example, some cantons share forensic data on a regular basis, including DNA matches, to generate intelligence, highlighting crime series and phenomena at the scale of larger geographic areas. Rossy *et al*. (2013) compared forensic data collected between 2009 and 2011 by Swiss police departments and showed that combining information from various types of traces (DNA, shoe marks, fingermarks, earmarks, tool marks, etc.) reveals links between criminal cases that would not have been detected otherwise. For instance, combining DNA and shoe mark data increases link detection efficiency in burglary cases. Thus connecting all types of traces should increase the detection of crime series and provide a more accurate overview of criminal phenomena. Network analysis has also proven fruitful with false identity documents (Baechler and Margot 2016) and counterfeit watches (Hochholdinger *et al.* (2019).

To represent DNA matches in a network format, the molecular elements (alleles) of the genetic profiles of individuals are set aside, to instead focus on matches and administrative data relevant to construct and interpret networks (herein, case number, location, date and type of criminal event, and other casework variables associated with DNA matches are called the "administrative dataset"). This way one can avoid the difficulty of addressing bylaws associated with the direct use of molecular genetic data.

The two fundamental elements (or nodes) that can be linked together are individuals and their associated criminal cases. Therefore, a link (or edge) between an individual and a case is made when the DNA of the individual is found at the crime scene. A group of individuals, along with their criminal cases, wherein every node (individual or case) is connected to at least one other node is called a component (Bichler 2019). As we will see, this atypical way of looking at the original DNA match data provides new insights on links between offenders and cases that can support investigations, in particular by providing intelligence related to unknowns (De Moor *et al.* 2017).

Any good source of data on these unknown offenders should also prove highly valuable for criminological research, since so little information about them is typically available. Forensic DNA is one of these rare sources. Some countries have now accumulated enough DNA data from crime scenes to analyze unknown offenders from a sociological perspective, thus beyond case-per-case assessments, in particular with social network analysis (SNA). The role in criminality of unknowns who left DNA traces is underscored by the study of Jeuniaux *et al*. (2017) aforementioned, and those by Lammers and coworkers in Holland (Lammers 2014; Lammers *et al.* (2012). Using crime gravity, specialization and geographic localization, these studies compared crimes committed by known vs. unknown individuals, with the goal of understanding why the latter stayed unknown for longer periods of time. They also show "that 65% of offenders, who left their DNA at two or more crime scenes, will be arrested within 8 years" and that "the more crime an offender commits, the more probable it is that he will be arrested". By relating crime location to DNA match data, Lammers and Bernasco (2013), showed that "the probability of arrest decreases as the number of police regions in which the offender commits his/ her crimes increases". Globally these studies showed that DNA can provide reliable links between unknown offenders and crime scenes (Lammers 2014; Lammers *et al.* (2012; Lammers and Bernasco 2013; De Moor *et al*. 2018).

Lately, De Moor *et al*. (2020) merged Belgian DNA data from unknown individuals with police file data on known offenders to assess the impact of including these unknowns in criminal networks analysis. They used degree and betweenness centrality, two SNA measurements, to compare two datasets. The first includes data from police files, from which they constructed a "reduced network". The second contained the same data completed with information on unknowns collected from DNA data. The authors found that the inclusion of unknown offenders in SNA increased the detection of co- offending, but did not change much how central in the network known offenders were. Their analysis also shows that the latter tend to be more central than unknown offenders. It should be noted that De Moor *et al*. (2020) calculated SNA measurements relative to the whole network, without addressing its discontinuity into

unconnected components, namely independent subgroups of co-offenders, as well as solitary offenders who are not networked and thus have no centrality values. Given the huge weight of solitary offenders in De Moor *et al*.s' data (43% of all individuals), the extent to which their measurements reflect the networked part of criminality is not clear.

Considering the above, the objective of the present study was to quantify the impact of unknown offenders on the criminal network structure, by making explicit the separation between component types (i.e., solitary offenders, co-offenders in small, medium and large social groups) to assess how the scale of SNA may change the conclusions. We characterized a 18-year forensic DNA match dataset (knowns + unknowns) which contains all matches that have been observed across all criminal cases subjected to DNA expertise in Québec, a vast territory (> 1.5 million km2) with more than eight million inhabitants. We then compared the results to those of a restricted dataset where unknowns were removed, to mimic typical police file data. Specifically, we addressed the following questions: How integrated within the criminal network unknown individuals are? How are unknowns distributed in the various components of the network? Are they more often associated with larger, smaller, or medium-sized components? Do they operate outside of these components, working mostly in a solitary fashion? Do they occupy strategic positions? Then, SNA measurements were quantified at the level of larger components, in agreement with De Moor *et al*. (2020) who propose that "future research could focus on only the biggest component(s)…" These questions were chosen while having in mind the reality of criminal investigations. For example, finding that unknown offenders are typically peripheral in their social group could influence the effort the police will invest to continue tracking them as time passes. On the other hand, if they tend to be central, it may be worth adopting investigative tactics that make better use of the networking information, and to put more effort in trying to determine how they elude detection. We show that individual-level SNA parameters provide important insights on the contribution of unknowns to networks. For example, their inclusion can prompt a reappraisal of the presumed role of known individuals belonging to the same component.

2. Method

2.1 Data source

The DNA match dataset used in this study was obtained from the *Laboratoire de sciences judiciaires et de médecine légale* (LSJML), which is the governmental agency in charge of forensic analyses for the 31 police departments in Québec (Ministère de la sécurité publique). The National DNA Databank of Canada (NDDB) was created in 2000 and contains the genetic profiles of individuals convicted of criminal offenses (the "convicted offender index" or COI) (Royal Canadian Mounted Police 2018A). The criteria to accept DNA profiles for search against the NDDB include crime type and the complexity of DNA mixtures (Milot *et al*. (2013). A long list of criminal offenses is admissible in the NDDB (see next section) (Royal Canadian Mounted Police 2019). A "criminalistics index" (CSI) was also created for tactical intelligence; it contains DNA profiles from crime scenes and is managed at the level of local forensic labs (LSJML in Québec) as part of the NDDB. The CSI is built on the CODIS infrastructure which is design to search the genetic profiles using the allele values and various search criteria for single or mixed source of DNA (see Milot *et al*. 2013 for an overview of the structure of the NDDB)[1].

The LSJML data used in this study is composed of DNA matches generated between July 2000 and July 2018. It involves 23,066 crime-scene samples obtained from 20,804 case files. Specifically, the dataset in our hands comprises a list of individuals and associated case files in which their DNA was detected. Other variables include crime type, birth date for known individuals, police bodies involved, dates of deposition of the genetic profile in the databank, suspect/offender identification dates, as well as a few others. Sensitive data, such as names and other sample-related information was anonymized or removed by the LSJML prior to analyses. Eight crime types are represented in the data, following their designation in the Canadian Criminal Code (Canadian Criminal code): (1) homicides, including murders and attempted murders, (2)

[1] To keep updated on the Canadian NDDB the reader should also consult http://www.rcmp-grc.gc.ca/nddb-bndg/index-accueil-eng.htm.

aggravated assaults (with a weapon), (3) sexual assaults, (4) robberies, (5) burglaries, (6) crimes grouped under the label “secondary” (drug or firearms possession, arson, impaired driving, etc.), (7) “rare secondary crimes” (e.g., counterfeit money production or distribution), and (8) minor crimes not admissible for the deposition of DNA profiles in the NDDB (e.g., theft under $5000). This research received approval from the research ethic committee of Université du Québec à Trois-Rivières (certification CER-19-262-07.11).

2.2 Precisions on data origin and structure

A DNA match occurs when two genetic profiles are identical or similar enough to suggest a common source: either they both come entirely from the same unique individual, or the profile of a given person fits into a DNA mixture of ≥2 contributors, while meeting certain criteria. Matches obtained from mixtures require further verification, and those declared as “valid” are entered in the match list. Three types of matches are included in this study, two of which being generated by searching crime scene genetic profiles against the NDDB. The first type is a match with the convicted offender index, i.e., providing the identification of an individual whose DNA profile was found on a crime scene. In our data, this becomes a link (edge in SNA jargon) between a known offender and a case file (i.e., between nodes in SNA jargon). The second type is a match between a DNA trace associated with a Québec criminal case file (i.e., processed at LSJML) and a DNA trace associated with a case file from another Canadian province (i.e., processed in another forensic lab). The third type of match occurs between DNA traces from two Québec case files, leading to the detection of serial offenders. In our data, the last two types of matches become a link (edge) between two criminal case files (nodes). Finally, it should be mentioned that all the matches used in our study are generally declared using 13 or 15 genetic loci.

The structure of our data has some implications that shall be exposed here. First, it is only from searching against the COI that a match can be obtained for an offender who left his/her DNA during a single offense (i.e., who is associated with a single case file in our data). This occurs when an offender who had previously been identified by other

means than DNA was convicted for a crime, and, consequently, had to provide his/her DNA for a deposition in the NDDB: if this offender later left DNA while committing another crime, the detection of this DNA would generate a match to the offender's genetic profile already deposited in the NDDB. These solitary known offenders are numerous (6645; Fig. 1) and excluded from our analysis.

Second, from the preceding point it should be understood that an unknown individual who left DNA in a single case would not appear in our data because his/her genetic profile exists only in a single case file and nowhere else, hence has no other occurrence to match with. In addition, it must be noted that 87 seemingly solitary unknown offenders linked to a single LSJML case file are actually repeat offenders, because they are also linked to criminal cases in other Canadian provinces (e.g., individual U10 in Fig. 1A). These inter-territorial matches (represented by dotted lines in Fig. 1) were also excluded from the analyses.

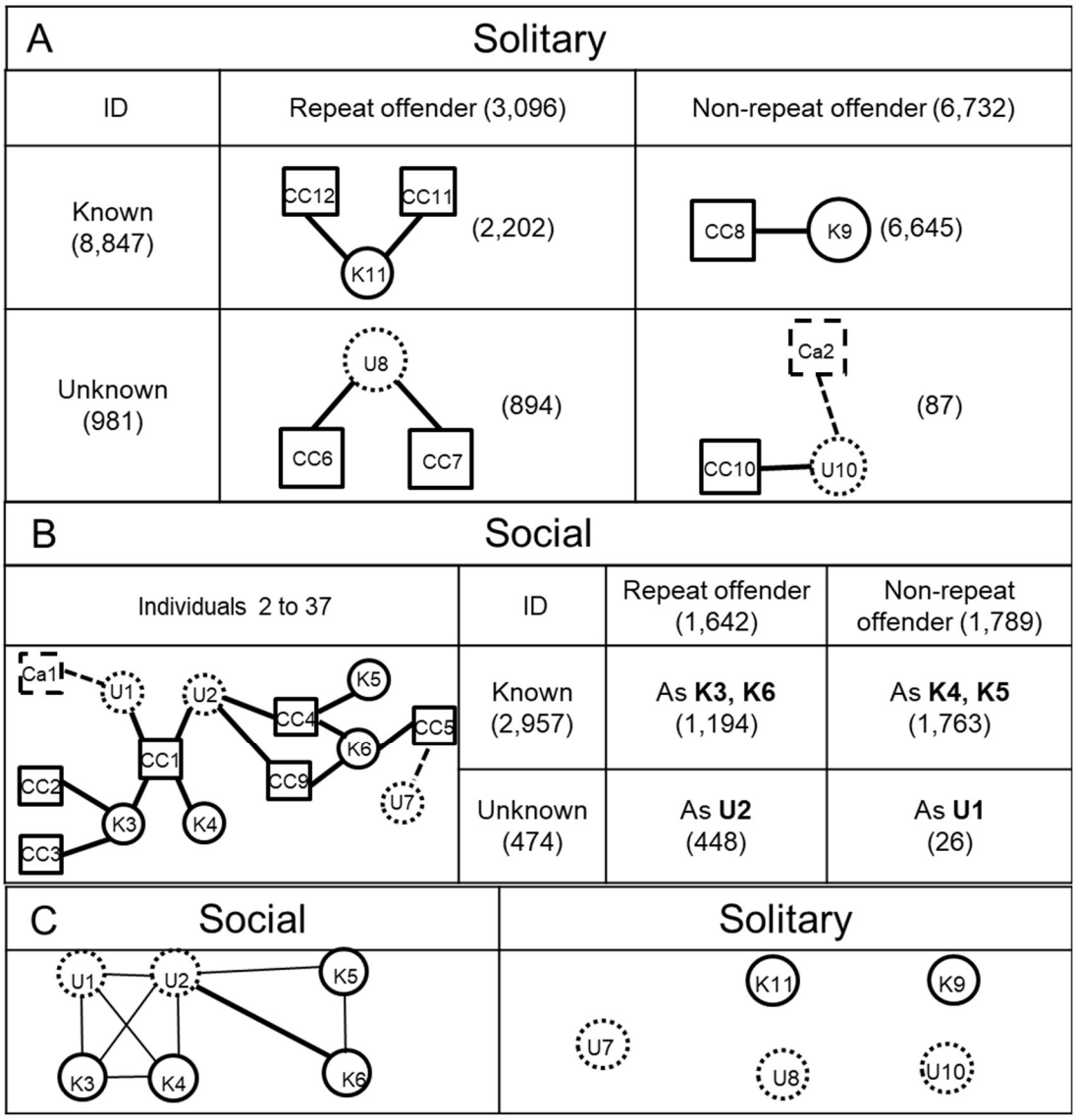


Figure 1: The distribution of known and unknown offenders in the various groups examined here. Also illustrated are examples of bimodal sociograms (social network diagrams using two types of nodes; individuals and case files), for solitary (A) and social (B) individuals (U: unknown, K: known, CC: criminal casefile, Ca: Canadian casefile). Dotted line shows links present in the DNA databank that do not exist in the administrative data used in this study, thus making individuals appear as non-repeat offenders. The number of observed cases for each category is given in brackets. Panel C shows the unimodal (individuals only) representation of the same components as in A and B. The case files are no longer visible, being embedded in the links the thicker link between unknown 2 and known 6 indicate that they share two cases together (known in graph theory as a weighted edge). All possible links occur between the individuals U1, U2, K3 and K4, who all participated in the same criminal case (CC1). This creates what is called a clique.

Social individuals occurred in components of 2–37 linked individuals. A total of 26 social individuals, which appeared as non- repeat offenders in LSJML data, were indeed linked to criminal cases from other Canadian provinces (e.g., individual U1 in Fig. 1B) and included in the analysis considering their connection to a component of ≥2 individuals. The consequences of this discrepancy between the selective inclusion of knowns and unknowns are discussed later, whenever relevant.

2.3 Network and component visualization

We reconstructed the criminal social network and its components from all edges and nodes contained in LSJML match data, using an in-house R script (R core team 2017). The network can be organized in two formats. In the bimodal format, a node is either an individual or a case file, as noted above, and edges are drawn between these two types of nodes (Fig. 1A and B). In the unimodal format, the case files are hidden to keep only individuals as nodes. Links (edges) are established when the DNA profiles of two individuals are found on the same crime scene (i.e., presumed co-offending cases; Fig. 1C).

2.4 Social network analysis measurements

We conducted social network analyses first at the whole network level, to assess general features, and secondly at the within-component level. In both cases, a comparison was done between the full vs. the restricted dataset. The first one included all individuals involved in DNA matches, while the second one excluded unknown individuals (i.e., those who left DNA traces but where no identified)[2].

[2] For more information about the SNA measurements used in this study see Appendix 1 in the Online Supplementary Material. (l'appendice est reproduit à la fin du chapitre II)

2.4.1 Network level analysis

The general properties of the criminal network were described by nine SNA measurements provided by the Gephi software (Bastian, Heyman, and Jacomy 2009) (degree, weighted degree, diameter, modularity, component number, clustering coefficient, number of triangles, average path length and density). These measurements were calculated for the full and restricted datasets. In addition, we generated 30 "random restricted datasets" by randomly removing from the full dataset a number of individuals corresponding to the total number of unknowns. This served to assess whether the differences observed between the full and restricted datasets were specifically due to unknowns or rather merely reflecting the removal of individuals independently of their status (known or unknown).

2.4.2 Within-component level analysis

To assess the integration of the known and unknown individuals we used four specific SNA measurements calculated in Gephi (Bastian *et al*. 2009). The *degree centrality* is the number of edges through which an individual is connected to others (Bichler 2019). It was used to assess whether the level of co-offending activity differed between knowns and unknowns, the minimum value of 1 indicating a peripheral individual, linked to only one other individual. The *betweenness centrality* measures how often a given individual falls on the shortest path between two others. It has been interpreted by Borgatti *et al*. (2018) as the potential for controlling information flow through a network (viewed as brokerage over multiple aspects by Morselli (2009), considering that the greater the betweenness the more control an individual has on the flow of information. Note that since the network is reconstructed from 18 years of longitudinal data, the betweenness centrality may not reflect the *current* social network of an offender. Actually, an individual could accumulate a large collection of co-offenders over his/her criminal career without necessarily interacting with more than one or a few of them at any given point in time. Since degree and betweenness centralities scale with network component size, we normalized their value as a function of the number of nodes in the component to which an individual belonged, following Freeman

(1978). For both parameters, normalized values range between 0 and 1. The third metric is based on triangles, defined as sets of fully interconnected triads of nodes (individuals) in components having ≥3 nodes. Thus, a triangle has three edges connecting three nodes. The maximum number of triangles for N nodes = N!/(6x(N-3)!). Thus, a 3-node component has 0 or 1 triangle, a 4-node component has between 0 and 4 triangles, a 5-node component could have between 0 and 10 triangles and so forth. The *clustering coefficient* is the proportion of observed triangles relative to the maximum number possible, which is achieved in a saturated network where all nodes are linked to each other. This measurement was used to assess how isolated or interconnected are individuals lying at the periphery of a network component. The fourth metric, the *egonet density*, qualifies the cohesiveness of the ego's neighborhood, in other words, how well connected to each other are the immediate neighbors of a focal individual. It is obtained by dividing the number of links between those neighbors by the total number of possible links (Bichler 2019; Borgatti *et al*. 2018). A zero value indicates that all the neighbors of a focal individual are connected only to him, while a value of one indicates that all these neighbors are interconnected with each other. Egonet density is assumed to correlate positively with the flow of information among neighbors (Bichler 2019), which in turn puts pressure on the ego to present a more consistent image of himself (Borgatti *et al*. 2018).

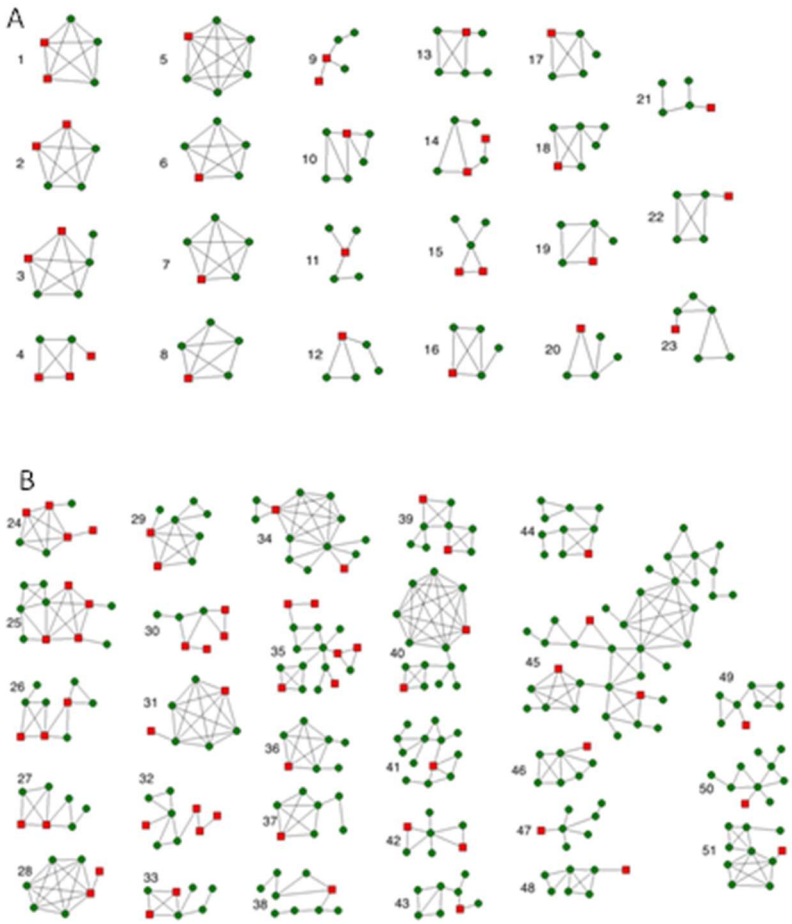


Figure 2: Sociograms of the 51 components used for the social network analysis (results in Fig. 4). A: components with 5 and 6 individuals; B: components with ε7 individuals. Green dots represent known individuals and red squares unknown ones. Sociograms layouts were drawn in UCINET (Borgatti, Everett, and Freeman 2002).

Note that the betweenness centrality and the clustering coefficient are not applicable to components of one or two individuals. For 3- or 4-individual components, the removal of eventual unknowns would cause most components to become too small as well to apply these measurements. Therefore, we calculated SNA metrics only for the larger components of ≥ 5 individuals, and with at least one unknown (N = 51; Fig. 2).

Finally, we evaluated the *centralization* for the two centrality measurements, which gives the average tendency of individuals of a given component to be more or less connected (centralization of degree) or central (centralization of betweenness) relative to all others in the network (Freeman 1978). Thus, centralization is relative to the whole

network including individuals in other components, while centrality is relative to individuals within a single component.

Comparing mean metric values for knowns vs. unknowns would not be well adapted to data having a power law distribution (Fig. 3). Therefore, normalized SNA measurements, denoted here by the general symbol k, were summed over the known individuals, both for the full (Σkf) and restricted (Σkr) datasets. The difference $\Delta k = \Sigma kr - \Sigma kf$, will be negative when the inclusion of unknowns contributes to increases metric values for the knowns.

2.5 Sensitivity analysis

Considering the cumulative and dynamic nature of the data, we conducted a sensitivity analysis. We calculated the degree, betweenness centrality and the clustering coefficient for subsets of the data limited to shorter periods: 2005–2011 and 2012–2018. This allowed us to verify whether our conclusions were consistent over network evolution time and to refine our interpretations when needed.

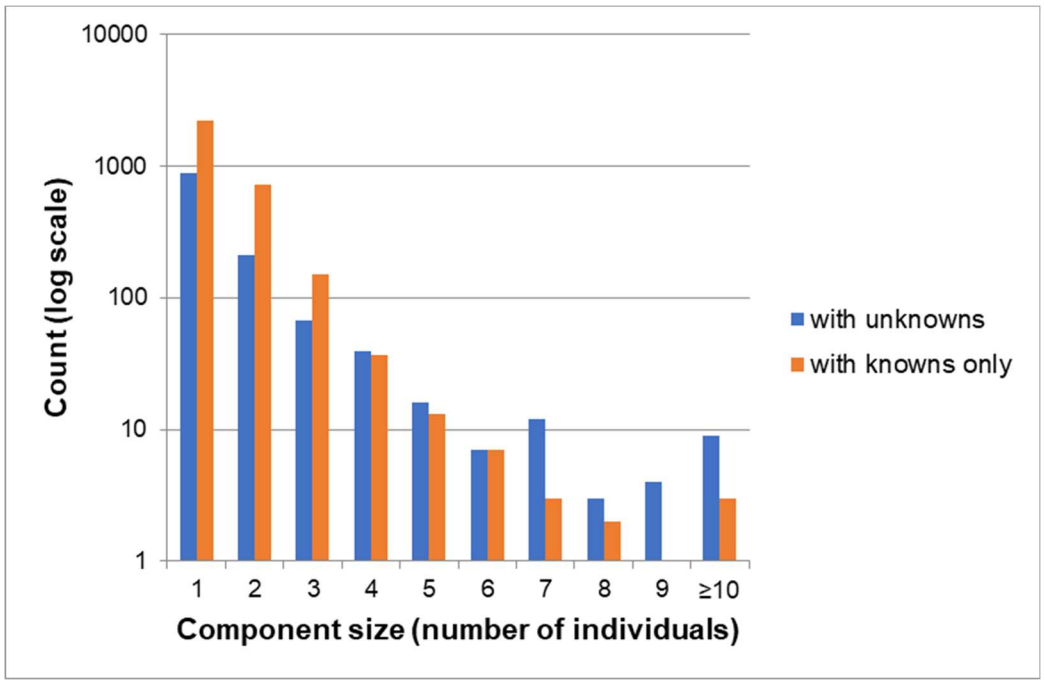


Figure 3: Distribution of components as a function of their size and inclusion or not of unknowns. Note the logarithmic scale of the Y-axis.

3 Results

3.1 Frequency of offender types

Fig. 1 shows how the 13,259 individuals in the data are distributed among the different categories: known vs. unknown, repeat vs. non-repeat, and social vs. solitary offenders. Solitary offenders make up 75% of the 11,804 known individuals. Among those, solitary known offenders with a single offense dominate (n = 6645; Fig. 1A); they represent 56% of the 11,804 known individuals, and 75% of 8847 solitary known ones. As for the 1455 unknown individuals, 67% of them were detected as solitary offenders. The lower proportion of non-repeat offenders among these solitary unknowns (9% compared to the 75% for knowns) is expected and mainly explained by the collection of data for those individuals having a single offense (as explained in Methods).

3.2 Network topology

The full dataset is composed of 23,066 edges organized into 11,146 components, a majority of which (9828) being represented by solitary actors (Fig. 1). Consequently, the whole network is quite fragmented and has a low density of 0.3%. The density when including only repeat offenders remains low at 0.8%. Component size ranges from one to 37 interconnected individuals (the full criminal network for components with ≥3 individuals including at least one unknown is illustrated in Fig. S1). Each component is associated with a number of criminal cases ranging from 1 to 78. A proportion of 7% of all edges were between two individuals sharing more than one crime. Note that graphs in our analyses are non-directed and edges unweighted by the number of shared crimes.

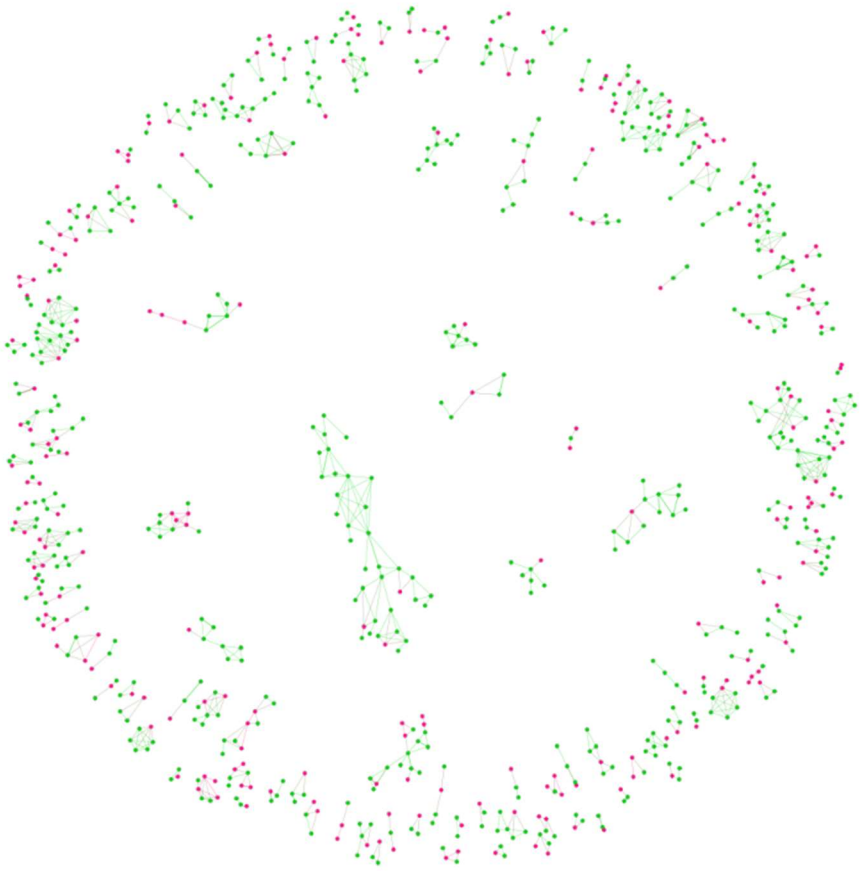


Figure S1: The 157 components with a minimum of three individuals including at least one unknown (red dots). Graph layouts were drawn using Hu's spatial algorithm (Hu 2005) using Gephi's default parameters (Bastian *et al*. 2009). In Hu's algorithm, the larger components appears central while the numerous small ones are peripheral.

3.3 Distribution of unknown offenders in components

Table 1 shows the distribution of unknown individuals in components. Let's recall that our analyses focus on *repeat* offenders, i.e. those who committed at least two crimes, for reasons explained in Method. Knowns represent 79% (5159) of individuals, while 1368 unknowns account for the remaining 21%, for a total of 6527 repeat offenders that, altogether, are included in 4414 components. The smallest components are those made of a single (i.e. solitary) offender, and the largest one is composed of 37 individuals, as aforementioned. A total of 1264 (29%) components include at least one unknown, and 370 of those 1264 components include two or more individuals (i.e. 213 small, 106 medium and 51 large components; Table 1). Table S1 provides a more detailed distribution of components according to the number of unknowns they contain while Fig. 2 shows sociograms of the 51 larger components used in SNA analysis.

The proportion of components with unknowns increases with component size: 27% for the small components, 36% for the medium components and 64% for the large ones (Table 1). Globally, a majority of components with ≥ 4 individuals contain unknowns while the reverse is true for smaller components (Fig. 3, Table 1).

Table 1: Distribution of unknown individuals in network components composed of social and solitary repeat offenders. The 6,732 solitary non-repeat offenders are excluded from the analysis (see text and Fig. 1).

Component group	Component size	Number of Individuals	Number of Unknowns	%	Number of Components	Number of components with unknown(s)	%
Small	1+2	**4984**	**1137**	**23 %**	**4040**	**1107**	**27%**
	1	3096	894	29%	3096	894	29%
	2	1888	243	13%	944	213	23%
Medium	3+4	**958**	**142**	**15 %**	**294**	**106**	**36%**
	3	654	85	13%	218	67	31%
	4	304	57	19%	76	39	51%
Large	**≥5**	**585**	**89**	**15 %**	**80**	***51***	**64%**
	5	145	22	15%	29	16	55%
	6	84	9	11%	14	7	50%
	7	105	23	22%	15	12	80%
	8	40	4	10%	5	3	60%
	9	45	10	22%	5	4	80%
	≥10	166	21	12%	12	9	75%
	All	**6527**	**1368**	**21 %**	**4414**	**1264**	**29%**

Table S1: Distribution of the 4,414 components composed of repeat offenders, according to the count of unknowns (N_u) and the component size (S_c). When $N_u = S_c$, (green cells) the component would simply not be detected without DNA information. Thus, 926 components would disappear if unknowns were not accounted for. In addition, removing DNA information for unknowns would result in a set of 201 unconnected solitary individuals for components of size $N_u = S_c$, + 1 (red cells) while they are indeed multi-offender components. Likewise, orange and dark grey cells correspond to components that would be reduced to dyads (62) and trios (31), respectively. In the large component group (blue cells), the removal of unknowns would result in various patterns (see examples in Figure 6).

Component	**Unknowns count (N_u)**						
Size (S_c)	**0**	**1**	**2**	**3**	**4**	**5**	**6**
1	2202	894	NA	NA	NA	NA	NA
2	731	183	30	NA	NA	NA	NA
3	151	51	14	2	NA	NA	NA
4	37	25	10	4	0	NA	NA
5	13	11	4	1	0	0	NA
6	7	5	2	0	0	0	0
7	3	5	5	0	2	0	0
8	2	2	1	0	0	0	0
9	1	1	1	1	1	0	0
10	2	3	0	0	0	0	0
11	1	0	0	0	1	0	0
13	0	1	1	0	0	0	0
14	0	0	1	0	0	0	0
17	0	0	0	0	0	0	1
37	0	0	0	1	0	0	0

By applying the global proportion of unknowns (21%) to the counts of individuals in the large component group, we would expect 585 × 21% = 123 unknowns if they

were randomly distributed across components of all sizes, while only 89 are actually observed. This difference is statistically significant, as shown by the simulation of random distributions of unknowns among components (Fig. S2). As visible in the top panel of Fig. S2, the 84 repeat offenders detected among the 89 unknowns in the larger components fits with the expected average (85). The total of 364 repeat unknown offenders detected for the medium size components is higher than the value expected under the hypothesis that knowns and unknowns are randomly distributed among components. Specifically, the average from 1000 permutations is 246 and 100% of these permutations resulted in values lower than 364. The surplus of unknown repeat offenders in medium components, relative to expectations, is "retrieved" from small size components. While the observed number of unknown offenders in small components is 894, 100% of permutations predict a higher number (average=1011). Permutation tests thus provide clear statistical support for the hypothesis of a non- random distribution of unknowns across component-size categories. In a second test using all the dataset (n = 13,259), hence including non-repeat and solitary offenders, the 89 unknowns from the large components are 41% higher than the expected value of 63 (Fig. S2, bottom panel).

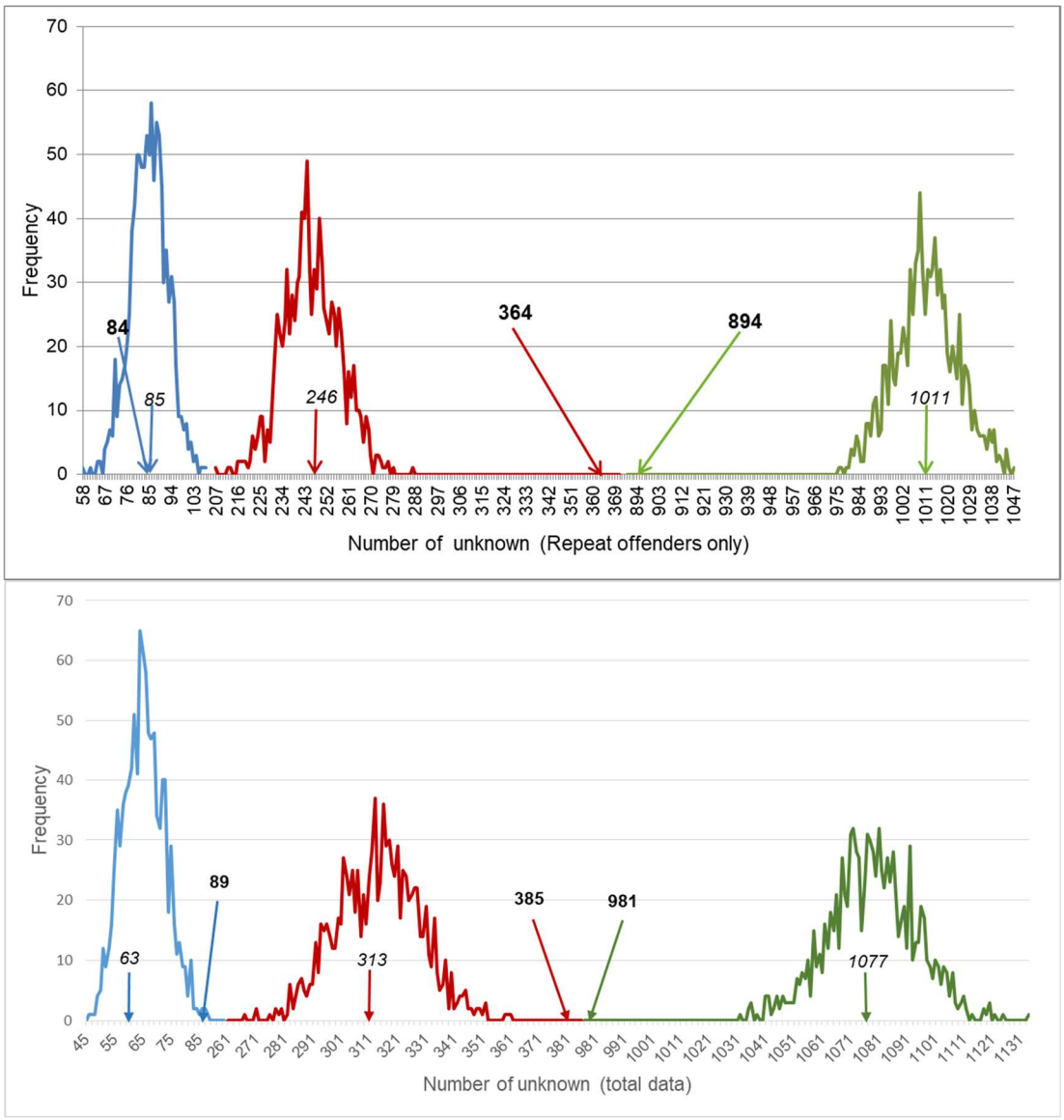


Figure S2: Distribution of the expected unknowns based on 1,000 simulations. The coloured lines show the values for; large (blue), medium (red), and small (green) components, under the hypothesis that they are randomly distributed among components. Numbers in black and bold indicate the values observed in LSJML data while those in italic indicate the average from simulations. Top panel: repeat offenders only; bottom panel: all individuals in the dataset.

3.4 Social network analysis

3.4.1 Network level

Basic network characteristics of the full and restricted datasets, and limited to the 51 large components, are summarized in Table 2, as well as those for the 30 randomized sets.

Table 2: SNA metrics for the full and restricted datasets with components having ≥5 individuals, compared to average values for the 30 datasets simulated by randomly removing 89 individuals.

Metric	Full dataset (knowns + unknowns)	Restricted dataset (knowns only, i.e. minus 89 unknowns)	Simulated datasets
Nodes	401	312	312
Edges	608	373	370.20 (±1.2)
Degree	3.03	2.39	2.37 (±.01)
Weighted degree	3.25	2.62	2.54 (±.01)
Density	.008	.008	.0077 ($\pm 8.2 \times 10^{-5}$)
Diameter	9	9	7.10 (±0.24)
Modularity	0.97	0.96	0.968 (±.0008)
Components*	51	71	75.20 (±0.99)
Isolates	NA	13	18.10 (±0.75)
Clustering coefficient	0.742	0.721	0.734 (±0.003)
Triangles	398	184	187.60 (±3.3)
Average path length	2.546	2.604	2.12 (±0.06)

* The rise in the component count is the consequence of known individuals becoming isolated and components splitted into smaller ones,

With more individuals, it is expected that the full dataset, with his 401 nodes, will be associated with greater values for degree, weighted degree and triangles, compared to the restricted dataset. Interestingly, peripheral unknowns and knowns (degree value of 1) occur in very similar proportions, respectively of 22.4% and 21.2% (20 unknowns, 66 knowns) (e.g., for unknowns see components 21, 31 and 48 in Fig. 2). The restricted dataset contains 13 individuals (known), who are now isolated because they were originally peripheral and linked to 13 of the 89 unknowns now removed (e.g., see components 9, 13 and 24 in Fig. 2). Consequently, a total of 33 unknowns are in peripheral-like position (20 +13), while the remaining 56 (63%) occupy more central positions in their component. On the other hand, the other measurements – i.e., density, diameter, modularity, average path length, clustering coefficient, which need not correlate positively to the number of nodes, are very similar when not equal for both sets. The removal of the unknowns also splits seven of the 51 large components into smaller ones (e.g., see components 10, 38 and 41 in Fig. 2). Thus, the restricted data set contains 58 components with > 1 individual +13 isolates, for a total of 71 components.

The 30 iterations with the random removal of 89 individuals results in similar measurements with some variation when compared to the restricted dataset (Table 2). This supports the hypothesis that unknowns are, at least, as well integrated as the knowns. In addition, two measurements support some deeper level for unknown's integration. Firstly, the average number of isolates is 18 (min 12, max 27) a value 38% higher than for the real restricted dataset where only unknowns were specifically removed. Consequently, there are fewer unknowns lying next to peripheral nodes than expected from their proportion in the network. Secondly, the average path length (APL) of 2.12 is lower compared to the value of 2.6 in the real restricted dataset. As the presence of unknowns (full dataset) slightly reduces the APL to 2.5, this shows that they are well integrated. On the opposite, the shorter APL of the simulated datasets shows that knowns, when integrated, increase the APL showing that they are not all fully interconnected but also largely peripheral. The shorter APL of simulated datasets is also a consequence of the randomly disrupted components that end up smaller and more numerous (75.2).

3.4.2 Within-component level

Normalized SNA statistics for degree and betweenness centrality, clustering, and egonet density, were examined at the component level for the full and restricted (i.e. without unknowns) datasets. The summation of statistics was done over the 51 large components, i.e. with ≥5 individuals and at least one unknown. Again, these components include a total of 89 unknowns (22%) and 312 knowns, for a total of 401 individuals.

In the full dataset, the unknowns show a total of 5% more links (degree) compared to the knowns, while they show similar betweenness centrality, clustering and egonet density (Fig. 4A). Removal of these unknowns had various effects on the network structure, causing a 19% and 10% decrease in normalized betweenness centrality and clustering, respectively (Fig. 4A). The slight change in egonet density is still within the boundaries of standard errors. Moreover the distribution of egonet density is almost identical for unknowns and knowns in the full and restricted datasets (Fig. S3). Although many links are lost in the restricted dataset, we observe a slight increase in degree for normalized values, showing that, in proportion, knowns are a little more connected when the unknowns are ignored.

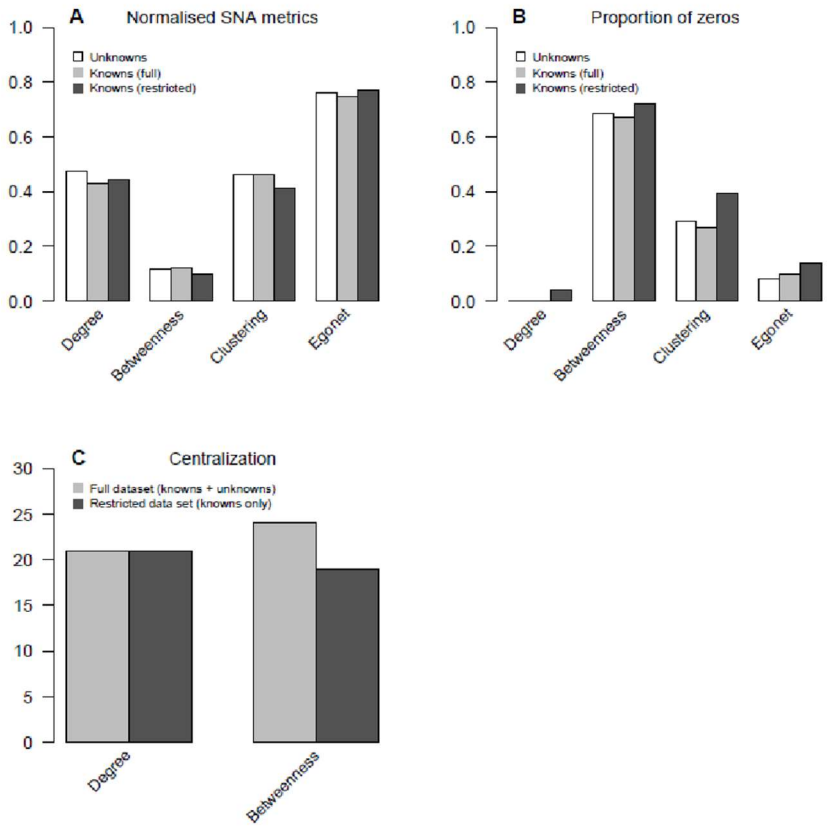


Figure 4: Social network analysis results for four normalized measurements (A), number of zero for the same measurements (B) and the centralization (C)

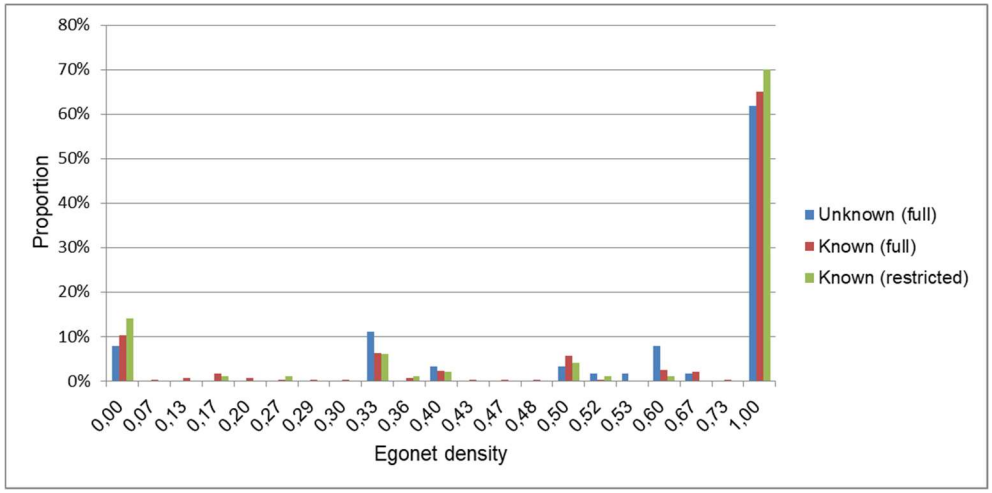


Figure S3: Egonet density distribution for the unknowns and knowns in the full and restricted datasets. For a value of 0, ego's alters are not connected together while they are all connected to one another when the density is 1.

The number of individuals with SNA metric values of zero also illustrates the non-negligible impact of unknowns on the network (Fig. 4B). When including unknowns (full dataset), all individuals are connected to at least another one, which is expected since this analysis starts with large components, hence all degree values must be > 0. When excluding unknowns (restricted dataset), 4% of knowns become isolated. Moreover, seven of the large components (14%) are fragmented into fourteen smaller ones. For betweenness centrality, clustering and egonet density, the restricted dataset respectively shows an increase of 7%, 46% and 40% in zero values. The variation in the SNA average measurements is shown in Table S2. Finally, the centralization of degree is identical for both datasets while that for betweenness centrality shows a decrease of 5% (Fig. 4C).

Table S2: Average values for the three SNA measurements with standard errors (SE) and Δ*K* associated with each one.

Metric	**Dataset**			
	Full		Resticted	
	Unknown	Known	Known	ΔK
Degree	0.48 (±0.03)	0.43 (±0.02)	0.44 (±0.02)	2%
Betweenness	0.12 (±0.02)	0.12 ±(0.01)	0.10 (±0.01)	-18%
Clustering	0.58 (±0.05)	0.58 (±0.02)	0.49 (±0.03)	-15%
Egonet density	0.76 (±0.04)	0.75 (±0.02)	0.77 (±0.03)	2.6%

The Δ*K* distribution of three SNA measurements is shown in Fig. 5. The average value of each metric for unknowns shows a negative relationship with Δ*K*. Therefore, the greater the difference between the summed metric values for knowns in the full vs. restricted datasets, the smaller the integration of unknowns to the network. In other words, the higher (and positive) the Δ*K*, the more known individuals appear as tightly interconnected when unknowns are removed. This occurs when unknowns themselves show a weaker integration to their network component (lower SNA metric values). In contrast, the lower (and negative) the Δ*K*, the less tightly known individuals are interconnected, and the more unknowns are.

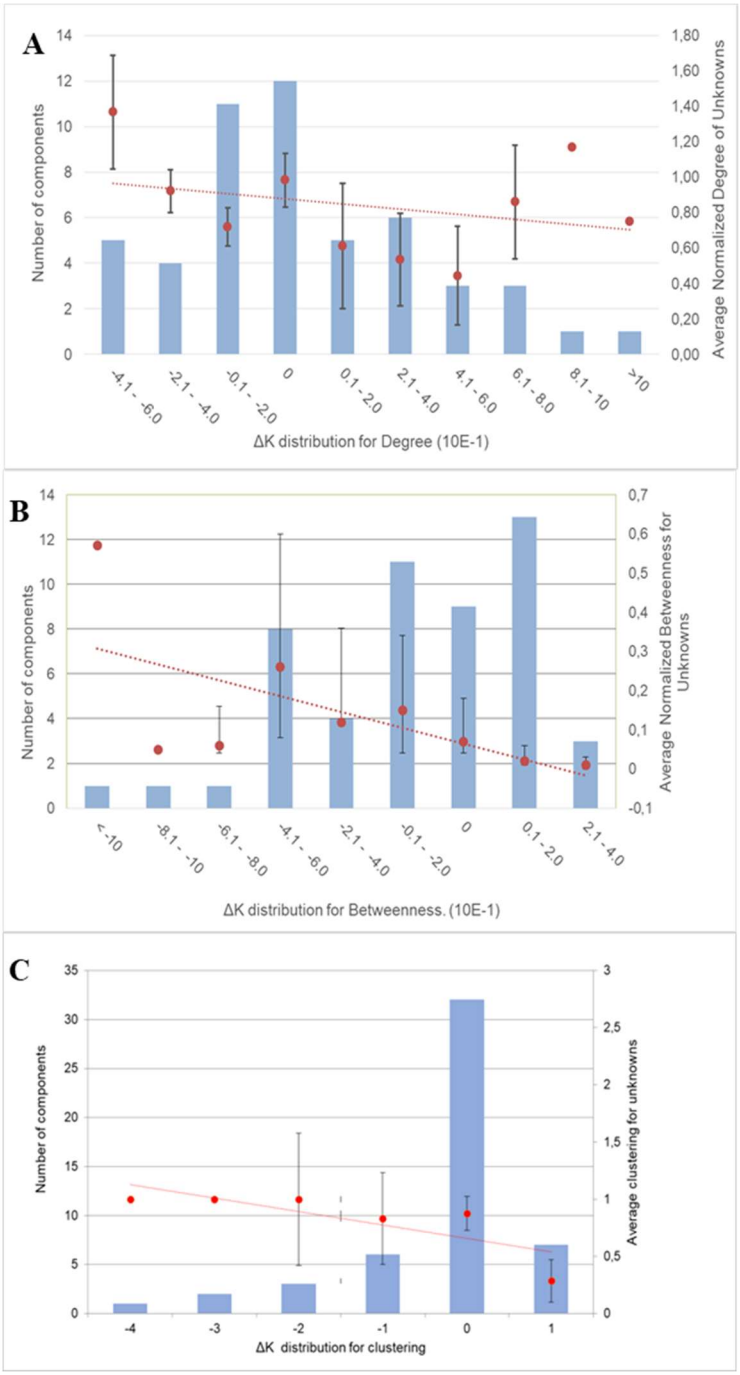


Figure 5: Distribution of the 51 large components as a function of Δ*K* for degree (A), betweenness centrality (B) and clustering (C). Δ*K* are grouped by magnitude. The average value of the measurement for the unknowns (red dots ± s.d.) is also reported (right Y-axis). The red lines are the slopes of metric values for unknown against the Δ*K* category. The correlation coefficients are A: -0.24, B:-0.54 and C:-0.17.

The strongest slope is observed for betweenness centrality (correlation coefficient $r = -0.54$; Fig. 5B) followed by the degree ($r = -0.24$; Fig. 5A). Thus, when unknowns are found in key positions, they may have a considerable impact on the component structure and on the weight of the known individuals surrounding them. On the opposite, peripheral unknowns contribute to positive ΔK values or have little effect. It is interesting to note the component proportions according to the negative, neutral or positive values of the ΔK. For betweenness centrality it is 51% of the components that are negatively affected (Fig. 5B).

3.5 Sensitivity analysis

SNA results for two 6-year data subsets show consistency with the entire 18-year dataset, albeit some discrepancies, when measurements are compared for knowns and unknowns (Fig. S4). The most salient differences are, firstly, the nearly twice higher betweenness centrality of unknowns relative to knowns for the 2012–2018 period, a difference not observed in 2005–2011 (Fig. S4C); secondly, the very high proportion (almost 70%) of known individuals having a zero value for the clustering coefficient for the 2012–2018 period, relative to unknowns (~20%; Fig. S4F).

3.6 Component dismantling

Inside some large components, centrality values of known individuals can change drastically when unknowns are removed, depending on their position within the component. In some cases, removal of an unknown situated in a “bridge position” dismantles the component. Fig. 6 shows two examples. Component 41 divides into two smaller ones with the removal of a single unknown, leading to a decline in the betweenness centrality of known individuals #7 (from 35 to 15) and #1 (from 41 to 6), with a large impact on the relative importance of these two individuals in the component (positive for #7 and negative for #1). In the second example, the removal of unknowns from component 25 causes known individual #4 to become the most central one, while he/she was previously in the shadow of three more central unknowns along with known #2, the latter being initially most central with a betweenness of 13.5. From the

sociograms in Fig. 2, one can identify many components for which the absence of the unknowns would cause important dismantling (e.g., component 38), while the majority would reduce to dyads or solo actors (e.g., components 9, 10, 11, 12, 24, 25, 26, 27, 29, 34, 35).

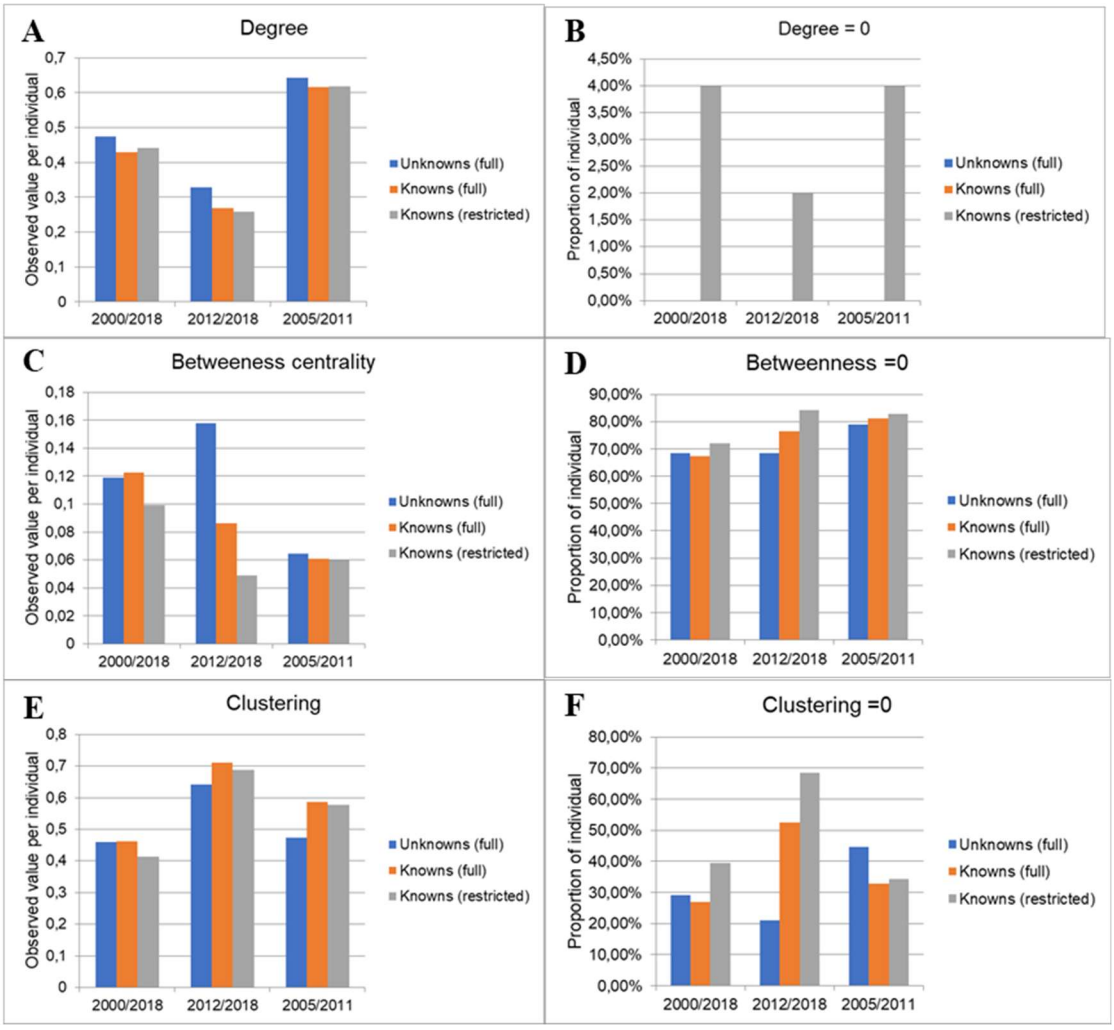


Figure S4: Sensitivity analysis results for Degree, betweenness centrality and clustering for the full (knowns and unknowns) and restricted (knowns only) datasets. In each panel, the data for the total study period (2000 to 2018) is compared to two subsets of 6 years (2012 to 2018, and 2005 to 2001). The three left panels show values per individual while the right panels present the proportion of the observed individual having a 0 value for each measurement. The N values used for proportion evaluation in each subgroup are respectively 89, 19 and 38 for the unknowns and 312, 51 and 122 for the knowns.

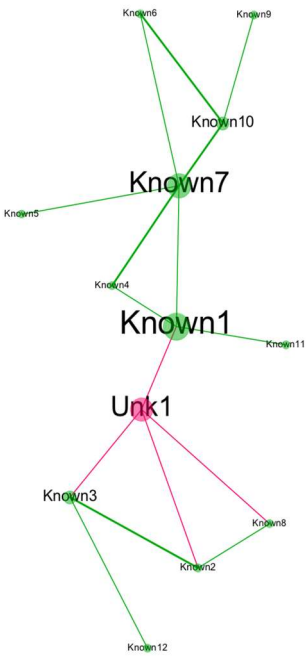
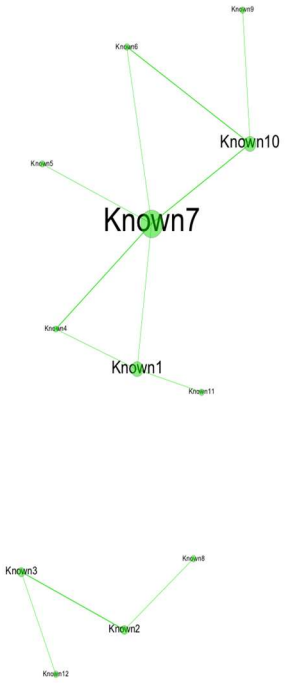
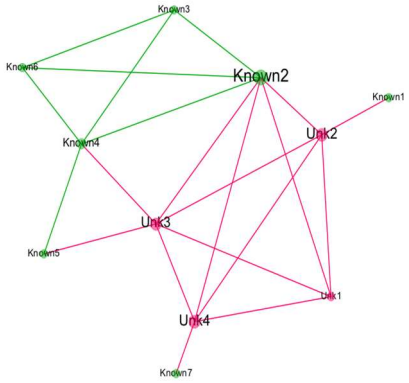
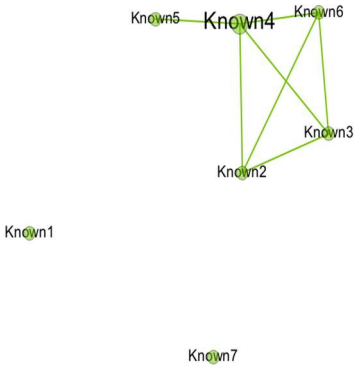
Component	Unknown and Known	Known Only
41 Known 7 35 —> 15 Known 1 41 —> 6		
25 Known 2 13,5 —> 0 Known 4 3,5 —> 3		

Figure 6: Two examples of component dismantling when unknowns are removed from the social network. Node size is proportional to betweenness centrality. Below each individual designation, the number shows the change in betweenness centrality. Component 41 qualifies for a betweenness ΔKn of $-1.9X10^{-1}$ and can be placed accordingly in Figure 5B along with component 25 with a ΔKn of $-1.8X10^{-1}$.

4. Discussion

Our results show that individuals known only from their DNA profiles, hence, who are unknown to investigators, occur in components of all sizes in the Québec criminal social network. Their ubiquity is, however, poorly accounted for in crime analysis, hence underexploited to orientate police operations, as attested by the small number of studies that have explored this potentiality (Jeuniaux *et al*. 2016; Lammers 2014; Lammers *et al*. 2012; Lammers and Bernasco 2013; De Moor *et al*. 2018; De Moor *et al*. 2020). As discussed below, social network analysis provides information about the proportion and position of unknowns in criminal networks and can serve as leverage to generate criminal intelligence and support criminological studies.

4.1 Distribution of unknowns

Most unknowns detected in LSJML data seemingly act alone or in dyads, as in the case of known offenders. Nonetheless, a non-negligible number of them are found in bigger components. Overall, 29% of components include unknowns, a proportion increasing respectively to 36% and 64% for components of medium (3–4 individuals) and large (≥5 individuals) size. The density of such a fragmented network is very low with only 0.3% of theoretically possible among- individual links actually observed. This important fragmentation perhaps makes it difficult for police investigators to identify a large proportion of unknowns, because we should expect that few known individuals hold information on a given unknown. We found that the number of unknowns (84) in large components was almost equal to the mean value expected (85) under a random distribution of repeated offenders across components (Fig. S2). However, when considering the proportion of unknowns (21%), we see that there are 32% fewer unknown individuals in the larger components. The larger proportion of unknowns observed in the medium size components and the smaller value observed in the solo actors group compared to the random average, suggest that these offenders seem attracted to some co-offending to the detriment of solo acting. On the other hand, if considering the general proportion of 21% the difference between the observed and proportionate value is much less different but switch to a lower value for the medium

size components and slightly higher for the solo actors. This suggests that it is more difficult for unknowns to elude detection when they commit crimes with more co-offenders (also supporting the above hypothesis about the greater difficulty to identify unknowns in very fragmented networks). If so, this could be explained, for example, by denunciations or greater criminal activity, leading to greater exposure. Actually, Lammers, *et al*. (2012) found that offenders committing more crimes have a higher probability of being arrested.

More importantly, these results show how the numerous known solitary non-repeat offenders could bias the random distribution of unknown that should be compared with repeat offenders only which are finally present in the large components in the same proportion as expected, while they seem over-represented in the complete data.

4.2 Potential impact on criminal investigations

Our results show the important impact unknowns may have when not considered in criminal investigations or network analysis. First, 14% of dyads with at least one unknown disappear, while the remaining 86% reduce to solitary, known actors. Likewise, 58% of medium size components (3–4 individuals) reduce to dyads, 17% to solitary actors, while two components composed only of unknowns disappear (Table S1). Larger components are affected in more various ways by the removal of unknowns, depending on the position of the latter. Therefore, DNA match data can bring a good amount of co- offending information that is absent from typical police files. This is true for 18% of all components in LSJML data and 63% of those with unknowns.

Second, applying SNA measurements to larger components reveals that unknowns contribute to degree, betweenness, clustering and egonet density values roughly in accordance to their proportion. Consequently, their removal has consequences on SNA measurements for known individuals. The betweenness centrality of the latter is most affected, with a decrease of 19%, which interestingly is a value close to the proportion of unknowns (21%). Moreover, ~50% of Δk values are negative, indicating that unknowns occupy either central or peripheral positions about as often as known individuals do, a conclusion further supported by the analysis of datasets generated by the random

removal of individuals, instead of that of unknowns specifically. Remarkably, De Moor *et al*. (2020) reached an opposite conclusion with Belgian DNA match data. This seems largely explained by (1) the fact that 43% of unknowns occur as solitary actors in their data, hence not connected to any known offender (vs. 0% here), and (2) not taking into account the discontinuity of the network (as explained in Introduction). Consequently, De Moor *et al.*'s (2020) results must be interpreted differently than ours. In their case, a Δk would be determined not only by the position of unknowns in multi-offender clusters but also by the proportion of solitary unknown actors, thereby conflating the socialization of co- offending unknowns (i.e., network structure) with their general contribution to criminal activity (i.e., the total number of cases involving them), as these authors themselves underscore. In our case, Δk relates to the co-offending socialization of unknowns (actually, the present study did not aim to address the contribution of solitary actors to crime incidence).

Third, the rise of 46% in the proportion of zeros values for the clustering coefficient shows that peripheral unknowns are well connected to their neighborhood through triads. This large rise is even more striking when compared to the rise of 7% in zero values for betweenness centrality. These results are not that surprising since a network structure usually has more individuals (known and unknown) in peripheral zones compared to central ones, yet it shows that unknowns are as well integrated as the knowns in the global component's structure. Moreover, the egonet density shows that neighbors of a focal individual (the ego) are connected with a similar density whether this individual is known or unknown.

Overall, our results show that unknown offenders have been non-negligible actors in Québec's criminal network (as reconstructed from LSJML DNA match data). They are not particularly more marginal or central then known individuals. Therefore, incorporating them in crime analysis can generate additional intelligence that should be precious for police operations. For instance, our results show that some unknowns may deserve greater attention even when they show up in case files for which (known) offenders have already been arrested. Accounting for unknowns not only uncovers more

co-offending relationships, it also modifies the evaluation of the importance of known individuals relative to each other in the network (as in Fig. 6). In addition, it can reveal that a known offender is less socially marginal than previously thought and consequently may have more information to provide to the police than initially expected. Therefore, we see how considering unknown individuals at the social network level could guide police tactical decisions, resource allocation, prioritization of investigations, and types of interventions in the field. Therefore, social network analysis of DNA matches can directly support intelligence-led policing (Ratcliffe 2016; Ribaux 2016).

4.3 Network dynamics

Sparrow (1991) identified two main issues with SNA, namely the incompleteness of the data and network dynamics. The former has been tackled by several authors with methods involving random or non-random nodes removal (or addition) to disrupt the network structure and thereby evaluate potential biases in SNA measurements (Borgatti and Krackhardt 2006; Smith and Moody 2013; Smith *et al*. 2017; Frantz *et al*. 2009). DNA data has the immense advantage of providing substantial information on unknown offenders, yet it reflects a limited part of total criminal activity because many criminal investigations do not involve genetic expertise (Rossy *et al*. 2013; Ribaux and Wright 2014). Thus, DNA- based inferences could be generalized to the whole criminal population only if one assumed that the characteristics and behavior of unknown offenders are similar, whether or not their DNA is detected on crime scenes. Testing this assumption should be an important objective of future studies.

Regarding network dynamics, Morselli (2009) studied a large drug trafficking network, and Charette and Papachristos (2017) the evolution of the dyadic co-offending activity over an eight-year period in Chicago. These two studies show that the structure of networks builds over time from co-offending stacking. As stated by Gründ and Morselli (2017): "… co-offending relationship as a multiplex network built around dyadic dynamics". In other words, a number of large network components are indeed a historical representation of a collection of dyadic co-offending. To assess how network

dynamics could affect our conclusions, we conducted a sensitivity analysis that unraveled some variation in SNA measurements as a function of the time period covered. Firstly, the lower proportion of zero values for betweenness centrality (Fig. S4D) for the complete dataset (2000/ 2018), relative to subsets covering shorter time intervals, reflects the fact that when more people become connected, betweenness overall increases, and consequently fewer zero values are observed. By contrast, clustering is lower for the complete dataset (2000/2018), as it includes more peripheral individuals relative to subsets covering shorter time intervals (Fig. S4E). While these results call for caution in SNA interpretation, overall they do not lead to drastically different conclusions about the position of unknown offenders in the network. Besides, this issue of dynamism is not specific to networks based on DNA match data but holds for any social network reconstructed from crime data.

Nevertheless, our sensitivity analysis is only a coarse assessment of the issue of network dynamics. The latter must be kept in mind when interpreting static SNA parameters. For example, an individual may occupy a central place in a component for different reasons: he/ she may be the hub of a contemporary group of offenders, or alternatively have simply changed co-offender several times during his/ her criminal career, etc. Individuals connected by DNA matches could have started, maintained or desisted from criminal activities, at different ages, with or without the presence of peer offenders (Farrington *et al*. 2014; Eggleston and Laub 2002; Haynie 2001). Turning points in one's life, such as marriage, parenthood, or new employment, are known to influence the course of a criminal career (Uggen and Staff 2004; Pyrooz *et al*. 2017; Blomberg *et al*. 2012). In future studies, detailed analysis of how components develop through time should provide additional insights into criminality, hence valuable information to further feed criminological studies and intelligence-led policing. Such dynamic mapping of social relationships will also allow the study of turning points (Bichler 2019). Capturing those changes is difficult in longitudinal studies (Haynie 2001), and this should be especially true with DNA match data that can miss a part of an individual’s criminal history. However, this is certainly not impossible, but rather a stimulating challenge that could be addressed by joining multiple sources of forensic,

police, and criminological data to obtain a more complete picture of network dynamics and criminal histories.

4.4 Limitations of the study

For our analyses we filtered LSJML data to keep only repeat offenders for the reasons explained in Methods. The large representation of non-repeat known offenders in the NDDB could reflect that some of these individuals, who were betrayed once by their DNA, are well aware of the risk and adapt their behavior to reduce it when committing subsequent crimes (Beauregard and Bouchard 2010). Actually, about two thirds of the individuals who were identified by DNA in a single case had been previously included in the NDDB following a conviction relative to another crime not involving a DNA match (i.e., their genetic profile was uploaded in the Convicted Offender Index upon conviction). Consequently, they truly are repeat offenders but not seen as such in the DNA data. It is their ability or chance of not leaving DNA behind, or that their traces went undetected, that made them appear as single-case offenders. Moreover, while 55% of the individuals in the repeat offender group had their DNA detected in two cases, Farrington *et al*. (2014) found from self-reported criminal activity that only 2.5% of offenders confess a level of activity that low (i.e., of one or two crimes). This large difference suggests that a number of individuals linked to 1–2 cases in our data committed more crimes than this. In addition, a small group (1.3%) of unknown individuals occurring only once in LSJML data, hence absent from our SNA analyses, are actually repeat offenders occurring in criminal cases analyzed by other Canadian laboratories. Consequently, we missed the complete picture for some components in which individuals may be more active in other parts of Canada, and possibly with other individuals. Moreover, individuals with a higher forensic awareness of the risk to be caught by their DNA could elude detection by taking more care when committing crimes (Beauregard and Bouchard 2010), thereby potentially reducing the number of unknowns that could be detected otherwise.

DNA databanks can contain genetic profiles from unknown individuals who, if they were known, would prove to be irrelevant to investigations. Actually, it is not rare

to find on a scene background DNA from unknown people who have nothing to do with the crime, which we designate here as fortuitous donors. Telling their DNA profiles apart from relevant ones is not always possible. Nevertheless, it seems improbable to observe a fortuitous donor appearing more than once in a databank. Therefore, while they may stay unknown forever, these fortuitous donors are unlikely to cause major problems in network analysis. Since a fortuitous donor is generally linked to a single criminal case, most should appear either as solitary components and excluded from DNA match data (see Methods) or as peripheral individuals in co-offending components to which he may be erroneously associated. Future studies could explore this question more in depth, for example by re-conducting SNA after filtering clusters according to criteria related to individual criminal activity such as offense number or the type of environment where DNA traces occurred (since fortuitous DNA might be more likely to be found in some environments than others). In addition, various weights could be given to edges, e.g. as a function of co- offending frequency.

A related issue is that true unknown offenders whose genetic profile was detected in a single case file (i.e., non-repeat unknown offenders) will also be missing from DNA match data (let's recall that a match requires two DNA profiles). Had we gotten information about them, they would have contributed to increasing the values of degree and betweenness centralities for some knowns (and possibly some other unknowns). Thus, it is likely that SNA measurements for some known individuals are underestimated by the absence of these non-repeat unknown offenders in our data. The impact of this missing data on the network structure remains unexplored and future studies should address the issue. As we don't know the exact amount of data that fall in that category, two scenarios could stand out. A small amount of single-offense unknown individuals may not have an important effect on the general structure of the observed components and associated metrics, as true non-repeat unknown offenders are likely to have fewer social connections with other criminals and remain peripheral. If these unknowns are numerous, their absence from SNA analyses could bias SNA metrics for other offenders, and change the number of components detected, as well as their size distribution. More studies (e.g., simulations) are needed to assess whether this bias could

have a noticeable impact on between-group comparisons (e.g., knowns vs. unknowns in the SNA data), which may not be always the case, such as when the groups compared are affected similarly by the absence of non-repeat unknown offenders in the data. On the other hand, it could happen that an individual who is missing in SNA data was the only one that linked together two or more clusters. In the other studies that examined criminal networks reconstructed from DNA matches, non- repeat unknown offenders were also not included but there was no mention of the possible consequences of this on their findings (Jeuniaux *et al*. 2016; Lammers 2014; Lammers *et al*. 2012; Lammers and Bernasco 2013).

Contamination from crime scene workers may occur and represent a greater problem than background DNA from fortuitous donors. Indeed, if the same person leaves DNA traces on several crime scenes, those traces could be interpreted as coming from an unknown repeat offender (Ballantyne *et al*. 2013). Forensic laboratories address this issue by keeping databases containing DNA profiles from people (e.g., police officers, scientists) investigating crime scenes (Ansell 2013; ENFSI working group 2014), although such bases can be quite incomplete. At LSJML, a 'reporting scientist elimination database' has been operational since the very creation of the National DNA Data Bank of Canada. However, in Canada it is not mandatory for crime scene examiners or police officers to provide a sample of their own DNA for elimination purposes (Lapointe *et al*. 2015), as opposed to some other countries or regions (e.g., England and Wales; (Forensic Science Regulator 2014)). In a pilot project, the LSJML built such a database from volunteer crime scene workers, which led to the exclusion of 58 non-relevant DNA profiles detected in criminal cases, some being recurrent (Lapointe *et al*. 2015). This is a rather small number in comparison to the tens of thousands of profiles already accumulated in the databank. Yet, they sometimes can have important consequences for investigations, as illustrated by an example in Lapointe (2015) involving two aggravated thefts, an attempted murder case, and a breaking and entering event.

Finally, the statistical analyses performed by random permutations of knowns and unknown in the network, or by randomly removing individuals, were chosen because they are better adapted to observations that are not independent, which is typical of social network data (De la Rua 2004). Permutations clearly support that unknown offenders are not randomly distributed across component-size categories, with medium components containing significantly more of them than expected by chance, while the reverse is true for small components. It should also be noted that the LSJML dataset represents the complete information on individuals in DNA-reconstructed SNA components, and not a sample. These individuals do, however, constitute a "sub-population" of all offenders that committed crimes in Québec during the study period. This sub-population can be defined as the group of offenders who both committed crimes in Québec during the study period *and* left DNA traces that were detected. This group could be fairly representative of the whole offender population, but we cannot assume it at this moment. Some types of individuals could be less represented in our data, like the ones with more forensic awareness (Beauregard and Bouchard 2010). Addressing the representativeness issue will be an important step in future research on DNA-reconstructed criminal networks. Despite the limitations discussed above, our findings show that DNA matches convey unique and useful information to improve knowledge about unknown offenders that are part of criminal social networks. Our study is also of interest for police investigations, at least at the jurisdictional level covered by the data. Finally, it is no surprise that numerous questions emerge by following the path of unknowns in criminal networks as this field of research using DNA matches data is in its infancy. As stated by Morselli (2013) about criminal networks: "Access to a variety of data sources and the creation of new research designs have also been amongst the more serious challenges facing researchers in this area"…. "The first challenge begins accessing relevant data sources… in the real world".

4.5 Specificities of DNA-based social network analysis

DNA traces analyzed in a forensic laboratory represent typically a small fraction of crimes known to the police, which represents itself only a subset of total crime. DNA is also rarely found for some crime types, like frauds. However, when present, DNA can

establish strong links between individuals and crime scenes. Moreover, DNA profiles can be kept in databanks and thus provide forensic intelligence a long time after their initial analysis, sometimes up to decades after (Rossy *et al*. 2013).

Beyond the idiosyncrasies of criminality in each territory, legal, strategic and operational variation among jurisdictions will determine what forensic intelligence can be extracted from social network analyses based on DNA data. For example, criteria for including offenders in a DNA database vary from one country to another. In their study of criminal networks in the Netherlands, Lammers and Bernasco (2013) excluded DNA matches made prior to 2002 because, up to that year, only offenders facing a prison sentence of eight years or more were included in the DNA databank. A similar situation prevailed in Belgium up to 2014 (Jeuniaux *et al*. 2016).

5 Conclusion and future research

The dark figure of criminality is important but hard to address to set light on illicit activities behind closed doors (Morselli 2009). Despite its limitations, DNA match data allow to half-open those doors behind which unknown offenders hide. The DNA sequence remains constant throughout life, making it possible to reliably connect crimes committed by the same person at different times and places. DNA analysis is also highly standardized and reproducible, which helps to reconstruct social relationships from data coming from different jurisdictions. This provides an advantage relative to other kinds of forensic or investigative data. For example, documenting the contribution of unknown offenders to criminal networks based on shoe marks or CCTV images is certainly valuable, but we may expect less consistency in marks left by an offender than with DNA. Moreover, unlike DNA, these kinds of forensic traces are generally not shared nationwide in a common database.

In this study, we compared SNA measurements for known and unknown individuals to assess their relative importance in the social network of criminality in Québec, for the longest period covered so far in this type of analysis (18 years). As initially proposed by Rossy and Ribaux (2014) and then by De Moore (2020), we

focused on network components large enough to generate a comprehensive analysis of offender socialization. We found that unknown individuals who leave DNA traces on crime scenes appear to be integrated into criminal networks as strongly as known offenders are.

We only begin to see the potential of DNA match data to generate forensic intelligence and criminological knowledge. As aforementioned, Lammers and Bernasco (2013) used DNA matches to examine factors that may explain the probability of an offender to be arrested. This raises the broader issue of criminal trajectories, which it will be increasingly possible to study as DNA databanks cover larger time periods. In particular, several variables linked to criminal cases are routinely included in the LSJML data and open new ways to study the position of unknowns in networks from the temporal dynamic perspective of criminal careers. More generally, the incorporation of unknown individuals in social network analysis will help uncover a part of the big picture of criminal activities (e.g., structure of criminal organizations, co-offending patterns, criminal careers, central individuals to target, network dynamics, etc.). As proposed by Ribaux and Talbot-Wright (2014), this approach can be part of a sound forensic intelligence strategy, which should reveal its full potential when information from different types of traces or forensic analyses (e.g., drug profiling or false document analyses (Baechler and Margot 2016; Baehler *et al*. 2015; Morelato *et al*. 2014), as well as from police files, are linked together to support investigation strategies and policing at the operational level. By currently using the DNA ID information only for what it is, the judicial system is losing the power of intelligence that could be gathered by the networking structure of DNA matches focusing on unknowns.

Funding information

LL received financial support from the Forensic Research Group (U. du Québec at Trois-Rivières) and from the International Centre for Comparative Criminology.

Declarations of interest

None.

Acknowledgements

Carlo Morselli made an instrumental contribution to the designing of this study. Sadly, he passed away in 2020. He was professor at Université de Montréal and former director of the International Centre for Comparative Criminology. We dedicate this work to his memory. Also, special thanks to Prof. Martina Morris from University of Washington for insightful discussions about social network data analyses.

Appendix A

Definitions

Social network analysis studies used concepts and terms that we clarify here. Even the 'network' term needs some precision. Actually, a network is, as Borgatti et *al.* noted: "A *way of thinking about social systems that focuses our attention on the connections or relations among the entities that make up the system*" (Borgatti 2018). From this definition, the total set of data that is of interest in this article could be seen as constituting a single network where every individual has some probability of being connected to any other one, because they could share common criminal activities.

In the literature, 'network' also sometimes refers to a component, and defined as: "*A maximum set of nodes in which every node can reach every other by some path*" (Borgatti 2018). Here we define the network as composed of numerous components of various sizes; if a recently added criminal activity connects individuals that were before in separate components, the older components are now joined in a new larger one. Confusion may be observed in the literature and the reader should pay attention to the mixed use of the terms 'network 'and 'component'.

Networks can be studied by various statistical parameters where some are designed to characterize the network itself while others to measure the importance of a node in a component. There are many ways in which a node can be important. In this study, we used four measurements to assess the importance of unknowns.

The degree: "*The number of edges incident on a node or the number of nodes adjacent to another node*" (Borgatti 2018). Under a bimodal data visualization, the degree would express the number of criminal cases associated with an individual, while under a unimodal visualization, the degree is the count of neighbors in relation to a specific focal individual.

The betweenness centrality: "*A measure of how often a given node falls along the shortest path between two other nodes*" (Borgatti 2018; Freeman 1978). This measure gives a more accurate assessment of the importance of an individual in a component, compared to the degree. A high value of betweenness centrality refers to a node occupying a central position, making the link between some other nodes that are not connected. In a component where everyone is connected to everyone else, there is no central position so that betweenness centrality is zero for each node.

The clustering coefficient is the proportion of triangles in the component, namely sets of three nodes all connected, relative to the maximal possible number (Bichler 2019), and is "*One of the most important patterns in terms of understanding the inter-connectivity of nodes in real graphs*" (Durak *et al*. 2012). In a triangle, individual A knows B, B knowns C, and A knows C, resulting in a closed transitive structure. Clustering indicates the proportion of transitive patterns associated with each individual (Borgatti 2018). For example, an unknown offender at the component periphery could have a high clustering coefficient, revealing in which case his/her important inter-connectivity position despite a betweenness centrality of zero.

The egonet density is related to the study of ego networks, a type of analysis that focuses on a personal-network including the surrounding individuals seen as alters. The egonet density is the number of ties between ego's alters, divided by the total number of possible ties (Borgatti 2018), and takes values between 0 and 1. When all alters are connected to an ego but no ties exist between these alters, the egonet density is 0; on the opposite, when all the possible ties exist between the alters, the density equals 1.

Centralization: A calculation based on Freeman's formula (Freeman 1978) that gives: "The extent to which a network is dominated by a single node. Specifically, "the extent to which one node is much more central than all others" (Borgatti 2018). In any component or network these ratio values can be from 0 to 1 where 1 represents a situation where an individual completely dominates the network (or component) with respect to centrality. In this article we looked at the general impact of the unknowns removal by comparing the sum of the centralizations observed in the two data sets.

Appendix B. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.forsciint.2021.111142.[3]

References

Ansell, R. (2013) Internal quality control in forensic DNA analysis, *Accreditation Quality Assurance* 18, 279–289.

Baechler, S. and Margot, P. (2016) Understanding crime and fostering security using forensic science: The example of turning false identity documents into forensic intelligence. *Security Journal* 29, 4618–639.

Baechler, S., Morelato, M., Ribaux, O., Beavis, A., Tahtouh, M., Kirkbride, K. P., Esseiva, P., Margot, P. et Roux, C. (2015) Forensic intelligence framework. Part II: Study of the main generic building blocks and challenges through the examples of illicit drugs and false identity documents monitoring. *Forensic Science International* 250, 44-52.

Ballantyne, K.N., Poy, A.L. and Van Oorschot, R.A.H. (2013) Environmental DNA monitoring: beware of the transition to more sensitive typing methodologies. *Australian Journal of Forensic Science* 45, 323–340, https://doi.org/10.1080/00450618.2013. 788683

Bastian, M., Heyman, S. et Jacomy, M. (2009) Gephi: an open source software for exploring and manipulating networks. *International AAAI conference on weblogs and social media* 8, 361-362.

Beauregard, E. and Bouchard, M. (2010) Cleaning up your act: forensic awareness as a detection avoidance strategy. *Journal of Criminal Justice* 38, 1160–1166.

[3] Les tableaux S1, S2, S3 et les figures S1, S2, S3 et S4 supplémentaires, accessibles sur internet sont inclus ici dans le texte.

Bichler, G. (2019) *Understanding Criminal Networks: A Research Guide.* University of California Press.

Blomberg, T. G., Bales, W. D. et Piquero, A. R. (2012) Is educational achievement a turning point for incarcerated delinquents across race and sex? *Journal of youth and adolescence* 41, 202-216.

Borgatti, S.P., Everett, M.B. and Freeman, L.C. (2002) *Ucinet for Windows: Software for Social Network Analysis.* Analytic Technologies, Harvard, Massachusetts.

Borgatti, S.P., Everett, M.G. and Johnson, J.C. (2018) *Analyzing Social Networks*, Sage.

Borgatti, S.P., Carley, K.M. and D. Krackhardt, (2006) On the robustness of centrality measures under conditions of imperfect data. *Social Network* 28, 124–136.

Canadian Criminal Code, R.S.C, 1985, c. C-46, sect. 487.05.

Charette, Y. et Papachristos A. V. (2017) The network dynamics of co-offending careers. *Social Networks* 51, 3-13.

De la Rua, F. (2004) L'analyse longitudinale de réseaux sociaux totaux avec Siena. Méthode, discussion et application. *Bulletin de méthodologie Sociologique* 84, 5–39.

De Moor, S., Vandeviver, C. et Vander Beken, T. (2018) Are DNA data a valid source to study the spatial behaviour of unknown offenders? *Science and Justice* 58, 315-322

De Moor, S., Vandeviver, C. et Vander Beken, T. (2020) Assessing the missing data problem in criminal network analysis using forensic DNA data. *Social Networks* 61, 99-106.

Durak, N., Pinar, A., Kolda, T.G. and Seshadhri, C. (2012) *Degree relations of triangles in real-world networks and graph models.* in: Proceedings of the Twenty First ACM International Conference on Information and Knowledge Management. ACM.

Eggleston, E.P. and Laub, J.H. (2002) The onset of adult offending: a neglected dimension of the criminal career. *Journal of Criminal Justice* 30, 603–622.

ENFSI, D., (2014) Working Group DNA-database Management Review and Recommendations. ⟨http://www.enfsi.eu/about-enfsi/structure/ working groups/DNA⟩. (Accessed 12 September 2019).

Farrington, D. P., Ttofi, M., Crago, R. W. et Coid, J. W. (2014) Prevalence, frequency, onset, desistance and criminal career duration in self-reports compared with official records. *Criminal Behaviour and Mental Health* 24, 241-253.

Forensic Science Regulator, (2014) Codes of Practice and Conduct, Protocol: DNA Contamination Detection-The Management and Use of Staff Elimination DNA Databases, FSR-P-302, 49.

Freeman, L. C. (1978) Centrality in social network conceptual clarification. *Social Network* 3, 215-239.

Frantz, T.L., Cataldo, M. and Carley, K.M. (2009) Robustness of centrality measures under uncertainty: examining the role of network topology. *Computational and Mathemathical Organization Theory* 15, 303–328.

Gallupe, O. (2016) Network analysis, in: *The Handbook of Measurement Issues in Criminology and Criminal Justice* (under the direction of Huebner, B. M. and Bynum, T. S). John Wiley & Sons, pp. 555–575.

Gründ, T. et Morselli, C. (2017) Overlapping crime: Stability and specialization of co-offending relationships. *Social Networks* 51, 14-22.

Haynie, D. L. (2001) Delinquent peers revisited: Does network structure matter? *American Journal of sociology* 106, 1013-1057.

Hochholdinger, S., Arnoux, M., Delémont, O. et Esseiva, P. (2019) Forensic intelligence on illicit markets: the example of watch counterfeiting. *Forensic Science International* 302, 109868.

Hu, Y. (2005) Efficient, high-quality force-directed graph drawing. *Mathematical Journal* 10, 37–71.

Interpol, (2019) Global DNA Profiling Survey Results 2019. ⟨https://www.interpol.int/How-we-work/Forensics/DNA. (Accessed 28 July 2020).

Jeuniaux, P. P. J. M. H., De Moor, S., Robert, L., Renard, B., Stappers, C. et Vanvooren V. (2017) Reconstruction and study of offending trajectories through forensic evidence. Dans *The Routledge International Handbook of Forensic Intelligence and Criminology*. Routledge.

Jeuniaux, P. P. J. M. H., Duboccage, L., Renard, B., Van Renterghem, P. et Vanvooren, V. (2016) Establishing networks in a forensic DNA database to gain operational and strategic intelligence. *Security Journal* 29, 584-602. https://doi.org/10.1057/sj. 2015.31

Lammers, M. (2014) Are Arrested and Non-Arrested Serial Offenders Different? A Test of Spatial Offending Patterns Using DNA Found at Crime Scenes. *Journal of Research in Crime and Delinquency* 51, 143-167.

Lammers, M. et Bernasco, W. (2013) Are mobile offenders less likely to be caught? The influence of the geographical dispersion of serial offenders' crime locations on their probability of arrest. *European Journal of Criminology* 10, 168-186.

Lammers, M., Bernasco, W. et Elffers, H. (2012) How Long Do Offenders Escape Arrest? Using DNA Traces to Analyze When Serial Offenders Are Caught. *Journal of Investigative Psychology and Offender Profiling* 9, 13.

Lapointe, M., Rogic, A., Bourgoin, S., Jolicoeur, C. and Séguin, D. (2015) Leading-edge forensic DNA analyses and the necessity of including crime scene investigators, police officers and technicians in a DNA elimination database. *Forensic Science International Genetics* 19, 50–55.

Milot, E., Lecomte, M., Germain, H. et Crispino, F. (2013) The national DNA data bank of Canada: a Quebecer perspective. *Frontiers in genetics* 4, 249.

Ministère de la sécurité publique. (2013) La desserte policière municipale, provinciale et autochtone au Québec profil organisationnel. Bibliothèque et Archives nationales du Québec. ISBN978-2-550-72638-8.

Morelato, M., Baechler, S., Ribaux, O., Beavis, A., Tahtouh, M., Kirkbride, P. and Margot, P. (2014) Forensic intelligence framework—Part I: induction of a transversal model by comparing illicit drugs and false identity documents monitoring. *Forensic Science International* 236, 181–190, https://doi.org/10.1016/j.forsciint.2013.12.045

Morselli, C. (2009) *Inside criminal networks*. Springer.

Morselli, C. (2013) *Crime and networks*. Routledge.

Pyrooz, D. C., Mc Gloin, J. M. et Decker, S. H. (2017) Parenthood as a turning point in the life course for male and female gang members: a study of within-individual changes in gang membership and criminal behaviour. *Criminology* 55, 869-899.

Ratcliffe, J.H. (2016) Intelligence-led Policing. Routledge.

R Core Team, (2017) R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing. Vienna, Austria, (URL), ⟨https://www. R-project.org/⟩.

Ribaux, O. (2014) Police scientifique : le renseignement par la trace. Lausanne, Presses polytechniques et universitaires romandes.

Ribaux, O. et Talbot Wright B. (2014) Expanding forensic science through forensic intelligence. *Science and Justice* 54, 494-501. https://doi.org/10.1016/j.scijus. 2014.05.001

Rossy, Q. et Ribaux, O. (2014) A collaborative approach for incorporating forensic case data into crime investigation using criminal intelligence analysis and visualisation. *Science and justice : journal of the Forensic Science Society* 54, 146-153. https://doi.org/10.1016/j. scijus.2013.09.004

Rossy, Q., Ioset, S., Dessimoz, D. et Ribaux, O. (2013) Integrating forensic information in a crime intelligence database. *Forensic Science International* 230, 137-146. https:// doi.org/10.1016/j.forsciint.2012.10.010

Royal Canadian Mounted Police. (2018) National DNA Data Bank - History, ⟨http://www. rcmp-grc.gc.ca/nddb-bndg/histo-eng.htm. (Accessed 4 September 2018).

Royal Canadian Mounted Police. (2018) "Primary - Historical" in NDDB DNA Designated Offences - Section 487.04 of the Criminal Code of Canada (Revised 2018/08). ⟨http://www.rcmp-grc.gc.ca/nddb-bndg/form/ddo-did-eng.htm#prim-hist. (Accessed 4 September 2018).

Royal Canadian Mounted Police. (2019) National DNA Data Bank annual report 2018–2019, 13. ⟨http://www.rcmp-grc.gc.ca/pubs/nddb-bndg/index-eng.htm.

Smith, J.A. and Moody, J. (2013) Structural effects of network sampling coverage I: nodes missing at random. *Social Network* 35, 652–668.

Smith, J.A., Moody, J and Morgan, J.H. (2017) Network sampling coverage II: The effect of non-random missing data on network measurement, Social Network 48, 78–99.

Sparrow, M. K. (1991) The application of network analysis to criminal intelligence: An assessment of the prospects. *Social Networks* 13, 251-274.

Uggen, C. and Staff, J. (2004) Work as a turning point for criminal offenders. *Crime and Employment: Critical Issues in Crime Reduction for Corrections for corrections* 65, 141–168.

CHAPITRE III

DNA Databanks as a Source of Information about the Criminal Behavior of individuals who have been linked to Crimes But Not Identified by Police

Lavergne, Léo[a,b]; Boivin, Rémi [b,c], Baechler, Simon[a,b,e,f], Fiola, Karine[d];, Séguin, Diane[d], Lefebvre, Jean-François[d], Milot, Emmanuel[a,b]

a: Forensic Research Group and Département de chimie, biochimie et physique, Université du Québec à Trois-Rivières, Trois-Rivières, Québec, Canada.
b : Centre international de criminologie comparée, Québec, Canada.
c: École de criminologie, Université de Montréal, Montréal, Québec, Canada.
d: Laboratoire de sciences judiciaires et de médecine légale, Ministère de la Sécurité publique, Montréal, Québec, Canada.
e : Ecole des sciences criminelles, Université de Lausanne, Lausanne, Switzerland
f : Domaine Traces et Analyse criminelle, Police neuchâteloise, Neuchâtel, Switzerland

Le contenu de ce chapitre a été soumis une première fois le premier novembre 2022, puis le 15 décembre 2023 pour publication en anglais dans la revue « Canadian Journal of Criminology and Criminal Justice ».
Cette revue utilise un processus d'examen par les pairs.

3.1 Contribution des auteurs

Dans cet article, l'ensemble de l'analyse des données, incluant plusieurs niveaux de vérification de ces dernières, ainsi que la rédaction des textes a été effectuée par Léo Lavergne. Les 1000 itérations aléatoires utilisées dans la section traitant de la distribution comparative des types de délits entre les inconnus et les connus ont été obtenues par un script R créé par Emmanuel Milot. Ce dernier ainsi que Rémi Boivin ont révisé le manuscrit jusque dans la dernière version qui a aussi été soumise à Simon

Baechler. Karine Fiola et Jean-François Lefebvre ont procédé à l'anonymisation des données du LSJML qui couvraient une année supplémentaire par rapport aux données utilisées au chapitre deux. Diane Séguin directrice de la section biologie au LSJML a rendu possibles tous ces travaux en acceptant de développer l'approche réseau au LSJML.

3.2 Résumé de l'article

Les délinquants absents des dossiers de police imposent une limite à la compréhension des comportements criminels dans les études criminologiques et pour les opérations policières. Des études récentes ont démontré que les concordances des banques de données ADN, incluant des inconnus, avaient un potentiel pour aborder ce problème. En utilisant les informations des dossiers criminels de ces concordances, nous démontrons que les individus inconnus présentent des divergences comparativement à leurs complices connus. En utilisant 19 années de concordances ADN récoltés au Québec, Canada, nous avons pu évaluer l'activité criminelle des connus et inconnus en utilisant leurs comportements de récidive et de co-délinquance ainsi que la diversification, le niveau de gravité, les types de délits et l'intermédiarité des individus. Nous avons découvert que les l448 inconnus étudiés présentent des caractéristiques et comportements différents par rapport aux individus connus. Ils sont plutôt récidivistes solitaires, avec moins de cas, moins de violence et plus spécialisés si plus actifs. Nos résultats sont en accord avec d'autres études qui démontrent que l'activité criminelle des inconnus est conforme à l'hypothèse d'exposition. Cette découverte, associée à une approche d'analyse en réseau, est novatrice et pourrait avoir un impact plus important que prévu sur les stratégies d'enquêtes et les politiques avec des implications pour le renseignement.

3.3 Article complet (Anglais): DNA Databanks as a Source of Information about the Criminal Behavior of individuals who have been linked to Crimes But Not Identified by Police

Abstract

Perpetrators of offenses missing from police files limit the capacity to reconstruct delinquent behaviors for criminological research and operational purposes. Recent studies show that forensic DNA databanks could offer the potential to address this problem, through large-scale analysis of DNA matches, many of which involve unidentified offenders. By using information associated with the criminal cases involved in DNA matches, we demonstrate that unknown individuals present discrepancy when compared with known accomplices. From 19 years of DNA match data from Québec, Canada, we were able to assess unknowns' and knowns' criminal activity using co-offending and repeat offending behavior, along with diversification index, crime seriousness scale, crime types and betweenness centrality. We find that the 1,448 unknowns studied are somehow marginal when compared to the known individuals. They are more solitary repeat offenders, doing fewer cases with less violence and becoming more specialized if they present a more active behavior. Our results are in accordance with other studies, showing that unknown individuals' criminal activity fits in accordance with the exposure hypothesis. These findings associated with a network approach are innovative and may have a larger impact than previously expected on investigation and policing with implications for forensic intelligence.

1. Introduction

For decades, DNA analysis in forensic laboratories around the world, have helped identify individuals who have committed offenses, as well as detect serial crimes. The introduction of this technology in the forensic lab has been seen as a real revolution, but although complex, it offers us a final result which is obtained by simple comparison of genetic profiles giving matches.

These matches are mostly of two types. Firstly, the genetic profile from a crime scene is compared against a convicted offender database. If a match occurs the name of the offender linked to his profile in the database is transmitted to the authorities. The use of a nominative database is the key largely responsible for the success obtained by DNA analysis. In the second type of matches, the genetic profiles from the crime scenes are compared and repeat offenders leaving their DNA on numerous crime scenes could then be detected as serial offenders. It should also be noted that a laboratory may analyze multiple samples from a single crime scene, giving genetic profiles that could potentially come from different individuals. In this later case, the DNA matches data shows their potential for detecting co-offending. Finally it should also be noted that two subtypes of results emerge following the comparison against an offender database; either the genetic profile is associated to an individual, as shown earlier, or no match appears and the genetic profile can however still testify to the presence of an unknown individual at the crime scene. In short, DNA data used in criminological studies are simply the result of an association between an offense and a genetic profile, the latter coming from a known or unknown individual.

A few criminological studies have used DNA matches considering only the identified individuals. Analysing numerous sexual assault kits left unprocessed over many years and using the US federal CODIS national DNA databank, Lovell *et al*. (2017) were able to unveil patterns in serial sex offenders. Using information from the cases, they were able to show that aggressions committed by stranger-only offenders were different when compared to non strangers offenders. A study that helps better understand sexual assault behaviors. In a more recent study, using the same type of data they were able to study the reoffending history of sexual offenders (Lovell *et al*. 2020). They show that three groups could emerge from the data. The first one describes the offenders as a “sexual specialist”, with more violent and sexual felonies in their reoffending. The second one is described as a “high volume generalist” showing reoffending felonies of various types and the third one, labeled as “low-volume”, are offenders more active in street and gang felonies as drug dealing.

In other studies, the attention is brought to DNA that testifies to the presence of a still unidentified individual at the crime scene, an individual who could be described as an unknown. Under the administration of a DNA databank, which generally revolves around the detection and transmission of identification to the authorities (RCMP annual report 2018-2019), such individuals remain behind, with the exception of those found in serial offending, which are of interest for tactical purposes. After decades, these unknowns count over a thousand in the data used in this study which were obtained from the Laboratoire de science judiciaire et de médecine légale (LSJML) of the Québec province in Canada, the forensic laboratory serving police investigation for a population of approximately 8 million.

The most original contribution that DNA match data can bring to the study of careers, relative to other criminological approaches, is that this data includes offenders who are known only by their DNA (De Moor, Vandevider and Vander Beken 2018a). Unknown individuals are part of the dark data of criminology which could take various aspects and is a very difficult type of data to assess, but here with individuals only detected by their DNA we now, at least, know the types of crimes they commit, their period of activity and their co-delinquency. Questions about these topics are usually covered in traditional criminology studies but they are based solely on police data or material from interviews with inmates (Huebner and Bynum 2016) and thus include only information from identified offenders. However, co-offending also involves relationships with offenders who remain unknown, i.e., who were not identified, or even detected, during police investigations, and who may differ from known individuals in relevant aspects (De Moor, Vandevider and Vander Beken 2018b). These unidentified co-offenders are betrayed by their DNA found at the crime scene. The presence of unknown offenders in DNA data attracted the attention of many researchers while some others have taken these unidentified individuals into consideration, albeit indirectly. For example, Ouellet and Bouchard (2017) interviewed 172 newly arrested inmates who had been criminally active in the 35 months before their arrest to test the exposure and competence hypotheses. According to the exposure hypothesis, the longer an individual

is criminally active, the greater his/her risk of being detected, a risk that also increases with the level of crime seriousness and violence (Blumstein and Cohen 1987; Blumstein, Cohen, Piquero and Visher, 2010). Evaluating the latter aspect requires determining a scale of crime seriousness and assigning different crimes a position in this hierarchy. The competence hypothesis states that arrested offenders are 'failed' offenders who are on average less skilled than perpetrators of criminal offenses who elude police arrest (Jacobs and Wright 2006). In a study of individuals active in lucrative drug markets, Ouellet and Bouchard (2017) looked at the duration, intensity, and level of involvement and found that individuals involved in this type of criminal activity are less likely to be identified than those who commit more violent crimes. They also show that the risk of being arrested correlates negatively with per-crime payoff, an approach that was also used as a proxy for competence of criminal efficiency by Tremblay and Morselli (2000). The results of the study by Ouellet and Bouchard (2017) suggest that unidentified individuals may be overrepresented in crime types associated with increased competence, such as marketing drugs, which, as they are associated with lower levels of violence, leave them less exposed to arrest. The unidentified could therefore be expected to be more active in secondary crime types (SEC) in our data (see below). By contrast, and not unexpectedly, the duration of criminal activities was a strong predictor of the risk of offender identification. Ouellet and Bouchard's study (2017) is one of the few studies that provide information on unidentified offenders using standard methods. However, although they interviewed offenders about their undetected criminal activity, these individuals had been identified by the time of the study.

As is the case in other areas of criminological research, the study of criminal careers and criminal activity is impeded by the so-called "dark figure" of crime. Unreported or undiscovered crimes and their unknown authors are an issue that has been recognized for decades (Biderman and Reiss 1967), as data sources used to study criminal trajectories and co-offending generally do not contain any data on unidentified individuals. Moreover, the panel and life history approaches, used in the Ouellet and Bouchard's study (2017), require a substantial amount of work to collect data. Forensic traces, in some cases in association with police files, may provide insights into longer

periods of criminal activity while requiring a much smaller investment of time (Rossy, Loset, Dessimoz and Ribaux, 2013).

Despite the various tools available to analyze criminal activity, criminological research faces the recurrent issue of missing data due to the illicit and concealed nature of crimes as well as selection biases in investigative and analysis methods (Sparrow 1991; Piquero and Weisburd 2010). Some strategies have been developed to account for the dark figure of crime such as using random retrieval of individuals in criminal networks (Smith and Moody 2013). Forensic methods can also be used to analyze material traces that provide information about offenders who are unknown to police services (Rossy *et al.* 2013; Rossy and Morselli 2017). This approach can provide information about unidentified offenders, such as the characteristics of their crimes (type, seriousness, frequency, diversity), their level of socialization (i.e., solitary or co-offending behaviors), and the length of their criminal careers as well as turning points in those careers (Blumstein and Cohen 1987; Pyrooz *et al*. 2017; Uggen and Staff 2004; Blomberg, Bales, and Piquero 2012).

Additional insights about unidentified individuals can be provided by social network analysis of forensic trace data (Rossy and Morselli 2017), creating an additional way to study the dynamic aspects of criminal behavior examined in crime theory (Blumstein *et al.* 1988). Forensic DNA match data has been shown to be an exceptional source of information about unidentified offenders. In a pioneer study, Lammers, Bernasco and Eiffers (2012) used forensic DNA match data from the Netherlands to study crimes committed by identified and unidentified individuals. They found that 65% of unidentified offenders will be arrested within 8 years of their second offense, while 35% will be arrested either in more than 8 years or never, leading them to ask “why is it that some offenders are arrested swiftly and others only after many years, or never?” Lammers and Bernasco (2013) analyzed DNA matches and crime location data and concluded that “the probability of arrest decreases as the number of police regions in which the offender commits his crime increases. The general use of DNA data for criminological studies has been investigated by De Moor, Vander Beken and Van Daele (2017), who showed that considering unidentified individuals has advantages for

intelligence purposes and provides a valuable alternative approach in criminological research on criminal behavior, despite some limitations. De Moor *et al*. (2018a) used DNA results to study the spatial distribution of felonies in Belgium committed by unidentified individuals. They noted a difference in the spatial distribution of the data depending on whether it came from a police source or a DNA database. They put forward the hypothesis that local policies could be at the origin of these observations since certain districts could be more active than others in the search for DNA traces. In a third study, the same researchers found that when they join DNA data on unidentified individuals to police data, they could study five times more crimes and uncover four times more networks than with police data alone (De Moor *et al*. 2018b). More recently, De Moor, Vandeviver and Vander Beken (2020) merged DNA data from unidentified individuals with police case file data on known offenders to assess the impact of ignoring vs. accounting for unidentified individuals in network analysis. They show that, if DNA is taken into account, 44% more co-offending relations are observed in comparison with database solely based on police-recorded crime data.

Recently, DNA match data from cases in Québec, Canada, was used in conjunction with social network analysis (SNA) to take advantage of DNA traces left by unidentified individuals imbedded in their co-offending networks (Lavergne, Baechler, Boivin, Jeuniaux, Fiola, Séguin, Lefebvre and Milot 2022). SNA results differed depending on whether unidentified individuals were included in the analyses. When unidentified individuals were included, some were found to be located in strategic positions in criminal networks, suggesting that investigations involving these networks might have been hampered by the lack of this information in police files. The results also showed that unidentified individuals have as many important social connections as those who have been identified, suggesting that the possibility of their presence should be taken into consideration in investigative policing and in assessing networking criminal activities from a forensic intelligence perspective. As shown in Figure 1, an unidentified offender could be the only person connecting two social subgroups of known offenders. Unless DNA match data is considered, a police investigator would be unaware of this connection between the two groups of identified individuals and might reach a different conclusion about a criminal event than if she/he had known about it.

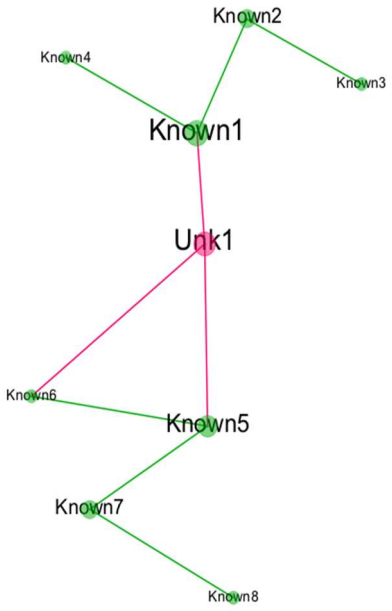


Figure 1: Example of a network where an unknown individual ("Unk1", red dot) stands at a bridge position between two groups of known individuals (green dots). The lines join individuals who did at least one offense together. Dot size is proportional to betweenness centrality (see Lavergne *et al*. 2022 for details)

In addition to information on social networks, DNA matches also contain information about the criminal activities of unidentified offenders (e.g., periods of activity, crime types and specializations). In the first study of criminal activities to incorporate DNA match data, Jeuniaux, Duboccage, Renard, Van Renterghen and Vanvooren (2016) analyzed the distribution of 13 crime types as a function of whether offenders were connected to network components (i.e. clusters of linked individuals emerging from co-offending cases) and the size of the networks involved, using DNA matches from 16 Belgians administrative districts. They found that 31% of all network components were associated with crimes committed in a single district, while 16% were associated with a single type of crime. However, their analysis did not distinguish whether the data involved was linked to unidentified individuals. These previous studies show the utility of examining DNA matches involving unidentified individuals, an

approach that could be used to address a variety of traditional research questions in criminological studies, but with a new perspective.

As an example, co-offending, which is a critical dimension of criminal career research in criminology, could benefit from the presence of these unidentified individuals found in DNA data. The fact that we can observe networked criminal activities in DNA data is a direct consequence of co-delinquency. However, without presuming to carry out a specific study of co-delinquency, we focus on an exploratory approach to find what could be observed with this type of data.

As the complexity of co-offending patterns raises fundamental questions, we basically keep in mind the presence of the unidentified individuals to compare them to the identified ones. In future studies results like those presented here could be integrated into in-depth research questions like those listed below but taking advantage of the addition of unidentified individuals. For example, under what circumstances do offenders team up with the same accomplices in multiple crimes (McGloin *et al.* 2008)? Does the likelihood of co-delinquency vary with type of crime (Gründ and Morselli 2017; Bright, Whelan and Ouellet 2022)? What factors determine how the criminal career of young offenders is pursued, e.g., as a member of gangs or acting alone (Pederson 2018)? What are the different types of adult criminal trajectories and possible increases in violent behaviors (Blumstein, Cohen, Das and Moitra 1988)? How do the characteristics of adolescent peer networks affect delinquency (Haynie 2001)? Creating study designs to address such questions is a challenge, as answering them requires considering numerous parameters, such as sources of data, age groups, and homophily, as well as determining whether data aggregation or individual analytic approaches are more effective.

All these questions and their related studies among many others have led to a great improvement in knowledge surrounding co-delinquency. Given in short, other studies show that co-offending relationships are rarely stable and are generally short-lived and transient (Weerman 2003). Arrest data from the Chicago police department over an 8-year period showed that 57.6% of individuals re-offended solo, 45.9% re-offending with varying co-offenders and a small percentage (5.3%) re-offending with

the same co-offender (Charette and Papachristos, 2017). Both Charette and Papachristos (2017) and McGloin *et al*. (2008) found that the criminal activity of individuals arrested more often by the police tended to involve more partners and was more likely to involve the same individuals as co-offenders. Additionally McGloin *et al*. (2009) also comes to the conclusion that during adolescence the majority of young people do not use the same accomplices. Gründ and Morselli (2017) investigated the overlap between co-offending networks and criminal activities and found that 10% of co-offender pairs commit multiple crimes, a value similar to the 11.7% documented by Charette and Papachristos (2017). Gründ and Morselli (2017) also reported that "almost half of those co-offending relationships are completely specialized around one crime type." Moreover, in the United Kingdom, for example, 10 to 20% of crimes are the work of more than one individual (Van Mastrigt and Farrington 2009) and 24% of all perpetrators of offenses are involved with co-offenders for at least part of their criminal career (Carrington 2002). Co-offending tends to be more common during adolescence and to decline with age or increased length of criminal career (Carrington 2002; McCord and Conway 2002; McGloin, Sullivan, Piquero and Bacon 2008; Reiss and Farrington 1991; Stolzenberg and D'alessio 2008). More recent studies concentrate on social network analysis (SNA) aspects as in Nieto, Davies and Borrion (2022) where the authors find that clustering (as triadic closures) could be a universal property of social networks.

Finally, co-offending is a complex behavior, the decision to co-offend is driven by numerous aspects of a protagonist's personality, social relation, (friendship, kinship), and their dynamics (Weaver and Fraser 2022), peer influence (Haynie, 2002), skills of various nature (Clare 2011), or lack of it as in the new field of cyber co-offending where IT skills are mandatory to perform (Kranenbarg 2022). Thus through a complex set of behaviors and influences, the co-offender will attempt to find the best social and personal advantages that he can obtain (Weerman 2003). The basic dynamics of co-offending could also be studied with DNA data while the more complex temporal trend of this behavior have been shown to be influenced by numerous and various life course events, such as parenthood (Pyrooz, McGloin and Decker 2017), or by a change in personal situation as scholarship, new interests coming up with maturation, peer

influence of new friends enhancing desistance or the fear of being caught (Knight and West 1975) just to mention a few.

Our research does not, and cannot, because of the limitation of the data, incorporate all of these aspects and distinctions of co-offending activities but opens the door to results obtained by essentially comparing the criminal activities of both types of offenders by addressing the following questions: i) Do identified and unidentified individuals participate in the same kinds of criminal activities? ii) Do their crime patterns (e.g., diversity, seriousness) change as a function of criminal behaviors undertaken solo or involving co-offenders? iii) Can crime patterns be correlated with features of the social networks with which the offenders are connected (e.g., network component density)? iv) Does the criminal behavior of unidentified individuals correlate with the predictions of the exposure and competence hypotheses?

By looking at such aspects of criminal activities, we hope to provide some answers to the question posed by Lammers *et al*. (2012) as to why some individuals who commit crimes remain unidentified by the police for long periods of time.

2. Methods

2.1 Data

The DNA match dataset for this study was obtained from the *Laboratoire de sciences judiciaires et de médecine légale* (LSJML), which is the governmental agency (ministère de la Sécurité publique) in charge of forensic analyses for the 31 police departments in Québec. The National DNA Databank of Canada (NDDB), under the administration of the Royal Canadian Mounted Police (RCMP), was put in force in 2000 and manages the reception, analysis and comparison of the genetic profiles of individuals who have been convicted of criminal offenses (the “convicted offender index” or COI). For more information on the criteria to accept DNA profiles and

management of the searches at the NDDB see Lavergne et al, (2022) and Milot, Lecomte, Germain and Crispino (2013).

The LSJML data used in this study are exclusively composed of DNA matches resulting from casework involving the local crime scene index (CSI). For known and unknown individuals, information related to their crime types comes exclusively from the LSJML CODIS databank administration, as we did not have access to police files and records to add more information about the known individuals. Thus, the types of activities of the two groups are comparable since their distribution would then only depend on the random deposit of individuals' DNA at crime scenes. The data collection extended from July 2000 to July 2019, a time frame large enough to observe numerous repeat and co-offenders without extending over several generations (Campana and Varese 2022). The total dataset involves 24,633 crime-scene samples obtained from 22,210 case files linked to 14,173 offenders of which 1,448 of whom were known only by DNA at the crime scene and, because they are unidentified, are therefore absent from police files and from the usual criminological studies. Because this study focused on crime types associated to repeat offenders and co-offenders, unique cases linked to one individual (n = 5,711) were eliminated. Using a restricted dataset allowed us to ensure that the results for identified and unidentified individuals were comparable and free from the biases that result from including the numerous offenders who are linked to a single identification and therefore provide no information on repeat and co-offending behaviors.

In short, data obtained from DNA matches consists of individuals linked to offenses for which DNA has been successfully detected and analyzed. Variables include case ID, individual numeric ID, crime type, the date of the event, known individual's date of birth, police bodies involved, dates of deposition in the databank and suspect/offender identification dates, as well as a few others. More sensitive data, such as names and other sample-related information were anonymized or removed by the LSJML prior to analyses.

2.2 Repeat offenders, co-offenders and crime types

An individual's criminal history can amount to a single offense or to several crimes (herein "repeat offenders"), perpetrated alone or with partners (Conway and McCord 2002). The career of a repeat offender can be a mix of solitary and co-offenses. Consequently, our dataset is organized in records (lines) uniquely defined by an offender-offense dyad. The repeat vs. non-repeat offender status is determined by whether or not an individual occurs in >1 data lines. For example, an offender associated with 3 criminal case files will have 3 data lines associated with him/her. A case file which appears in more than one line corresponds to a crime perpetrated by >1 offenders as the number of lines observed. At one extreme are repeat offenders who committed only solo offenses, while at the other extreme are those who have always worked in co-offending. Between these two extremes are individuals who exhibit various mixtures of solo and co-offending cases. When using the calculation as described above, the repeat offending count, represents the total set of an individual offenses done solo and co-offending. In this latter count, the number of individuals exclusively solo repeat offenders (with no co-offending), consist in the vast majority of the data. Similarly, the co-offending count includes individuals who only co-offend (with no solo offense) along with repeat offenders when co-offending. In the data descriptions and figures, the terms ``repeat offenders" or "co-offenders" refers to this latter type of calculation. In the other analysis where the term solo appears, solo and co-offending crimes are counted separately with no overlap.

Eight crime types are represented in the data, following their designation in the Canadian Criminal Code: (1) homicides (HO), including murders and attempted murders, (2) aggravated assaults (AA; with a weapon), (3) sexual assaults (SA), (4) robberies (ROB), (5) burglaries (BUR), (6) "secondary" crimes (SEC; e.g., drug or firearm possession, arson, impaired driving), (7) "rare secondary" crimes (RS; e.g., counterfeit money production and distribution), and (8) crimes for which DNA is "not admissible for depositions" (NAD) of DNA profiles in the National DNA Databank of Canada (e.g., theft under $ 5,000 CAD).Even if the offenses in this last category are not admissible at National level, they are part of our study since they represent a significant

proportion of minor offenses and still used to match cases at local level. These categories were uses to analyze the criminal activity of known vs. unknown individuals and as a function of their solitary vs. co-offending behavior, actually a kind of information that is absent from police files (De Moor *et al*. 2018b)

Analysis of the distribution of crime type data for the two groups is complex as the data are inter-dependent as a consequence of co-offending, which results in individuals, both identified and unidentified, appearing in more than one case file. Consequently, we analysed the differences between identified and unidentified individuals for their associated crime types through the permutations of the offender's status only, as identified or unidentified. That way, we retain the entire complex structure of the data composed of crime type, repeat offending and co-delinquency. This approach also respected the proportion of individuals in each group. From 1,000 iterations, we obtained the distribution of possible values for each crime type as related to whether the offender was identified or unidentified and whether the crime had been committed solo or had involved a co-offender. Comparison with observed (real world) data was made to determine if these values were over, under, or equal to random distribution.

2.3 Statistical analyses

2.3.1 Distribution of knowns and unknowns in offender groups

We determined the proportion of offender-offense dyads involving known and unknown individuals for three categorical factors: repeat vs. non-repeat, solitary vs. co-offenders, crime types. To assess the significance of differences observed, we randomly permuted the status (known or unknown) for cases in the DNA data file 1,000 times, while maintaining the number of crimes, the total number of identified and unidentified offenders, and the number of accomplices for co-offending per crimes. These permutations results gave us the random distribution of individuals that could be expected as being around 25% for each group.

2.3.2 Crime specialization

Crime specialization, and its counterpart, diversification, can be assessed by a number of analytical approaches. Sullivan, McGloin, Ray and Caudy (2009) compared four of them applied to the same data set: forward specialization coefficient (FSC), diversity index (D_i), latent class analysis (LCA), and the multilevel item response theory-based approach (ITR). They show that LCA and IRT are more appropriate for investigations associated with the subculture of violence, FSC is more appropriate for examining crime types in sequence, and D_i to measure the diversity at the individual level. The latter is thus relevant to compare criminal activities between known and unknown individuals (Sullivan *et al.* 2009). D_i is defined as "the probability that randomly paired members of a population will be different on a specified characteristic" (Lieberson 1969) or, within the criminological context, as "the probability that any two offenses drawn randomly from a given individual's set of offenses belong to two different offending categories" (Piquero, Paternoster, Mazerolle, Brame and Dean 1999; Mazerolle, Brame, Paternoster, Piquero and Dean 2000). Following Agresti, A. and Agresti B. (1978), we calculated D_i as:

$$D_i = 1 - \sum_{i=1}^{k} pi^2 \qquad (1)$$

Where p_i is the proportion of crime type i. D_i ranges from 0, i.e., when all crimes committed by an individual belong to the same type, to a maximum value $D_{i,\max} = (k\text{-}1)/k$ bounded by the amount of crime categories (k) seen for that individual (Agresti, A. and Agresti B. 1978).

2.3.3 Crime Seriousness

Beyond their number, the seriousness of crimes committed by an individual could also differ between known and unknown offenders. The measurement of seriousness requires some qualitative scale, such as those used in self-reported delinquency

interviews (Elliott, Huizinga and Ageton 1985; Piquero, Macintosh and Hickman 2002) or determined by a social survey (Wolfgang 1985), since crime seriousness can depend on social-cultural factors (Thurstone 1927). With DNA matches we implement the global crime seriousness scale by evaluating it with two quantitative parameters. One challenge is to determine a scale of crime seriousness and to assign different crimes along this hierarchy. Thus, for each offender, we weighed the number of crimes and their seriousness, using an adapted log scale (Boivin 2021). It is generally accepted in society that a burglary is less serious than a homicide. Inspired by Conrad K.J., Riley, Conrad K. M.; Chan and Dennis (2010), we gave seriousness scores (S) to crime types on a power scale ($S=2^x$, where $x \in \{0,1,2,3,4,5,6,7,8,9\}$). Non-admissible (NAD), rare secondary (RS), and secondary (SEC) offenses were scored respectively S=1, 2, and 4. These correspond to the most minor offenses. Burglaries (BUR) were given a slightly higher score (S=8) due to their more intrusive nature. Most serious crimes, i.e. those involving some form of violence against people, were given increasingly high scores, as a function of their general degree of violence relative to each other: S=64, 128, 256, and 512, respectively for robberies (ROB), armed assaults (AA), sexual assaults (SA), and homicides (HO). Seriousness scores were then summed up for each offender. For example, an individual who did 50 burglaries would have a total score of 400, and one with a single homicide a total score of 512.

Seriousness is also addressed using a network approach to analyze DNA matches data gives the opportunity to work out comparisons with parameters like the degree values and the component density obtained from social network analysis (SNA). Of the four parameters that were analyzed by Lavergne *et al.* (2022), we focused on the degree values evaluated for components in a unimodal format, where every link is a connection between two individuals sharing a criminal activity (i.e., one or more crimes done in co-offending). Therefore the degree values represent the global number of accomplices counted for an individual in the course of his career. For comparisons between known and unknown individuals, crime seriousness is assessed against the degree value as the number of accomplices.

2.3.4 Criminal activity timespan and crime frequency

From criminal event dates reported in LSJML data, we calculated the time, in days, elapsed between the earliest and latest criminal event recorded for each individual. Values were then transformed in years (e.g., 2.05 or 2.87 years). A value of 0.00274 years, i.e. one day, was attributed to individuals with either a unique or multiple offenses that all occurred on the same day (a rare situation). We took the data from the beginning of 2000 up to March 2016, three years before the last case registered. We then compared the average period of criminal activity for the known and unknown individuals up to three years and more. We also calculated an individual's activity frequency as the number of crimes he/she committed per year, which translate as crime frequency.

2.3.5 Co-offending and social network analysis (SNA) parameters

To assess how the position of an individual in a criminal social network may be related to its involvement in co-offending (or vice versa), we measured the correlation between co-offending and betweenness centrality for the known and unknown individuals. The latter measures the position of an individual in the criminal social network, measured as "how often this given individual falls along the shortest path between two other individuals" (Borgatti, Everett and Johnson 2018; Freeman 2017) (Fig. S1). This parameter takes a zero (0) value for all peripheral individuals in a network while more central individuals present various higher values depending on their position in relation to the other isolated ones surrounding them. The higher the number of isolated individuals surrounding a specific one, the higher value of betweenness will be obtained.

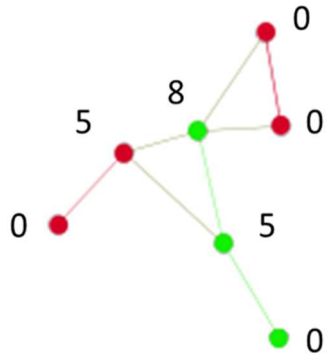


Figure S1: A component composed of three knowns (green dot) and four unknowns (red dot) showing betweenness centrality values for every individual.

The co-offending status of the individuals is evaluated by the number of cases they have done with accomplices. We defined three groups, the ones with at least one case (>0 cases), the one with at least two cases (>1) and the latter group with three cases and more (>3).In addition we examined the relationship between Di and the component density (*d*) as in McGloin and Piquero (2010). From the criminal social network study published by Lavergne *et al.* (2022) were 54 components made of ≥5 interconnected individuals have been selected, the density was evaluated as in Bichler (2019):

$$d = l/(N(N-1))/2 \qquad (2)$$

Where *l* is the number of links observed between the *N* nodes (individuals) of the component. Then, we examined the relationship between *d* for a component and the average *Di* of individuals in the component.

3. Results

3.1 Total offense count

The average number of criminal events per individual is nearly identical for known (2.5) and unknown (2.4) offenders (n=17,471 events). Unknowns represent 20.6%

(1448/7014) of individuals in the dataset (Table 1). The total count for each crime type was >900, except those not admissible for depositions into the DNA databank (NAD=429): HO=1195, AA=1303, SA=1067, ROB=1813, BUR=9685, SEC=1028, RS=951. Thus, all types were well represented in the data. Burglaries, generally being by far the most frequent crimes seen in criminological studies (Bright *et al*. 2022), thes same is true in forensic DNA data with their proportion of 55.4%.

3.2 Distribution of knowns and unknowns in offender groups

The proportion of individuals who did multiple offenses, and including at least one in co-offending, is 23% (1298/5566) for knowns and 32% (470/1448) for unknowns (Table 1). As expected, the average value obtained from the permutations is 25%, which is represented by the values of 1405/5566 for the knowns and 364/1448 for the unknowns; Fig. S2).

The difference between the groups becomes very large when keeping only co-offenders (repeat or not): 41% (1298/3188) for knowns and 95% (470/495) for unknowns (Table 1). This suggests that nearly all unknown individuals who committed at least one offense with accomplices are recidivists. By contrast, the majority (59%) (1890/3188) of known co-offenders were involved in a single (detected) crime while co-offending which is the consequence of the DNA being compared to an offender database giving numerous identification over time.

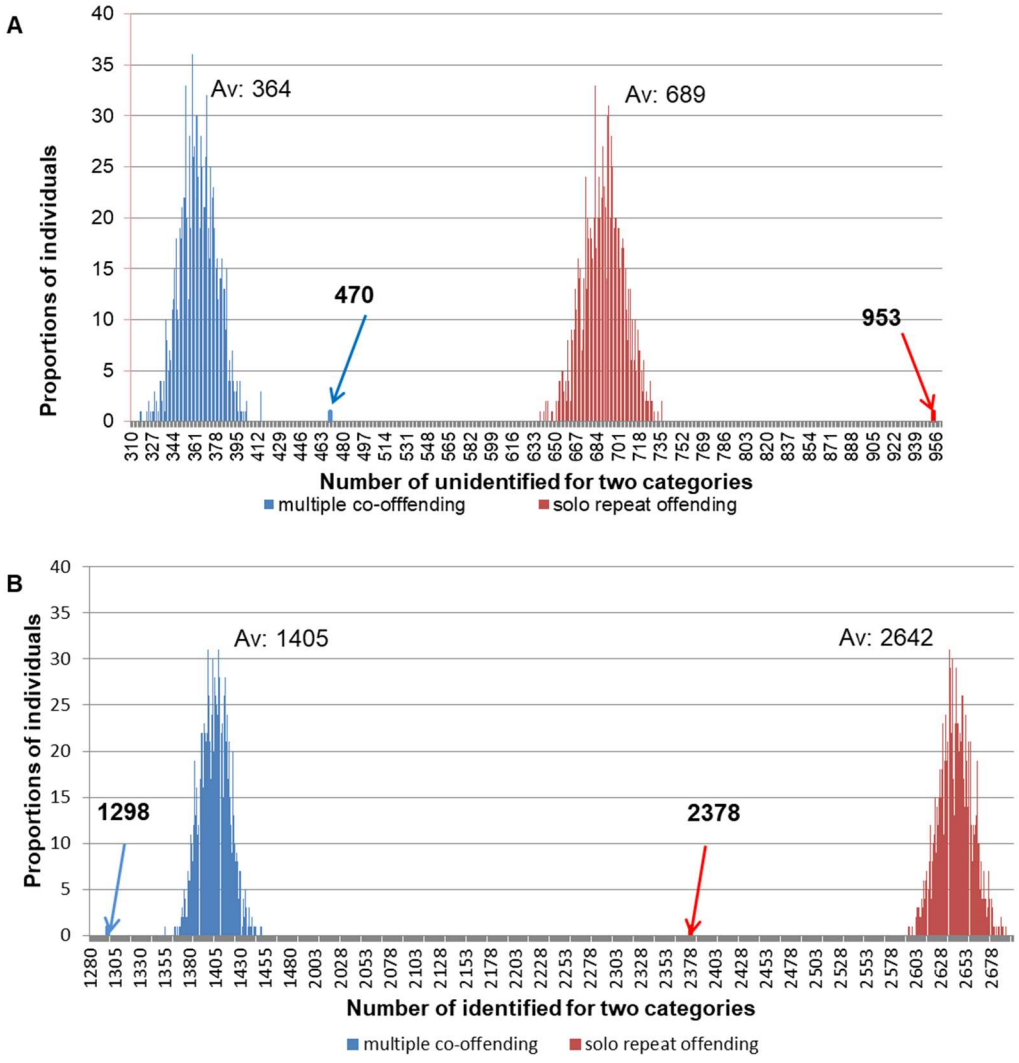


Figure S2: Distribution of the random iterations for two behaviors categories for the unknowns (A) and the knows (B). The average values obtained from the iterations are shown at top of the distributions, the observed values from the data in bold, positioned with arrows.

Among the knowns with at least two offenses, 43% (2378/5566) are found in the solo category, which is close to the average of 47% (2642/5566) expected for both groups from the permutations (Table 1, Fig. S2). This proportion rises to 66% for unknowns (953/1448) in the data. As shown in Fig. 2A, unknown offenders with exactly two offenses are largely responsible for this higher proportion, while those with more

offenses keep a lower profile (22%) compared to the knowns, 44% of whom having >2 criminal cases in their history.

Table 1: Distribution of individuals in Québec forensic DNA match data according to their identification status (known vs. unknown) and their co-offending activity. Overall, the 7,014 individuals were involved in 17,471 criminal events.

Status	**Nb. of reported offenses**	**At least one crime done with co-offender(s)**		**Total**
		yes	no	
Known offender	1	1890	N/A[a]	1890
	>1	1298	2378	3676
	Total	3188	2378	5566
Unknown offender	1	25[b]	N/A[a]	25
	>1	470	953	1423
	Total	495	953	1448
Total		3683	3331	7014

a: For reasons explained in the limitations section, solitary identified offenders linked to a single case were set aside to be in accordance with the unidentified for which this data is not available; hence no count is reported for them. b: Unidentified being observed with a single case in Québec since these individuals are repeat offenders with at least one other case in another Canadian province. This value is under-represented since unidentified co-offenders on a single case do not appear in the database.

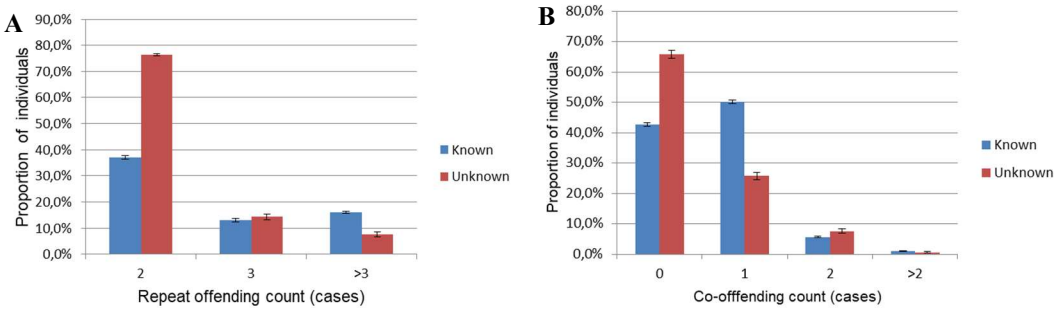


Figure 2: Proportion of individuals in the two groups (± SE.) as a function of the number of repeat offending cases (panel A) and co-offending cases (panel B). In panel A knowns sum to 66% and unknown to 98% with solo cases not included (see Table 1). In panel B both knowns and unknowns sum to 100% with solo cases included (under 0 co-offending).

As shown in Fig. 2B, appearing under zero co-offending, a noticeable difference is observed for the solitary repeat offenders with a proportion much higher for unknowns (66%) than knowns (43%). By contrast, 50% of knowns are involved in a single co-offending case, a proportion that decreases to 25% for unknowns. Finally, less than 10% of offenders are associated with 2 or more co-offending cases, both for knowns and unknowns.

3.3 Distribution of knowns and unknowns: crime types

The frequency distribution of crime types differ between knowns and unknowns (Table 2). Unknowns are less active in major crimes as seen in homicide, armed assault, robbery and burglary when they are solitary offenders. Still, no variation is observed between knowns and unknowns in the sexual assault category which is mostly observed in solitary offending. Nevertheless, when we sum up the number of those serious crimes for the two groups of individuals, the solitary unknowns get 166 cases against 264 for the knowns (values from cases/100 individuals). This pattern is not observed in the co-offending cases where the two groups share about the same number of cases (102 for the unknowns vs 98 for the knowns).

Table 2: Distribution of crime types among known and unknown offenders and as a function of their social behavior.

Offense (count)	Observed data (total counts)				Values per 100 individuals			
	Solitary offenders		**Co-offenders**		**Solitary offenders**		**Co-offenders**	
	known (N=3499)	unknown (N=1338)	known (N=3189)	unknown (N=495)	known	unknown	known	unknown
Homicide	278	60	756	101	8	4	24	20
Armed assault	498	135	603	67	14	10	19	14
Sexual assault	579	231	226	31	17	17	7	6
Robbery	967	174	599	73	28	13	19	15
Burglary	6897	1626	929	233	197	122	29	47
secondary	531	232	217	48	15	17	7	10
Rare secondary	321	203	352	75	9	15	11	15
Non admissible	226	143	53	7	6	11	2	1
Total	10297	2804	3735	635	294	209	117	128

However the homicide numbers in the co-offending group are the second-highest values for knowns and unknowns and show higher values compared to the solitary offenders homicide cases. On the other hand, for the three groups of minor felonies (SEC, RS and NAD) the sum-up case numbers for the unknowns always outdo the ones for the knowns. For the solitary cases, the unknowns gets 43 against 30 for the knowns and for the co-offending cases the same comparison gives 26 to 20. Globally, it could also be observed that burglaries are the most frequent felonies across all groups while robbery and armed assault have intermediate values. Less severe crime types (secondary, rare secondary, non-admissible) are those where, according to their proportion, unknowns are more active, with the exception of the non-admissible cases in co-offending and the burglary in solo offending.

Table 3: Percentage of 1,000 permutations that shows a larger number of cases of a given type compared to real values for the unknowns (HO: homicides, AA: aggravated assaults,, SA: sexual assaults, ROB: robberies, BUR: burglaries, SEC: secondary offenses, RS: rare secondary offenses, NAD: offenses non-admissible for depositions in the DNA databank. A 100% value means that the real value is smaller than those of all 1,000 permutations, while a 0% value indicates that the real value is greater than those of all permutations. The corresponding percentage of known offenders is thus 100 minus the value of unknowns.

	HO	AA	SA	ROB	BUR	SEC	RS	NAD
Solo cases	86.4	32.1	0.0	100	94.4	0.0	0.0	0.0
Co-offend. cases	100	100	100	100	63.4	80.0	93.10	95.0
Total cases	100	100	0.3	100	94.5	0.0	0.0	0.0

Unknowns' criminal activity differs markedly according to felony types. For three of the eight crime types, 100% of permutations result in larger counts than real values observed (Table 3, total cases) showing that these individuals are, in respect of their proportion compared to the known offenders, less active for homicide, aggravated assaults and robberies, when compared to the expected values observed with the permuted individuals. On the other hand, still according to their proportions with regards to the known individuals' permutations show that unknowns are more active in secondary, rare secondary and non-admissible felonies.

3.4 Crime Seriousness

Using the crime seriousness scale described above, the average value of the offenses done solo is 133 (±208) for the knowns and 90 (±149), i.e., 32% lower, for the unknowns. For the cases done in co-offending, average seriousness for the knowns is 194 (±218) and 161 (±224), i.e., 17% lower, for the unknowns. Plotting crime seriousness against the number of offenses committed by an individual shows that, when compared to the average, knowns' crime seriousness increases when the number of offenses increases, which is true for solitary and co-offending cases (Fig. 3). The reverse is true for unknowns. Solo unknown offenders show a slighter decrease from 90 to 74

while the drop is more prominent in the co-offending cases type which starts from 161 to end at 88.

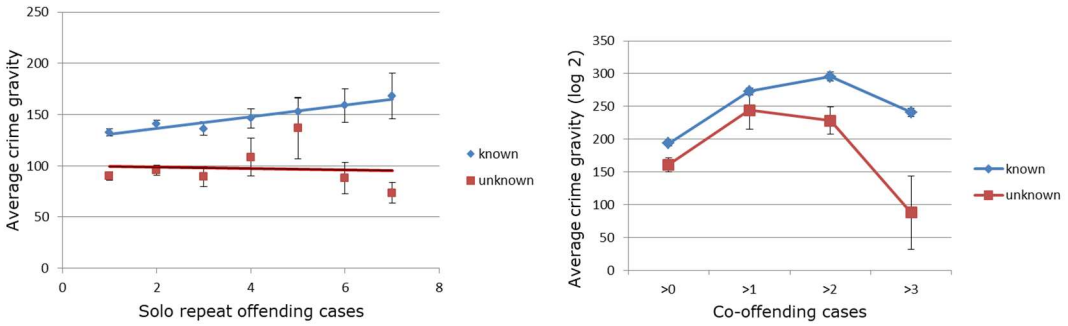


Figure 3: Average crime seriousness (± SE) for the known and unknown individuals. Panel A: cases carried out in solitary (N=3500 and 1338 offenses for knowns and unknowns, respectively). Panel B: for cases carried out in co-offending (Total N values: knowns (3185), unknowns (495). The first value under >0 represent the total set of data

3.5 Social network analysis

Figure 4A shows a positive correlation between average crime seriousness for known and unknown offenders and their degree value in social network components. The same observation has been made for crime specialisation (Di)(Fig. 4B). Crime seriousness and diversification thus increases with the number of accomplices that an individual has had in his/her delinquent life. Unknown individuals show a higher average of crime seriousness and diversity when their crime involve more accomplices.

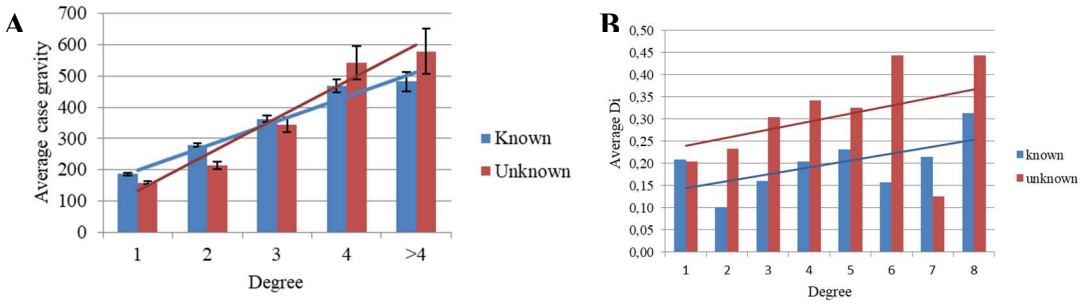


Figure 4: Average crime seriousness (± SE) according to degree value as evaluated from SNA where the degree is the number of connections an individual has with others.

3.6 Crime specialization

The criminal activity of unknown offenders is more diversified (mean D_i=0.22 ±0.25; max: 0.75) than that of known offenders (mean D_i=0.16 ±0.23; max: 0.8). Most of this difference is explained by co-offending activity (unknowns: mean D_i=0.33 ±0.24; knowns: mean D_i=0.15 ±0.24; Fig. 5), as there is almost no difference for the solo offending activities. On the other hand, in the repeat offender subgroup, both knowns and unknowns become more specialized as the number of their offenses increases (Fig. 6A). The same is observed when co-offending increases albeit less markedly for known offenders (Fig. 6B).

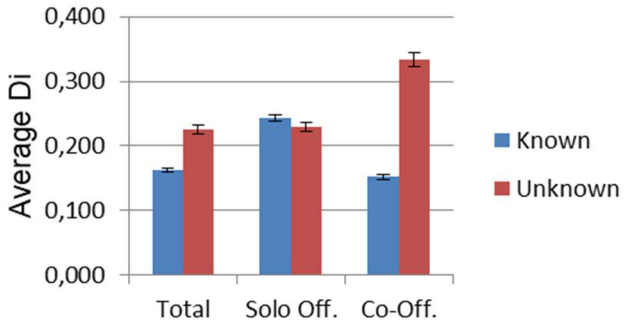


Figure 5: Average D_i (± SE) for known and unknown offenders for the total dataset and for individuals showing solitary offending and co-offending behavior.

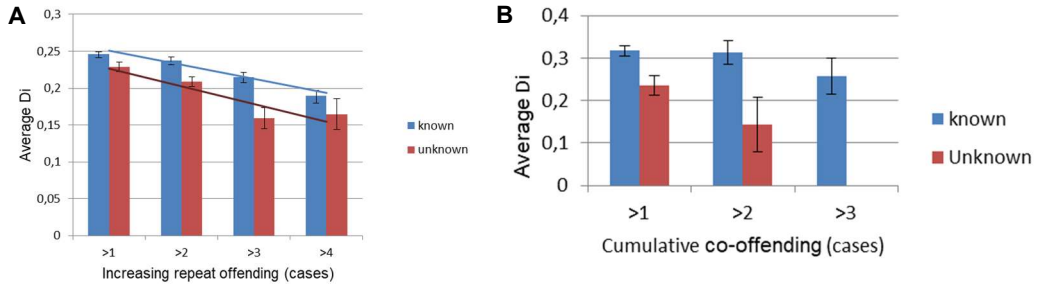


Figure 6: Average D_i (± SE) for repeat offending (panel A) and co-offending (panel B) According to the cumulative number of cases. Note that the average D_i for unknown when co-offending in > 3 cases is zero (see Fig.5) and that repeat offending includes co-offending cases when these account in the repeat offending list.

Another way to visualize the increasing specialization with the number of offenses is to look at the proportion of individuals with a D_i of zero, i.e., those specialized in a single crime type (Fig. 7). A different pattern was uncovered in the co-offending subgroup where the unknowns reached 100% with >3 cases while the known gets less specialized as the number of co-offending cases increases.

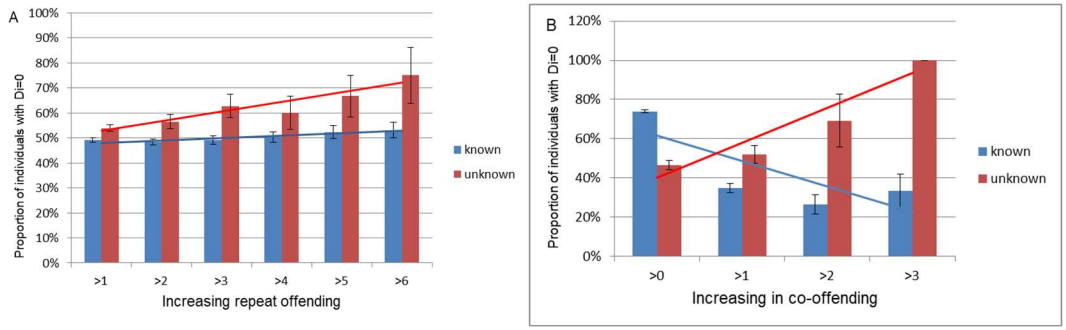


Figure 7: Proportion of Di=0. Panel A: for cumulative repeat offending cases. Panel B: for cumulative co-offending cases. All unknowns are fully specialized when active in >3 co-offending cases.

3.7 Criminal activity timespan and crime frequency

The time elapsed between the earliest and the latest crime of offenders exposes another difference between knowns and unknowns. The timespan of an offender's criminal activity ranges from one day to 26.9 years (mean=4.7, SD=±4.9, med= 3.08). We cannot exclude that the true criminal lifespan is longer than this time period, since an individual may have committed undetected offenses or may have not finished their criminal career as of 2019 (i.e., the upper limit of our data set). Strikingly, a high proportion of offenders show a criminal activity timespan equal or shorter than one year (49% and 37% of knowns and unknowns, respectively; Fig. 8). For a longer timespan (≥1 years), 62% of the criminal activities are associated with the unknown and 51% of the knowns). This translates into different per-year crime frequencies for all individuals (Table 4). For the known and unknown individuals, the value has been evaluated firstly for the complete data, secondly for the individual active only within a year and thirdly for those active over a year.

A salient pattern is that known individuals are more than twice active within a year compared to unknown individuals. For the longer period of activities both groups are closer but as the unknowns spread their activities over numerous years, as seen in Fig. 8, they kept a lower frequency in crime per year.

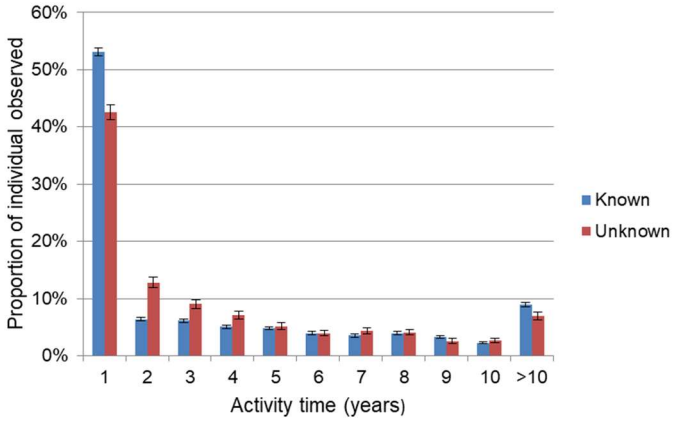


Figure 8: Percentage of known and unknown individuals according to their criminal activity windows in years. (1; for up to a year, 2; from one up to 2 years, and so on)

Table 4: Average crime/year according to crime activity windows as described with more details in Fig. 8.

	Criminal activity window		
Identity	All	<= 1 year	>1 year
Known	140 (±179)	262 (±16)	0,85 (±0,92)
Unknown	46 (±125)	107 (±10)	0,79 (±0,89)

3.8 Co-offending and SNA parameters

Figure 9 shows an increase in the number of co-offending cases as betweenness centrality (from SNA) increases. The trend is much more pronounced for knowns, with a slope of 16.8 than for unknown offenders with a slope value of 4.7.

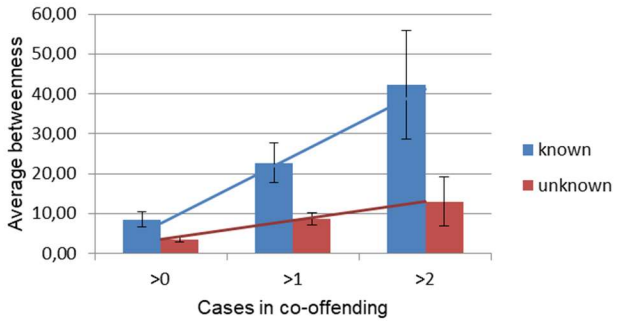


Figure 9: Average betweenness centrality as a function of the number of co-offending cases (S.E shown). Using the co-offending information gives a fine-tuning on the observation of Lavergne *et al.* (2022) that unknowns were as central as knowns in network components.

When the average Di is plotted against the component density, we observe a negative correlation for the knowns (r= -0.324), similar to the one obtained in McGloin and Piquero (2010). (r= -0.159 and -0.280 respectively from the group average and the Tobit regression model). On the other hand, our unknown individuals show almost no correlation with a value of r= -0.008.

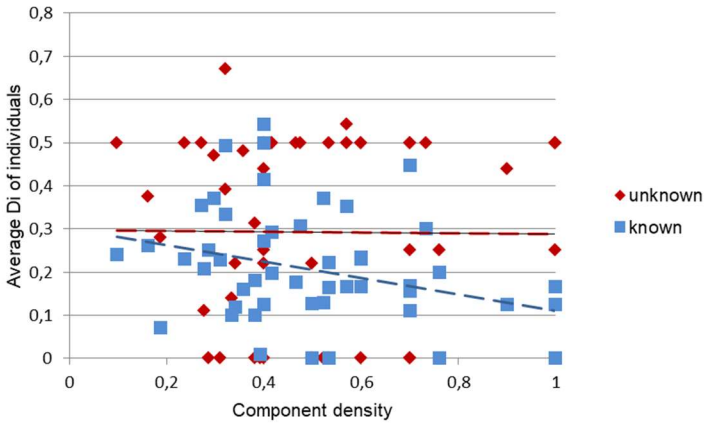


Figure 10: Average Di value of the individuals in each component according to the component density. (54 components, N= 330 knowns and 91 unknowns)

4. Discussion

The study of criminal activities has a long history spanning over a century. Various methodological approaches have been used, sometimes fuelling controversy owing to the challenge of characterizing such a complex reality (Piquero, Farrington and Blumstein 2003).Tracking unknown criminal offenders and their impact on criminality is a major challenge in understanding this complex reality. The information available about them is extremely scarce, as coined in the quite à propos expression « dark side of criminality ». In a previous study (Lavergne *et al.* 2022) it was shown how forensic DNA matches represent an exceptional source of information to document the place occupied by unknown offenders in criminal social networks. A few recent studies in other countries likewise used DNA match data as a novel source to expand our knowledge about unidentified criminals and their characteristics (Lammers 2014; Lammers *et al.* 2012; Jeuniaux *et al.* 2016; De Moor *et al.* 2017; 2020). In the present study, we further uncovered and quantified differences in the criminal careers of knowns and unknowns. One salient finding is that unknown individuals show specific criminal activity patterns that could help them elude arrest.

Missing data can take various forms and have several impacts for criminological studies. A number of approaches have been proposed to evaluate these impacts mostly using random removal of data (Borgatti, Carley and Krackhardt 2006; Smith and Moody 2013; Smith, Moody and Morgan 2017; Krause, Huisman, Steglich and Snijders 2020; De Moor *et al.* 2020). On the other hand, working with DNA left on crime scenes gives access to a large number of individuals absent from police files for which types, location and other information about criminal activities is available. It's a rare source from which unknown individuals, usually falling in the missing data, are tractable. The present study builds on the detailed criminal information found in the LSJML data to compare the criminal activity of known and unknown offenders.

4.1 Exposure and competence hypotheses

A noticeable larger proportion of unknowns than knowns commit only solo crimes over their criminal career. This observation does not exclude the possibility of co-offending activities inside this group which might be not detectable in our DNA data. Nevertheless, assuming that being active with accomplices - some of which being possibly known to the police - exposes an offender to more visibility or denunciation, this result provides support for the exposure hypothesis, which predicts that the longer an individual is criminally active, the greater his/her risk of being detected. The hypothesis also takes into account that some crimes are more at risk than others (Ouellet and Bouchard 2017). Moreover, 22% of unknowns committed more than two crimes, compared to 44% for the knowns, further suggesting a lower exposure of the former.

The competence hypothesis, often associated with the exposure hypothesis, remains understudied in criminal careers (Ouellet and Bouchard 2007) This hypothesis suggests that the offenders included in traditional criminological research are those who failed and are behind the bars, while those who are skilled remain unidentified by the police. Competence in crime is a complex issue that could be addressed using many life aspects of the criminal individual and his human capital. To potentially reduce the complexity and limitations of such studies, an approach of merging criminal earnings with offending frequency in order to create a measure of payoffs per crime defines such a criminal efficiency (Tremblay and Morselli 2000). Thus basic criminal earning could define some level of criminal efficiency and, as Tremblay and Morselli (2000) noticed, "*...some offenders are more skilled than others in increasing their monthly criminal earnings without, at the same time, significantly raising their risks of apprehension*"., Our results show that the unidentified are more active in secondary offenses such as drug and firearm markets, where the rates of identification and arrest are likely to be lower. Since burglars with better income have a clearance rate negatively associated with the crime trip length (Vandeviver, Van Daele and Vander Beken 2015), it is counter-intuitive that dealers which are more sedentary, selling within a radius of 7 km from their residence (Morselli and Royer 2008), could also earn good income. An expected lower arrest rate for drug dealing could be explained by two factor, first they

are based on close reliable networks (Morselli and Royer 2008) and second, to the fact that enforcement of drug market are difficult due to the size of their networks and the budget and organization needed to counter them (Baveja, Batta,Caulking and Karwan 1993). On the contrary, unlike crimes that do not attract attention, homicides have one of the highest solve rates (Alderden and Lavery 2007).

Another salient finding is the difference in the distribution of the types of crimes committed by known vs. unknown offenders. Our analysis uncovered three patterns in that matter. Firstly, unknowns are less active in major crimes against the person, whether they work in solo or with accomplices, with the exception of sexual assaults where they slightly preponderate. Actually, it has been demonstrated that felons more active in violent crimes will receive more attention from police forces (Heller and McEwen 1973), as this could be a possible police bias toward a higher presence of unknowns in less violent crimes. In that matter, it may thus be surprising that unknowns are predominant in sexual assaults (ref. to Tab. 3) which are violent crimes. Yet, it has been demonstrated that this type of aggression is typically committed alone (Van Mastrigt and Farrington 2009), which ultimately supports the exposure effect due to the violence of the crime.

Secondly, unknowns commit fewer burglaries when they act in solo, but more crimes of this type when co-offending, which could result from a peer influence effect (Akers, Ronald L. 2017). Being part of the social learning theory, the influence of peers could be described as; “Adolescents who report that their friends are delinquent tend to report higher levels of delinquency than adolescents with fewer or no delinquent friends” (Haynie 2002).

Thirdly, by contrast unknowns commit more secondary and rare secondary crimes, which are less violent and comprise the so-called small felonies such as drug dealing, car theft, extortion, and others. These observations further support the exposure hypothesis: as shown by Ouellet and Bouchard (2017), criminals that are more active in profitable

market crimes such as drug dealing, car theft and extortion are more likely to avoid arrests. These correspond exactly to the "(rare) secondary offense" category in the current study. Moreover, these crimes are generally more profitable financially, a feature associated with the competence hypothesis (Tremblay and Morselli 2000) which states that arrested offenders are 'failed' offenders who are on average less skilled than perpetrators who elude police arrest (Jacobs and Wright 2006).

Thus, by being more active in crime types where you can earn incomes, unknowns are perpetrators who are more likely to avoid detection as explained by the competence hypothesis. An example of competence could be seen in burglars that do not decide randomly on a target house, their decision is based on a rational decision-making process (Vanderviden at al 2015) which in turn could be seen as a specific competence in burglaries (Clare 2011). One could hypothesize that unidentified individuals make informed choices about their criminal activities. Moreover, The above mentioned findings could be seen as information supporting intelligence to better understand the unknown individuals along with supporting policing and operationalizations. By knowing better what these individuals usually commit as crimes, it would be possible to be more effective in tracing them.

When comparing the proportions of co-offending data observed for the knowns (57%) and the unknowns (34%) in DNA matches data (Table 1), to other current criminological studies using other types of data than DNA, they are about the same. It's 20 to 44% that was observed in Carrington (2002), 30% in Van Mastrigt and Farrington (2009), 35% in Hodgson (2007) and 40% in McCord and Conway (2002) among young offenders. The discrepancies between studies likely partly reflect the variation among jurisdictions in the patterns and type of criminality under investigation. In other studies the methodology used could pinpoint more precise information about co-offending as in Gründ and Morselli (2017) and Charette and Papachristos (2017) where reuse of accomplices and specialization has been under investigation.

4.2 Crime specialization

It has been shown that crime specialization may be difficult to assess using large aggregated data while it may be more commonly observed in shorter criminal careers (Sullivan, McGloin, Pratt and Piquero 2006), or at earlier stages of longer ones (Blumstein *et al.* 1988). It may also change with age, gender, or co-offending (Mazerolle *et al.* 2000; Charette and Papachristos 2017; Gründ and Morselli 2017). Using the global data of this study, one could observe that both groups had about the same level of specialization as repeat offenders, while showing a notable difference when co-offending. On the other hand, much more discrepancy could be observed between unknowns and knowns when looking more closely within their level of repeat and co-offending. More precisely, unknowns get more specialized if involved in more repeat and co-offending cases. Again a possible peer influence in the first felonies, as conversely, individuals in both groups get more specialized as they are active in more and more solo cases in longer criminal careers (Clare 2011). An observation that has also been reported for co-offending (Gründ and Morselli 2017).The pattern differs strongly when looking only at fully specialized individuals (Di=0). Their proportion increases more rapidly with the number of offenses for unknowns than for knowns, when acting in solo. Indeed, all unknowns associated with more than 3 cases are totally specialized. Clare (2011) suggests that a higher specialization in burglaries relates with skills and success in obtaining benefits from crimes while avoiding arrest, as such offenders would be more efficient and aware of risks. This is further supported by the study of Lammers *et al.* (2012), who showed from DNA match data that unknown individuals committing the same crime multiple times had a lower probability of arrest. On the other hand, the presence of numerous accomplices produces an opposite effect by increasing crime diversity (Di) which induces more exposure according to the competence/exposure hypothesis.

The diversification index is known to have two limitations. Firstly, it gives the degree of specialization with no relation to the crime type. Secondly, since D_i cannot, by definition, be calculated for individuals with a single offense, the latter were excluded

from this analysis. Nevertheless, as stated by Sullivan *et al.* (2009) "D_i has fewer conceptual and measurement constraints than does the FSC…" which gives D_i more freedom to uncover various specialization patterns at the individual level (Mazerolle *et al.* 2000). The dependency of D_i on the number of crime categories (*k)* led some authors to suggest a normalization of its values ($D_{i,\text{norm}}=D_i/D_{i,\text{max}}$), so that $D_{i,norm}$ ranges would be {0,1} (Amemiya 1963). However, other authors warn against normalization because it changes the amount of information conveyed by this metric, as well as its interpretation (Agresti A. and Agresti B. 1978; Sullivan *et al.* 2006; Budescu, D. and Budescu, M. 2012; Lieberson 1969).

4.3 Crime seriousness

Whether known and unknown offenders differ by the seriousness of their crime remains an unexplored question in criminology (De Moor, S., Vander Beken, T. and Van Daele, S. 2017). For both known solo and co-offenders, crime seriousness is positively correlated with the number of offenses, while the pattern for the unknown group is non-linear. For both groups, crime seriousness increases with the number or accomplices. When the number of their accomplices raises ≥3, both groups show the same level of crime seriousness. As aforementioned, solo acting unknowns, commit on average, less serious offenses than the knowns. As the number of their accomplice's increases, a peer ripple effect might be at play, especially with known accomplices active in more violent crimes. It is generally recognized that the influence of peers encourages individuals to commit more crimes and more violent ones (Zimring 1981; Haynie 2001; Conway and McCord 2002; Hodgson 2007; McGloin and Piquero 2009). Moreover, the exposure hypothesis, mentioned above, is linked to the crime seriousness which in turns is highly related to social and cultural environment as for gender perception (Rossi, Simpson and Miller 1985) and more specifically for sexual assaults (Kwan Y., Ip and Kwan P. 2000), educational attainment (Rossi, Waite, Bose and Berk 1974) and drug-related felonies would scale very differently depending on whom or where surveys are addressed (Kwan, et al. 2000).

4.4 Criminal activity timespan and crime frequency

This study shows that known and unknown offenders in Québec are rather equally criminally active. However, cases associated with unknowns are more spread over time than for knowns. As a result, unknowns appear less visible in global crime data, as in lab reports and DNA matches sent to investigators, giving them less exposure. Would they appear in crime statistics, at any defined period, they would present a lower density of criminal activities. Having to cope with the censored data, we set apart for the last three years (2016-2019) of the data and consequently we only considered individuals with some chance of completing a three-year cycle of criminal activities to better visualize the results between one day and three years and over. Nevertheless, even with this adapted approach we couldn't expect to have the perfect big picture as, working with a recent complete set of data, some of the individuals are still pursuing their criminal career.

4.5 Co-offending and SNA parameters

The negative relationship between the diversity index and component density observed for the known group agrees with the pattern reported by McGloin and Piquero (2010) in their study on network redundancy. Such a relationship is not observed here for the unknown group. However, the smaller co-offending data subset of the unknown available for this analysis, including only 53 large components, could partly explain the observed discrepancy as unknown are very predominant in solo activities (which are absent here), a category of data where the Di didn't show any correlation with density according to McGloin and Piquero (2010). The explanation for why unknowns maintain a similar level of specialization at different component density is more straightforward in the case of those specialized in a single crime type (Di = 0): A higher component density is associated with redundant networks where the individuals (more interconnected) share the same limited information while individuals in non-redundant networks (less interconnected) have access to diverse knowledge, opportunities and skills, which is obviously true for central individuals (McGloin and Piquero 2010;

Morselli and Tremblay 2004). Moreover it has also been demonstrated that in delinquent friendship, “High density in combination with a delinquent peer network translates into greater delinquency involvement” (Haynie 2001) thus giving more exposure to the individuals involved and more opportunities to be caught. Thus, it could be argued that some unknowns could become more at risk, when integrated in components showing a higher density, while the ones with complete specialization (Di = 0) oppose this tendency.

The average betweenness centrality increases with increasing involvement in co-offending, which is expected. Yet, the rise is much steeper for known offenders. This observation is in accordance with Haynie (2001), on data from a year in the classroom (1994-95), where the “central position in a delinquent peer network is associated with an increased delinquency index.” Thus giving again to the knowns more exposure. Considering previous finding (Lavergne *et al.* (2022) that unknowns are not less integrated than knowns in network components, and indeed contribute to increasing betweenness values for the knowns, this suggests that unknowns find ways to keep a low profile as they could try to maintain interrelations with others to a minimum and consequently ended with less betweenness centrality minimizing exposure. Moreover, is it being networked and central that drives the individual to more co-offending or is it their propensity for co-offending that ends up making them more networked and central? An issue that remains open with the data we have on hand.

Betweenness centralities assessed with DNA data or other type of information creates networks that are dynamic, in a way that they are built up through time, and should be interpreted differently compared to the usual “broker” position (Morselli and Roy 2008; Bichler 2019).The individual with a higher betweenness centrality score doesn’t have necessarily maintain social relationships with all others connected to him/her in the network. These accomplices may have been integrated at very different periods in the dynamic of the component history. Considering this aspect, it is expected that betweenness centrality correlates positively with the number of co-offending cases.

Therefore, even though a fair proportion of unknowns (34%) are involved in co-offending, they establish, in the long run, less contact as it is shown by their lower betweenness centrality built over a large period.

4.6 Limitations

Using DNA match data as a source of information about the criminological aspect of an offender's criminal activities has limitations. However, because the data is obtained from a common source (in the same way that data can be obtained from a common databank) the identification status about identified and unidentified individuals is unaffected by the various factors that influence whether DNA will be included in the data bank, such as the quality of the sample, and provides a rare opportunity to analyze the criminal activities of unidentified individuals. Therefore, criteria used to select eligible samples are key to obtaining comparable results between the two groups. However, all aspects of criminality in the population could not be studied as the data is limited to information from crime scenes that involve good DNA. In addition to good quality, potentially irrelevant DNA is filtered under investigation by searching for witnesses, family members or post-event contamination (Lapointe, Rogic, Bourgoin, Jolicoeur and Séquin 2015) in order to obtain the most probable DNA attributable to the perpetrators.

Having data on eight crime categories is rather rare in criminological studies, yet it is not without limitations here. Actually, two of these are somewhat catchall categories: “secondary” and “rare secondary” offenses hide a large spectrum of crime types, such as drug dealing, extortion, arson, or forgery, which are associated with a diversity of offenders’ skills, motivations, behavior, or else (Tremblay and Morselli 2000). Our individual-based analyses would benefit from a better breakdown of crime categories, especially to further investigate if discrepancies were observable for the unknowns that are more active in solo offending. Our catchall categories could possibly be responsible for the same level of activity seen in the “rare secondary” crime category (solo 15%, co-offending 15%) and for the slightly similar results in “secondary” offenses (10% in co-

offending, 17% in solo). The same observation is applicable for the knowns that are already showing less activity in both crime categories and behaviors.

Comparing this study to those based on Québec police files (e.g., Gründ and Morselli 2017; Charette and Papachristos 2017), our 19 years of DNA match data, with his 22,210 cases represent a rather small amount of criminal information, as for the sole year of 2021, 262,033 felonies of all categories have been registered in the Québec Province[1]. Incidentally, the vast majority of these case files had no genetic expertise and consequently no DNA matches to include unknowns. Nevertheless, our DNA match data could be analyzed using the same approaches as in Gründ and Morselli (2017) and Charette and Papachristos (2017) giving the benefit of including unknown individuals. Moreover, even though DNA matches data provide only part of the criminal activities (Beauregard and Bouchard 2010), looking for discrepancies between known and unknown with DNA can still provide sensible results when the two groups are treated equally from the same data source. Moreover, offender's DNA coming from an individual, whatever his level of forensic awareness, is left accidentally as any other type of trace regardless of the crime types or their behavior (Margot 2014), which could be extended to whether or not their ID is known because offenders mainly try to conceal their identity as in sexual assaults (Beauregard and Bouchard 2010).

It should be noted that identified teenagers (<18 years old) are underrepresented in our data, a consequence of bylaws included in the NDDB that restrict disclosure of teens' names under some circumstances. While this specific subgroup is frequently scrutinized in criminological studies, the findings presented here reflect criminal activities of a mixed adult and teen population where the latter represent a smaller proportion and could be somehow overrepresented in the unknown group.

A specific bias that could be present in a DNA dataset is the fact that DNA data may probably come mostly from careless felons only while we do not know if some

[1]https://www.quebec.ca/gouvernement/ministere/securite-publique/publications/statistiques-criminalite-quebec

individual act under some awareness concerning scene contamination by their DNA (Beauregard and Bouchard 2010; Chopin, Beauregard and Bitzer 2020). This could be the situation for some convicted felons who learn from their trials or other inmates how to be careful in their future activities (Stevens 2008), which is less of a concern regarding unknown offenders, who should not have gone to prison yet. Hypothetically, their ability to elude identification may partly be linked to a more educated state in a forensic awareness way.

Our DNA match data set contains information about co-offending known individuals with a single criminal case. However, as explained in Methods, information about single-offense unknown individuals cannot be retrieved from the same data. Consequently, our interpretation and conclusion concerning the co-offending activities should be interpreted accordingly and with care. Even though these individuals are present at a level of 27% (1890/7014) their influence shouldn't be up to the level where it would interfere significantly on the main aspect of the unknown's prominent solo acting results.

4.7 Future research

The present study, along with a few others (Jeuniaux *et al.* 2016; Lammers *et al.* 2012; DeMoor *et al.* 2020; Lavergne *et al.* 2022), represent first steps in the study of criminal careers based on DNA matches. Altogether, they have allowed the identification of either new research questions deserving deeper investigation in the future, or previously known questions on which DNA match data can bring new insights. As an example of the former, conducting more detailed studies of the seriousness of the crimes committed by offenders just before and after the moment of their identification by the police could shed light on what factors help to betray unknown offenders or, adversely, prevent them from being detected. Criminal activities of known individuals during the period when they were still unknown (i.e., before their identification) could be addressed in a comparison way to the post-identification

criminal career and to the others that remain unknown under the aspects of crime volume and violence, forensic awareness and deterrence.

The network structure surrounding the unknowns can also reveal patterns useful for police investigations. To deepen this type of research, the addition of police file data to DNA matches would provide a more complete picture of criminal networks. The addition of other types of trace data can further support forensic intelligence (Rossy *et al.* 2013; Ribaux and Talbot Wright 2014). Comparing several DNA databanks around the world is of interest to understand the strengths, limitations, and biases associated with DNA match data and analysis policy along with DNA database dynamic evolution (Leary and Pease 2003). The latter approach could serve to describe what information and data features should best lead to insightful DNA-based network analysis. Networks are built around co-offending dynamics (Gründ and Morselli 2017), which may be missed when looking at them from a static perspective, and calls for approaches based on dynamic network visualization (Lavergne *et al.* 2022; Leary and Pease. 2003). Large data sets like the LSJML one have the potential to reveal many aspects of the role of unknown offenders in the building up of criminal trajectories in various contexts and in the evolution of co-offending networks through time. In that matter, future research could also incorporate elements of criminal psychology (e.g., see Haynie 2002) for additional insights into how co-offending networks arise and evolve.

5. Conclusion

Our analysis of LSJML data using crime-related variables and social network measurements show that discrepancies may exist between offenders arrested by police forces and those who elude identification. One salient finding is that unknowns appear to restrict their activities more than known offenders do: they are more active in solo offending, commit on average fewer offenses per unit of time, and tend to specialize, hence reducing the diversity of crimes they commit, perhaps thereby improving their skills. They are also less inclined to do violent crimes. They are more active in crime

providing a greater monetary return. From this last observation, a question may emerge as whether or not those financially profitable crimes are regularly associated with forensic DNA traces, which would make the latter particularly useful to generate investigation leads when they could be linked to co-offending. From the unknowns we also observed that they spread their activities over a larger period of time, exhibiting a lower frequency of crime per year. Finally, they possibly collect co-offenders over time at a slower rate than the knowns do. The unknowns get involved in more serious crime when they are active with more accomplices or integrated in a denser component, which supports Haynie's (2001) hypothesis that peer influence enhances violent delinquency. Overall, these observations seem coherent with the exposure and competence hypotheses. According to Ouellet and Bouchard (2017), unknown offenders (partly?) include individuals who maintain a low profile by committing crimes that attract less attention by the police. Our study illustrates how the population-level analysis of DNA match data can help better understand why some offenders remain unknown to police forces for longer periods of time and our findings about the unknown's behavior could be of help at the investigation level in police departments. Finally, the similarities observed between our results about diversity index, crime seriousness, and solo vs co-offending behavior, for the known individuals and sometimes for the unknowns, and the one observed in traditional criminological studies would support the future use of DNA data for general analysis with always the advantage of including unknown individuals.

6. References

Agresti, Alan and Barbara F. Agresti (1978) Statistical analysis of qualitative variation. *Sociological methodology* 9, 204-237.

Akers, Ronald L. (2017) *Social learning and social structure: A general theory of crime and deviance*. Routledge.

Alderden, Megan A. and Timothy A Lavery. (2007) Predicting homicide clearance in Chicago: Investigating disparities in prediction across different types of homicide. *Homicide Studies* 11, 115-132.

Amemiya, E. C. (1963) Measurement of economic differentiation. *Journal of Regional Science* 5, 85-88.

Baveja, Alok, Rajan Batta, Jonathan P. Caulkins and Mark H. Karwan (1993) Modeling the response of illicit drup markets to local enforcement. *Socio-Economic Planning Sciences* 27, 73-89.

Beauregard, Eric and Martin Bouchard (2010) Cleaning up your act: Forensic awareness as a detection avoidance strategy. *Journal of Criminal Justice* 38, 1160-1166.

Bichler, Gisela (2019) *Understanding Criminal Networks: A Research Guide*, University of California Press.

Biderman, Albert D. and Albert J. Reiss (1967) On exploring the "dark figure" of crime. *The Annals of the American Academy of Political and Social Science* 374, 1-15.

Blomberg, Thomas D., William D. Bales and Alex R. Piquero (2012) Is educational achievement a turning point for incarcerated delinquents across race and sex? *Journal of youth and adolescence* 41, 202-216.

Blumstein, Alfred and Jacqueline Cohen (1987) Characterizing criminal careers. *Science* 237, 985-991.

Blumstein, Alfred, Jacqueline Cohen, Somnath Das and Soumyo D. Moitra (1988) Specialization and seriousness during adult criminal careers. *Journal of Quantitative Criminology* 4, 303-345.

Blumstein, Alfred, Jacqueline Cohen, Alex Piquero and Christy A. Visher (2010) Linking the crime and arrest processes to measure variations in individual arrest risk per crime (Q). *Journal of Quantitative Criminology* 26, 533-548.

Boivin, Rémi (2021) *Petit traité d'analyse criminelle,* Les Presses de l'Université de Montréal. ISBN 9782760643147

Borgatti, Stephe P., Kathlenn M. Carley and David Krackhardt (2006) On the robustness of centrality measures under conditions of imperfect data. *Social Networks* 28, 124-136.

Bright, David, Chad Wheland and Marie Ouellet (2022) Assessing variation in co-offending networks. *Global Crime* 23, 101-121

Budescu, David V. and Mia Budescu (2012) How to measure diversity when you must. *Psychological Methods* 17, 215.

Campana, Paolo and Federico Varese (2022) Studying organized crime networks: Data sources, boundaries and the limits of structural measures. *Social Networks* 69, 149-159

Carrington, Peter J. (2002) Group crime in Canada. *Canadian Journal of Criminology* 44, 277.

Charette, Yanick and Andrew V. Papachristos (2017) The network dynamics of co-offending careers. *Social Networks* 51, 3-13.

Chopin, Julien, Eric Beauregard and Sonja Bitzer (2020) Factors influencing the use of forensic awareness strategies in sexual homicide. *Journal of Criminal Justice* 71, 101709.

Clare, Joseph (2011) Examination of systematic variations in burglar's domain-specific perceptual and procedural skills. *Psychology, Crime and Law* 17,199-214.

Conrad, Kendon J., Barth B. Riley, Karen M. Conrad, Ya-Fen Chan and Michael L. Dennis (2010) Validation of the Crime and Violence Scale (CVS) against the Rasch measurement model including differences by gender, race, and age. *Evaluation review* 34, 83-115.

Conway, Kevin P. and Joan McCord (2002) A longitudinal examination of the relation between co-offending with violent accomplices and violent crime. Aggressive Behavior: *Official Journal of the International Society for Research on Aggression* 28, 97-108.

De Moor, S., Vander Beken, T. et Van Daele, S. (2017) DNA Databases as Alternative Data Sources for Criminological Research. *European Journal on Criminal Policy and Research* 23, 175-192.

De Moor, Sabine, Christophe Vandeviver and Tom Vander Beken (2018a) Are DNA data a valid source to study the spatial behavior of unknown offenders? *Science and Justice* 58, 315-322.

De Moor, Sabine, Christophe Vandeviver and Tom Vander Beken (2018b) Integrating police-recorded crime data and DNA data to study serial co-offending behaviour. *European Journal of Criminology* 15, 1-20.

De Moor, Sabine, Christophe Vandeviver and Tom Vander Beken (2020) Assessing the missing data problem in criminal network analysis using forensic DNA data. *Social Networks* 61, 99-106.

Elliott, Delbert. S., David Huizinga and Suzanne S. Ageton (1985) *Explaining delinquency and drug use*, Sage Publications Beverly Hills, CA.

Gründ, Thomas and Carlo Morselli (2017) Overlapping crime: Stability and specialization of co-offending relationships. *Social Networks* 51, 14.

Haynie, Dana L. (2001) Delinquent peers revisited; does network structure matter? *American journal of sociology* 106, 1013-1057.

Haynie, Dana L. (2002) Friendship networks and delinquency: The relative nature of peer delinquency. *Journal of quantitative criminology* 18, 99-134.

Heller, Nelson B. and Thomas McEwen (1973) Applications of crime seriousness information in police departments. *Journal of criminal justice* 1, 241-253.

Hodgson, Beth (2007) Co-offending in UK police recorded crime data. *The Police Journa*l 80, 333-353.

Huebner, Beth M. and Timothy S. Bynum (2016) Crime in the lifecourse. Dans *The handbook of measurement issues in criminology and criminal justice*. Wiley.

Jacobs, Bruce A. and Richard Wright (20060 *Street justice: Retaliation in the criminal underworld.* Cambridge University Press.

Jeuniaux, Patrick M. H., Leen Duboccage, Bertrand Renard, Pierre Van Renterghen and Vanessa Vanvooren (2016) Establishing networks in a forensic DNA database to gain operational and strategic intelligence. *Security Journal* 29, 584-602.

Knight, Barry J. and Donald J. West (1975) Temporary and continuing delinquency. *British Journal of Criminology* 15, 43.

Kranenbarg, Marleen Weulen (2022) When do they offend together? Comparing co-offending between different types of cyber-offenses and traditional offenses. *Computers in human behavior* 130, 107-186

Krause, Robert W., Mark Huisman, Christian Steglich and Tom Snijders (2020) Missing data in cross-sectional networks–An extensive comparison of missing data treatment methods. *Social Networks* 62, 99-112.

Kwan, Ying Keung, Wai Cheung Ip and Patrick Kwan (2000) A crime index with Thurstone's scaling of crime severity. *Journal of criminal justice* 28, 237-244.

Lammers, Marre (2014) Are Arrested and Non-Arrested Serial Offenders Different? A Test of Spatial Offending Patterns Using DNA Found at Crime Scenes. *Journal of Research in Crime and Delinquency* 51, 143-167.

Lammers, Marre and Wim Bernasco (2013) Are mobile offenders less likely to be caught? The influence of the geographical dispersion of serial offenders' crime locations on their probability of arrest. *European Journal of Criminology* 10, 168-186.

Lammers, Marre, Wim Bernasco and Henk Eiffers (2012) How Long Do Offenders Escape Arrest? Using DNA Traces to Analyze When Serial Offenders Are Caught. *Journal of Investigative Psychology and Offender Profiling* 9, 13.

Lapointe, Martine, Anita Rogic, Sarah Bourgoin, Christine Jolicoeur and Diane Séguin (2015) Leading-edge forensic DNA analysis and the necessity of including crime scene investigators, police officers and technicians in a DNA elimination database. *Forensic Science International Genetics* 19, 50-55.

Lavergne, Léo, Rémi Boivin, Simon Baechler, Patrick Jeuniaux, Karine Fiola, Diane Séquin, Jean-François Lefebvre and Emmanuel Milot (2022) Determining the impact of unknown individuals in criminality using network analysis of DNA matches. *Forensic Science International* 331, 11142.

Leary, Dick and Ken Pease (2003) DNA and the active criminal population. *Crime Prevention and Community Safety* 5, 7-12.

Lieberson, Stanley (1969) Measuring population diversity. *American Sociological Review* 34, 850-862.

Lovell, Rachel, Wenxuan Huang, Laura Overman, Daniel J. Flannery and Joanna Klingenstein (2020) Offending histories and typologies of suspected sexual offenders identified via untested sexual assault kits. *justice and behavior* 47: 470-486

Lovell, Rachel, Misty Luminais, Daniel J. Flannery, Laura Overman, Duoduo Huang, Tiffany Walker and Dan R. Clark (2017) Offending patterns for serial sex offenders identified via the DNA testing of previously unsubmitted sexual assault kits. *Journal of Justice* 52, 68-78

Margot, Pierre (2014) Traceology: The trace as the fundamental vector of police science/forensic science. *Revue Internationale de Criminologie et de Police Technique et Scientifique*. 67, 72-94.

Mazerolle, Paul, Robert Brame, Ray Paternoster, Alex R. Piquero and Charles Dean (2000) Onset age, persistence, and offending versatility: Comparisons across gender. *Criminology* 38, 1143-1172.

McCord, J. and K. P. Conway (2002). Patterns of juvenile delinquency and co-offending. *Crime and social organization* 10, 15-30.

McGloin, Jean Marie and Alex. R. Piquero (2009) 'I Wasn't Alone': Collective behaviour and violent delinquency. *Australian & New Zealand Journal of Criminology* 42, 336-353.

McGloin, Jean Marie and Alex. R. Piquero (2010) On the Relationship between Co-Offending Network Redundancy and Offending Versatility. *Journal of Research in Crime and Delinquency* 47, 63-90.

McGloin, Jean Marie, Christophe Sullivan, Alex R. Piquero and Sarah Bacon (2008) Investigating the stability of co-offending and co-offenders among a sample of youthful offenders. *Criminology* 46, 155-188.

Milot, Emmanuel, Marie Lecomte, Hugo Germain and Frank Crispino (2013) The national DNA data bank of Canada: a Quebecer perspective. *Frontiers in genetics* 4, 249.

Morselli, Carlo and Julie Roy (2008) Brokerage Qualification in Ringing Operations. *Criminology* 46, 71-98

Morselli, Carlo and Marie-Noële Royer (2008) Criminal mobility and criminal achievement. *Journal of Research in Crime and Delinquency* 45, 4-21.

Morselli, Carlo and Pierre Tremblay (2004) Criminal achievement, offender networks and the benefits of low self-control. *Criminology* 42, 773-804.

Nieto, Alberto, Toby Davies and Hervé Borrion (2022) "Offending with the accomplices of my accomplices": Evidence and implications regarding triadic closure in co-offending networks. *Social Networks* 70, 325-333

Ouellet, Frédéric and Martin Bouchard (2017) Only a matter of time? The role of criminal competence in avoiding arrest. *Justice Quarterly* 34, 699-726.

Pederson, Maria L. (2018) Do offenders have distinct offending patterns before they join adult gang criminal groups? Analyses of crime specialization and escalation in offence seriousness. *European Journal of Criminology* 15, 680-701.

Piquero, Alex R., Oster Paternoster, Paul Mazerolle, Robert Brame and Charles W. Dean (1999) Onset age and offence specialization. *Journal of Research in Crime and Delinquency* 36, 275-299.

Piquero, Alex R., David P. Farrington, Alfred Blumstein (2003) The criminal career paradigm. *Crime and Justice* 30, 359-506.

Piquero, Alex R., Randall Macintosh and Matthew Hickman (2002) The validity of a self-reported delinquency scale: Comparisons across gender, age, race, and place of residence. *Sociological Methods & Research* 30, 492-529.

Piquero, Alex R. and David Weisburd (2010*) Handbook of quantitative criminology*, Springer.

Pyrooz, David, C., Jean Marie McGloin and Scott H. Decker (2017) Parenthood as a turning point in the life course of male and female gang members: a study of within-individual changes in gang membership and criminal behaviour. *Criminology* 55, 869-899.

Reiss Jr Albert J. and David P. Farrington (1991) Advancing knowledge about co-offending: Results from a prospective longitudinal survey of London males. *Journal of Criminal Law & Criminology* 82, 360.

Ribaux, Olivier and Benjamin Talbot Wright (2014) Expanding forensic science through forensic intelligence. *Science and Justice* 54, 494-501.

Rossi, Peter H., Jon E. Simpson and JoAnn L. Miller (1985) Beyond crime seriousness: Fitting the punishment to the crime. *Journal of Quantitative Criminology* 1, 59-90.

Rossi, Peter H., Emily Waite, Christine E. Bose and Richard E. Berk (1974) The seriousness of crimes: Normative structure and individual differences. *American Sociological Review* 39, 224-237.

Rossy, Quentin, Sylvain Loset, Damien Dessimoz and Olivier Ribaux (2013) Integrating forensic information in a crime intelligence database. *Forensic Science International* 230, 137-146.

Rossy, Quentin and Carlo Morselli (2017) The contribution of forensic science to the analysis of crime networks. Dans *The Routledge international handbook of forensic intelligence and criminology*. Routledge.

Smith, Jeffrey A. and James Moody (2013) Structural effects of network sampling coverage I: Nodes missing at random. *Social Networks* 35, 652-668.

Smith, Jeffrey A., James Moody and Jonathan H. Morgan (2017) Network sampling coverage II: The effect of non-random missing data on network measurement. *Social Networks* 48, 78-99.

Sparrow, M. K. (1991) The application of network analysis to criminal intelligence: An assessment of the prospects. *Social Networks* 13, 251-274.

Stevens, Dennis J. (2008) Forensic science, wrongful convictions and American prosecutor discretion. *The Howard Journal of Criminal Justice* 47,31-51.

Stolzenberg, Lisa and Stewart J. D'Alessio (2008) Co-offending and the age-crime curve. *Journal of research in crime and delinquency* 45, 65-86.

Sullivan, Christopher J., Jean Marie McGloin, Travis C. Pratt and Alex R. Piquero (2006) Rethinking the "norm" of offender generality: Investigating specialization in the short-term. *Criminology* 44, 199-233.

Sullivan, Christopher J., Jean Marie McGloin, James V. Ray and Michael S. Caudy (2009) Detecting specialization in offending: comparing analytic approaches. *Journal of quantitative criminology* 25, 419-441.

Thurstone, Louis L. (1927) A law of comparative judgment. *Psychological review* 34, 273.

Tremblay, Pierre and Carlo Morselli (2000) Patterns in criminal achievement: Wilson and Abrahamse revisited. *Criminology* 38, 633-657.

Uggen, Christopher and Jeremy Staff (2004) Work as a turning point for criminal offenders. *Crime and employment: Critical issues in crime reduction for corrections.* 65, 141-168.

Vandeviver, Christophe, Stijn Van Daele, and Tom Vander Beken (2015) What makes long crime trips worth undertaking? Balancing costs and benefits in burglars' journey to crime. *British Journal of Criminology* 55, 399-420.

Van Mastrigt, Sarah B. and David P. Farrington (2009) Co-offending, age, gender and crime type: Implications for criminal justice policy. *The British Journal of Criminology* 49, 552-573.

Weaver, Beth and Alistair Fraser (2022) The social dynamics of group offending. *Theoretical Criminology* 26, 264-284

Weerman, Frank M. (2003) Co-offending as Social Exchange. Explaining Characteristics of Co-offending. *British journal of criminology* 43, 398-416.

Wolfgang, Marvin E. (1985) *The national survey of crime severity*, US Department of Justice, Bureau of Justice Statistics.

Zimring, Franklin E. (1981) Kids, groups and crime: Some implications of a well-known secret. *Journal of Crime Law & Criminology* 72, 867.

CHAPITRE IV

PRODUCTION DE RENSEIGNEMENT CRIMINEL GRÂCE À L'ANALYSE DYNAMIQUE DES CONCORDANCES ADN EN RÉSEAUX JUMELÉS À DES DONNÉES POLICIÈRES.

« L'ampleur de certaines enquêtes criminelles implique de mettre en œuvre des démarches structurées de traitement des informations collectées, afin de maîtriser le flux et de conserver une vue d'ensemble. … (Rossy 2016).

4.1 Introduction

Les expertises génétiques utilisées pour l'identification des criminels placent parfois les enquêteurs devant la situation où des individus ne sont connus que par leur seul ADN, trouvé sur une scène de crime, l'identité de ces derniers est donc inconnue. Pour pallier ce manque d'information, on peut se servir des concordances ADN afin de reconstituer le réseau social criminel entourant les inconnus (Jeuniaux *et al.* 2016). C'est ainsi que De Moor *et al.* (2020) ont intégré les données ADN aux données policières de Belgique pour évaluer le positionnement des délinquants inconnus dans les réseaux de la criminalité. Nous proposons de nommer cette approche *Social network analysis from DNA* ou SNDNA. L'approche SNDNA n'est toutefois pas seulement intéressante pour retracer des inconnus, mais elle peut aussi fournir des informations inédites sur les activités de co-délinquance des individus d'identité connue des services de police, en révélant leurs liens avec des inconnus. Tel que présentée dans les chapitres II et III de la présente thèse (Lavergne *et al.* 2022, soumis), cette approche a été appliquée aux données québécoises pour mettre au jour des caractéristiques associées aux inconnus dont l'ADN seul révèle la présence.

Premièrement, il a été démontré que ces inconnus sont bien intégrés dans les réseaux de co-délinquance et qu'ils n'occupent donc pas, en moyenne, une position marginale dans ces réseaux. Cela signifie que d'autres délinquants, connus eux des

policiers, en savent peut-être plus sur les inconnus qu'on ne le pense généralement (De Moor *et al.* (2017). Deuxièmement, nous avons montré que les inconnus sont souvent complices avec les connus. L'étude criminologique de la co-délinquance a montré que les délinquants récidivent parfois avec les mêmes complices, et ce, pour diverses raisons. Par exemple, ces complices peuvent posséder des compétences complémentaires, avoir développé une amitié, ou être en mesure de profiter plus rapidement des occasions se présentant inopinément, puisqu'ils ont une expérience criminelle partagée (Weerman 2003; Charrette et Papachristos 2017; Gründ et Morselli 2017).

À la lumière des connaissances exposées dans les deux précédents chapitres, il sera ici question d'explorer le potentiel de l'approche SNDNA pour générer du renseignement criminel tactique en soutien aux enquêtes autour des individus inconnus qui restent à identifier. Pour ce faire, l'approche SNDNA sera appliquée à des cas concrets de crimes sur lesquels la Sûreté du Québec a enquêté dans le passé.

D'une part, les concordances ADN nous informent sur les inconnus et la structure sociale de co-délinquance qui les entoure. D'autre part, les données policières sur ces mêmes délits apportent une foule d'informations sur les circonstances qui les entourent (lieu, moment précis, etc.), les observations rapportées par des témoins ou fournies par les caméras de surveillance (p. ex. : nombre et apparence physique des délinquants), les traces retrouvées sur la scène (digitales, de pas, d'outils, etc.), les résultats d'expertises scientifiques sur celles-ci et sur d'autres éléments pertinents à l'enquête. De plus, les données policières permettent de compléter le portrait du réseau de co-délinquance. En effet, des liens entre des individus connus ont pu être détectés par les policiers dans des dossiers où l'expertise génétique n'est pas intervenue.

La quantité et la diversité des données disponibles varient selon l'enquête, de même que leur lien avec les délinquants connus et inconnus. Le couplage des données policières à celles des réseaux reconstruits par l'ADN a donc le potentiel de créer du renseignement en ajoutant une panoplie d'informations supplémentaires associées aux co-délinquants proches ou directement liées aux inconnus. La production de

renseignement s'effectue donc selon une approche plus globale. Il n'est plus question de gérer les concordances ADN au cas par cas, dans l'unique but de transmettre des identifications en provenance de la Banque nationale de données génétique (BNDG) aux enquêteurs, ni pour ces derniers de traiter isolément leurs enquêtes. L'intégration de l'ADN et de ses inconnus devient un nouvel outil tactique pour faire avancer des enquêtes, de même qu'une source d'informations ouvrant de nouvelles pistes de recherche en criminologie, sur la co-délinquance impliquant des individus inconnus, qui font souvent défaut dans ce champ de connaissances.

D'une manière plus concrète, la production de renseignement peut se structurer en utilisant des schémas relationnels[15]. L'utilisation de schémas relationnels en soutien aux enquêtes remonte aussi loin que les travaux de Wigmore qui, déjà en 1913, utilisait des graphes pour représenter les « relations de causalité entre les prémisses et les conclusions formulées en cours d'enquête. » (Rossy 2016 dans Morselli, Boivin). Dans les nombreuses années qui nous séparent de cette période, l'utilisation des graphes a évolué et s'est faite plus méthodique, et on a aussi fait intervenir davantage de types de relation (Harper & Harris (1974), Morris (1986)). Dans ce contexte, les données de sources variées peuvent être analysées selon une approche quantitative de paramètres décrivant les réseaux et leurs éléments (individus, liens), connue sous le nom d'analyse des réseaux sociaux (ARS) (Borgatti *et al.* 2018; Bichler 2019). C'est ainsi que l'on peut quantifier la position d'un individu dans un réseau grâce à des mesures correspondant à différentes facettes de cette position (centralité ou marginalité, nombre de contacts, etc). (Freeman 1977). Le lecteur consultera l'annexe A du chapitre II pour en savoir plus sur les paramètres d'ARS qui ont été utilisés dans le premier volet de cette thèse et dont certains seront considérés plus loin pour le montage de renseignement.

Déjà en 1991, Sparrow avançait l'idée de profiter des banques de données criminelles de plus en plus complexes pour structurer ces informations en des réseaux

[15] Un schéma relationnel est une représentation de lien entre des individus. Selon le contexte dans lequel on les utilise; psychologie, informatique, affaires, économie etc, ils sont de structures différentes. Dans notre situation ils représenteront des liens entre individus et leurs délits criminels.

auxquels on pourrait appliquer les ARS. Plus près de nous, Morselli (2009, 2013) a été un pionnier de l'utilisation de l'ARS pour l'étude des réseaux criminels. Les travaux de Rossy *et al.* (2013) constituent un exemple pionnier d'intégration de données provenant de dossiers de police (lieu, période etc), de traces matérielles faisant l'objet d'analyses forensiques et de celles provenant des services d'identité policière et impliquant de nombreux types de délits. En analysant des données couvrant trois ans et six cantons suisses, ces auteurs ont par exemple démontré que le jumelage des données ADN à celles de traces de chaussures (ces dernières figurant dans les dossiers de police) s'avérait particulièrement efficace pour faire des recoupements entre cambriolages; traces d'ADN et de chaussures sont en effet fréquemment retrouvées ensemble dans ce type de délit. Ce sont ces recoupements entre dossiers qui créent du nouveau renseignement forensique. Plus tard, Rossy et Ribaux (2014) ont démontré que l'utilisation de schémas relationnels permettait de saisir visuellement l'ensemble des informations pertinentes pour le renseignement criminel, en présentant divers types de liens entre individus, événements ou lieux.

Dans la présente étude, nous proposons une approche d'intégration globale des concordances ADN aux données obtenues des services de police afin d'en tirer une vision plus complète des réseaux impliquant des individus connus, porteurs d'information, qui sont en co-délinquance avec les inconnus. Nous avancerons ici l'hypothèse que plus un inconnu présente de délits en co-délinquance, plus les chances augmentent de trouver des informations pouvant conduire à son identification. Nous y évaluons le potentiel d'amélioration du renseignement criminel en ajoutant des informations policières pertinentes concernant les délits de ces individus connus, comme les lieux, le type de délit, les dates, et autres. Ainsi, il pourrait être possible de retrouver, par association et comparaison, des délits auxquels aurait pu participer l'inconnu et ainsi ouvrir de nouvelles pistes d'enquête.

Notre choix d'utiliser le modèle basé sur les concordances ADN mises en réseau s'impose de lui-même, car on peut présumer que la structure de lien ainsi reconstruite reflète largement la co-délinquance où l'on retrouverait des individus inconnus. Le

jumelage des données en provenance de deux entités administratives, scientifique et policière, à des fins d'analyse de réseaux, n'est pas sans complexité. Il convient alors d'utiliser des schémas relationnels où les individus et les événements sont disposés de manière ordonnée, en présentant les liens qui les unissent pour mettre en évidence les éléments pertinents pour l'identification des inconnus. Le montage des informations sera présenté en utilisant un schéma relationnel adapté, inspiré de deux modèles présentés dans Rossy (2016) qui seront détaillés plus loin.

4.2 Méthode

4.2.1 Données sur les individus et les délits criminels

Les données policières utilisées dans cette étude ont été obtenues grâce à l'aimable collaboration de la Sûreté du Québec (SQ). M. Éric Chartrand, analyste criminel à la SQ, a extrait des données policières concernant les délits, suspects ou perpétrateurs figurant dans les composants du réseau ADN retenues pour la présente analyse. Au besoin, M. Chartrand a aussi consulté les dossiers d'enquêtes obtenus des postes régionaux de la SQ. M. Chartrand fut présent aux différentes étapes de l'analyse pour s'assurer que les règles de sécurité et de confidentialité des données soient appliquées selon les termes de l'entente de recherche signée avec la Direction de l'amélioration continue de la SQ.

Les concordances ADN obtenues du Laboratoire de sciences judiciaires et de médecine légale (LSJML) et compilées en date de juillet 2019 ont été utilisées pour sélectionner quatre composants parmi les 51 ayant ≥5 individus identifiés (Lavergne *et al.* 2022). Les critères de sélection étaient : 1) composant ayant au minimum un inconnu et 2) un maximum de 10 individus; 3) présentant un nombre moyen et varié de délits, question de ne pas surcharger le composant en. Par exemple, trouver des composants associés à un ou plusieurs vols qualifiés, puisque les individus actifs dans ce type de délit plus violent ont habituellement derrière eux une carrière criminelle caractérisée par un ensemble plus ou moins volumineux de délits plus mineurs comme les vols d'automobiles (Blustein *et al.* 1988), qui, parfois nombreux, auraient le potentiel d'apporter davantage de liens associés provenant de tous les

dossiers de police ajoutés. Il est à noter que cette observation est aussi valable pour les individus ayant un homicide à leur actif (Blustein *et al.* 1988). Le Tableau 1 présente les caractéristiques des composants retenus, qui au nombre de quatre devraient présenter suffisamment d'exemples de liens de co-délinquance sur des délits variés, incluant des inconnus sans pour autant surcharger le travail de collaboration avec la Sûreté du Québec.

Tableau 1 : Description des composants de base, constitués de concordance ADN, utilisés pour la production de renseignement. Les numéros de composants correspondent à ceux attribués au chapitre II.

No composant	Nbre inconnus	Nbre connus	Nbre délits	Types de délit*
36	2	7	13	HO= 1, AS= 11, ND= 1
44	2	3	16	AA= 1, CA= 1, SR= 13, AD= 1
87	1	4	10	AS= 1, VQ= 6, CA= 3
366	1	3	4	AA= 2, AD= 1, ND= 1

*HO : homicide, AS : agression sexuelle, VQ : vol qualifié, AA : agression armée, CA : cambriolage, SR : secondaire, ND : non désignée, AD : autre désignée (Vol de plus de 5000$, incendie criminel, trafic d'armes, contrefaçon, etc).

4.2.2 Construction de schémas relationnels par utilisation du sociogramme

Les composants choisis ont été traités dans le logiciel Gephi (Bastian 2009) afin de visualiser leur structure en sociogramme. Les Figures 1A et 1B montrent respectivement les sociogrammes bimodaux (individus et dossiers criminels) et unimodaux (individus seuls) pour les quatre composants retenus. Ces sociogrammes seront utilisés comme

schéma relationnel pour la production du renseignement. On prend ici la mesure de ce qu'un sociogramme apporte en visualisation lorsque l'on compare les données de base du Tableau 1 aux Figures 1A et 1B.

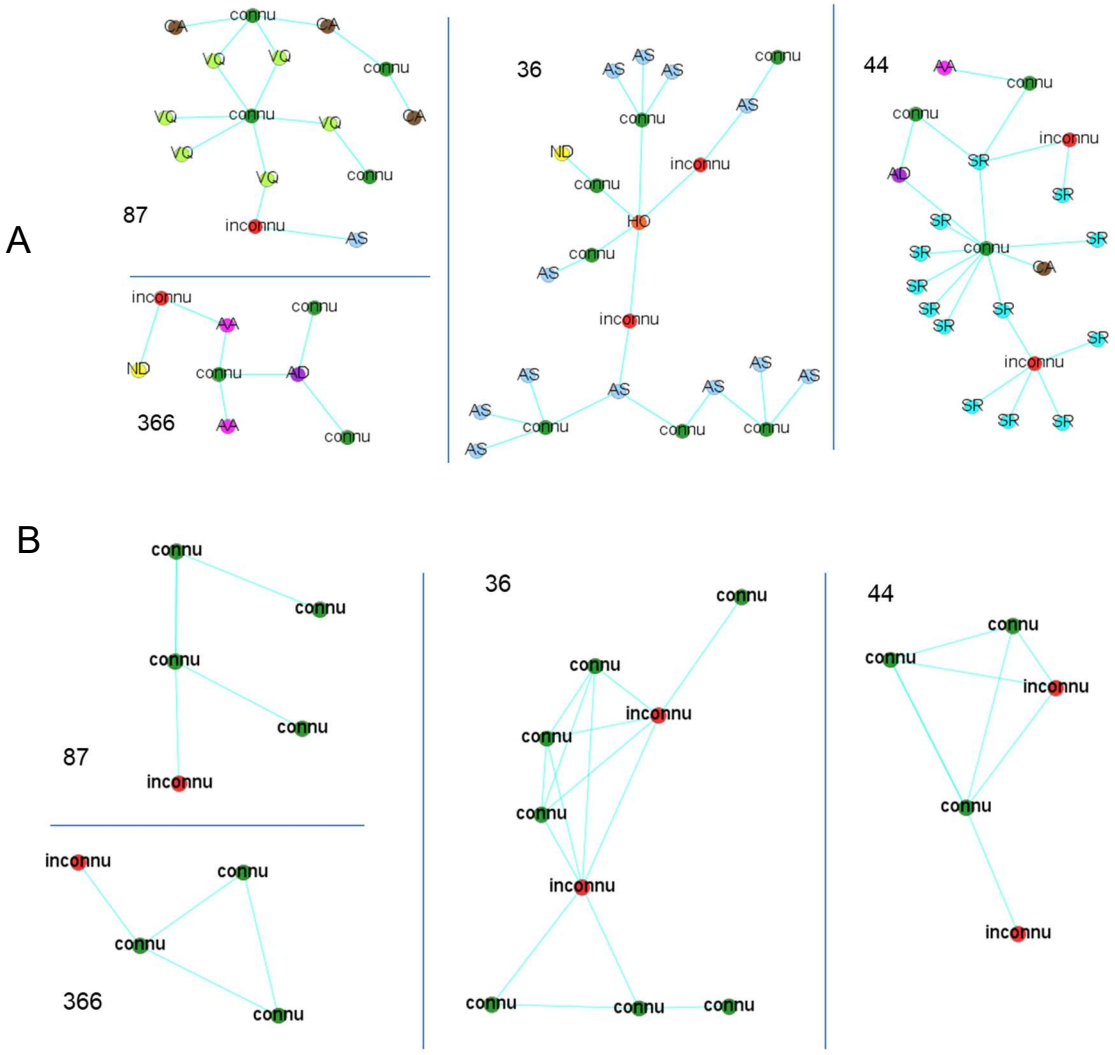


Figure 1 A et B : Sociogrammes reconstitués grâce aux concordances ADN, pour les composants 36, 44, 87 et 366. A : format bimodal avec les individus et les délits. B : format unimodal avec uniquement les individus. Les codes de couleur des divers éléments (individus et délits) et des liens sont expliqués dans la section « Caractéristiques des schémas ». Voir le Tableau 1 pour les codes d'infractions.

Les recommandations générales de Rossy (2016), mises en pratique dans Rossy et Ribaux (2014), ont été suivies d'une manière simplifiée, pour ce qui a trait à la représentation schématique des éléments (individus, délits, liens). Des couleurs

différentes ont été attribuées aux nœuds (individus et types de délits) ainsi qu'aux liens selon leur provenance (SQ, LSJML). L'approche de Rossy (2016) est plus élaborée et conçue pour des schémas relationnels complexes ayant de nombreux types d'éléments (nœuds) de nature très différente (individus, lieux, véhicules, etc.)

Le modèle présenté dans la Figure 1 relève du premier des quatre types de schéma proposés par Rossy (2016), soit le « schéma de réseau » où les individus constituent les éléments principaux, auxquels s'adjoignent des délits, ce qui correspond parfaitement aux données fournies par le LSJML. De plus, la dynamique temporelle doit être prise en considération. En effet, les composants ADN n'apparaissent pas en entier d'un seul coup, mais se développent dans le temps, au fur et à mesure que des individus et des délits s'ajoutent. C'est pourquoi, en plus du modèle « schéma de réseau », nous nous sommes inspirés du modèle « schéma de série » (Rossy 2016). L'idée derrière ce modèle est de reconstruire la série des événements (délits) selon leur chronologie, en utilisant les dates où ils surviennent (voir l'Annexe A sur la gestion des données). Ainsi, notre modèle combine les schémas de réseau et de série, afin de cerner la séquence d'apparition dans le temps des événements autour des inconnus, par leurs activités de co-délinquance.

Les données policières proviennent de la Sûreté du Québec. Elles contiennent deux types d'informations absentes des données du LSJML : d'une part des informations de base décrivant les délits criminels et les individus et, d'autre part, des informations plus techniques provenant de la scène de crime. Celles-ci complémentent les cas de co-délinquance déjà présents dans les composants reconstruits par l'ADN (Fig. 1A et 1B), se greffant comme de nouveaux liens aux délits et individus déjà en place. L'ordre d'ajout de ces informations présenté à la Figure 2 se décline comme suit.

Dans un premier temps :

a) on démarre avec le composant de base obtenu des données de concordances ADN du LSJML (Fig. 2A). À noter : les liens bleus correspondent aux données génétiques du LSJML.

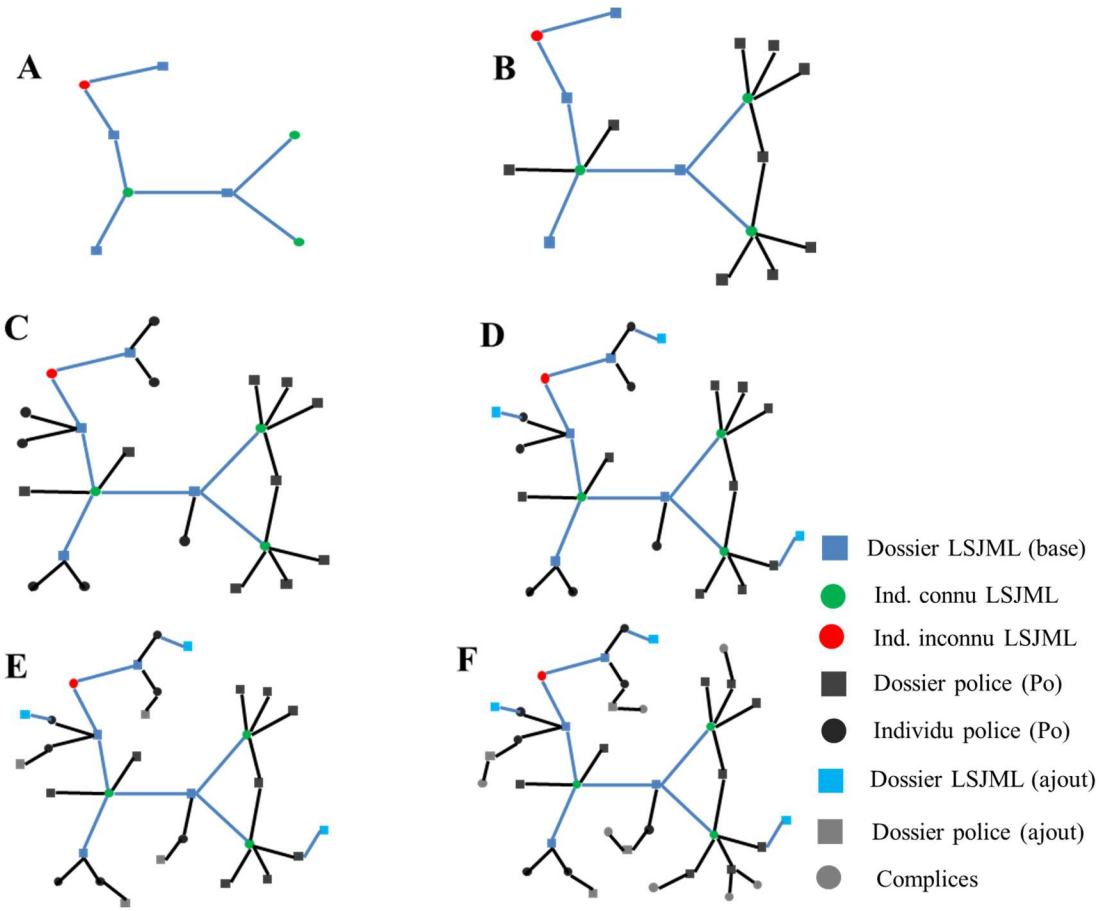


Figure 2 : Exemple d'ajout successif de couches d'informations aux composants ADN de base. A : le composant de départ est obtenu à partir des concordances ADN impliquant quatre individus, dont trois connus et un inconnu. Pris ensemble, ces individus cumulent quatre délits criminels à leur actif et sont contenus dans les données du LSJML. B : D'autres délits impliquant ces mêmes individus, et connus grâce aux données policières, sont ajoutés au schéma. C : Les données policières peuvent également contenir de l'information sur d'autres individus qui sont associés aux délits figurant dans le composant de départ. Ces individus (accusés ou suspectés) sont alors ajoutés. D : aux délits déjà contenus dans les données du LSJML, on ajoute les autres délits qui impliquent les individus ajoutés à l'étape précédente (C) et qui proviennent des données du LSJML. E : On ajoute d'autres délits connus des policiers concernant les individus ajoutés en C. F : On ajoute enfin tous les complices qui n'apparaissent pas encore dans le schéma à ce stade mais qui, on le sait, sont liés à certains de ces délits.

Ensuite on ajoute :

b) les autres délits criminels associés aux individus connus déjà présents dans le réseau ADN. Pour rappel, ces délits ne contiennent pas de résultats ADN liés aux individus déjà présents et sont donc absents des données du LSJML (carrés noirs

dans la Fig. 2B). À noter les liens noirs relient les données obtenues des services de police

c) les individus associés aux délits criminels figurant déjà dans le réseau ADN (ronds noirs, Fig. 2C). Il s'agit ici d'individus sans lien avec les résultats ADN obtenus pour ces délits; ce sont donc les données policières qui permettent de les lier à ces crimes

Dans un deuxième temps, on fait un retour vers les données du LSJML pour ajouter :

d) les dossiers ADN provenant des délits criminels qui pourraient être associés à ces nouveaux individus ajoutés à l'étape précédente (carrés bleu pâle, Fig. 2D);

e) les délits criminels associés aux individus ajoutés précédemment en C (carrés gris, Fig. 2E).

f) les autres individus considérés comme complices dans les dossiers de police ajoutés (ronds gris, Fig. 2F).

Cette série d'ajouts d'individus et de délits criminels peut s'effectuer en boucle, avec une augmentation de l'information contenue dans le schéma à chaque étape. Dans le cas présent, puisqu'il s'agit d'un exemple, une seule itération de recherche a été effectuée. Avec des banques de données bien organisées, un analyste pourrait procéder à ces étapes de manières automatiques sur un grand ensemble de données.

4.2.3 Données circonstancielles, de traces et autres informations d'enquête

Jusqu'ici les composants ne contiennent que les informations suivantes : nombre et positionnement (liens) des individus et délits criminels, de même que le statut des individus (connu, inconnu). Les informations d'enquêtes sont propres à chaque délit et proviennent des interventions policières sur les scènes de crime, ou du processus général de l'enquête. (Rossy et coll. 2013). En voici quelques exemples :

-Traces de pas

-Traces digitales

-Rapports d'analyses balistiques

-Résultats d'analyses toxicologiques sur des substances illicites

-Listes de contacts téléphoniques ou de réseaux sociaux

-Images de caméras de surveillance

-Témoignages

Si ces informations sont en lien avec des individus ou des délits associés au composant, elles peuvent s'y intégrer d'une manière similaire en créant de nouveaux liens entre des individus ou des délits. Par exemple, si des traces de pas semblant provenir de la même source sont observées dans deux cambriolages non liés par ADN, il est alors possible d'ajouter un lien entre ces deux délits. De plus, si les mêmes profils de traces de pas[16] sont aussi observés dans des délits ne figurant pas encore au composant, on peut les y ajouter en les reliant aux dossiers du schéma ayant les mêmes traces. Le même processus d'enrichissement du schéma pourrait se faire avec des ajouts tirés de la liste précédente et on peut facilement imaginer que les communications téléphoniques entre individus seraient une source des plus riches pour établir des liens supplémentaires autour des individus présents dans le composant de base. À titre d'exemple, Soudijn *et al.* (2022) ont mis au jour un imposant réseau criminel grâce à l'analyse de millions de messages texte envoyés depuis 4 158 téléphones portables, dont les trois-quarts étaient associés à des trafiquants de drogues synthétiques. Ainsi, avec les nouveaux liens associés à des données circonstancielles on ajoute des couches d'informations qui s'intègrent, dans les interrelations autour des inconnus et, possiblement, ces informations pourraient aider à les identifier. Avec davantage d'individus ajoutés par des ajouts individu/délit, on augmente la possibilité d'obtenir plus de renseignements circonstanciels qui pourraient faire apparaître davantage de liens informatifs. Il faut toutefois préciser que les informations circonstancielles ou autres ne sont pas nécessairement intégrées dans des banques de données d'une manière les rendant facilement accessibles pour l'analyse, comme le démontrent Rossy *et al*. (2013) pour la Suisse.

4.2.4 Dynamique des réseaux

Les composants ADN représentent une compilation d'activités criminelles (Gründ et Morselli 2017), dont la structure s'est développée au fil du temps, en fonction de la

[16] Même profil de traces signifiant qu'elles peuvent provenir de la même source.

séquence d'occurrence des délits en co-délinquance, auxquels se greffent les délits commis en solitaire par les mêmes individus. L'intégration du « schéma de série », mentionné précédemment, au schéma relationnel permet de visualiser la dynamique d'un composant dans le temps. Gephi (Bastian *et al.* 2009) permet d'utiliser les dates des délits et de choisir la ou les périodes à visualiser. Les paramètres de Levallois publiés sur le site de Gephi[17] ont été utilisés pour créer une version dynamique des composants. Puisque nos données s'arrêtent en juillet 2019, nous avons choisi la date arbitraire du 1er janvier 2020 comme date de fin d'analyse qui permet d'inclure, sans perte, toutes les données jusqu'à 2019 inclus. L'échelle de temps (timeline) créée à partir des dates associés à chacun des délits permet d'étudier le composant, dans une forme simplifiée, au fil de l'évolution des ans, et de choisir la fenêtre temporelle autour des inconnus en visualisant l'apparition des complices et des délits qui leur sont associés. La dynamique d'un composant peut alors être examinée sous un angle plus informatif et de surcroît plus efficace, pour retrouver une période ayant davantage d'intérêts, comme une concentration plus élevée de délits autour d'une période temporelle et une zone géographique.

4.2.5 Caractéristiques des sociogrammes

Afin de bien visualiser les divers éléments (individus et délits) présents dans les sociogrammes, un code de couleurs pour les individus, les délits et les types de liens a été mis en place (Fig. 3).

[17] https://seinecle.github.io/gephi-tutorials/generated-html/converting-a-network-with-dates-into-dynamic.html

Figure 3 : Couleurs associées aux éléments des sociogrammes. Les cinq premiers éléments présentent les divers statuts associés aux individus, et les autres ceux associés aux types de délit. Les individus connus et les suspects sont tous deux en vert, les suspects étant plus pâles. Ces couleurs sont utilisées dans l'ensemble des sociogrammes illustrés dans ce chapitre. Outre ces couleurs associées aux éléments, les couleurs des liens nous informent sur l'origine des éléments. Les liens bleu pâle, indiquent que les dossiers sont liés par des concordances ADN du LSJML. Les liens noirs indiquent des informations obtenues des services de police tandis que si les individus ont été identifiés par ADN et par l'enquête, le lien sera orange (Fig. 4).

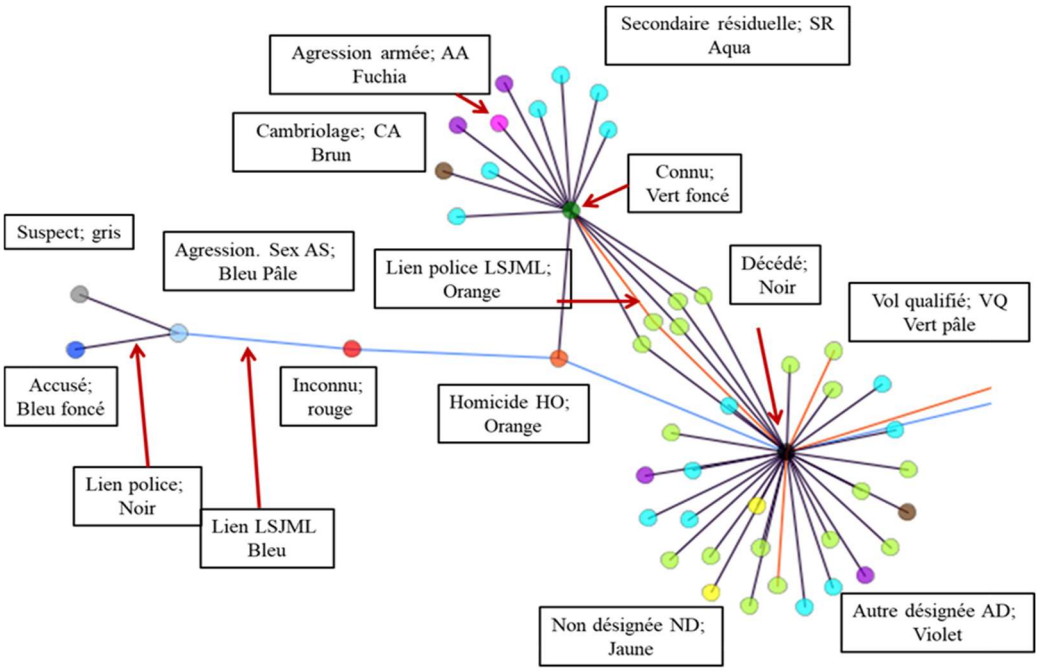


Figure 4 : Détail du sociogramme du composant 87. Voir la Figure 3 pour le code de couleurs.

4.3 Résultats

4.3.1 Ajout d'informations sur le type délit criminel et l'individu

Les Figures 5A à 5D montrent les résultats de la première étape d'ajout d'informations (étapes B et C de la Fig. 2) aux quatre composants retenus, présentés à la Figure 1A. Les données policières permettent d'ajouter de nombreux délits (175) connus de la police pour chacun des individus connus présents dans ces composants. À ces délits s'ajoutent onze individus supplémentaires identifiés dans les dossiers ADN associés aux délits ajoutés. Les sections qui suivent détaillent l'analyse de chaque composant séparément.

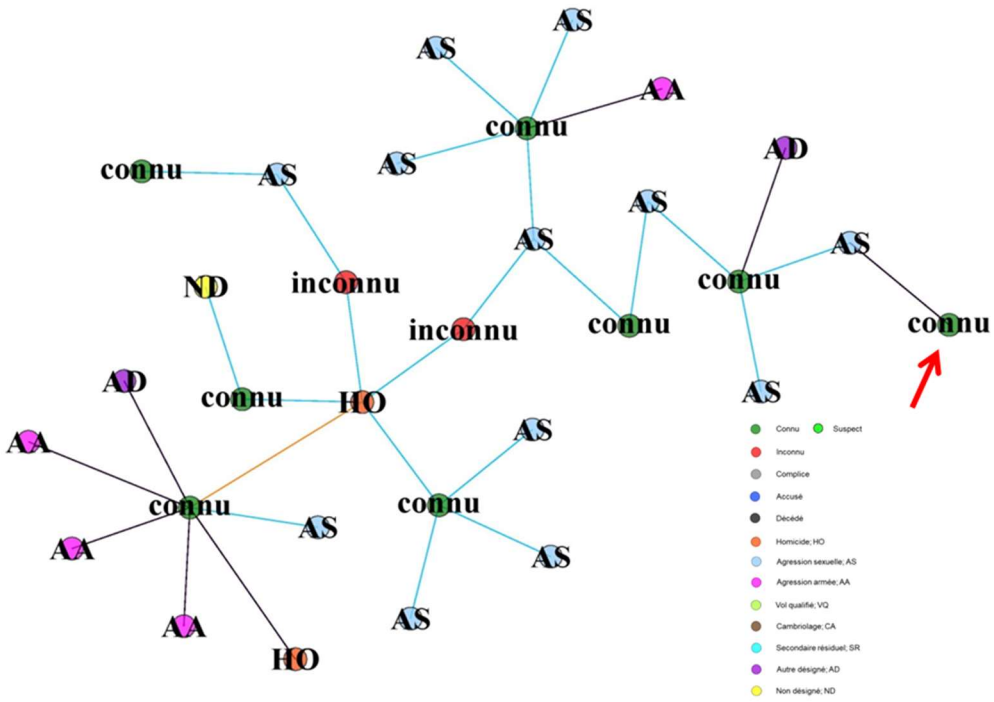


Figure 5A : Le composant 36 en format sociogramme bimodal après le jumelage avec les données policières (étapes B et C de la Fig. 2). L'individu ajouté est indiqué par une flèche rouge.

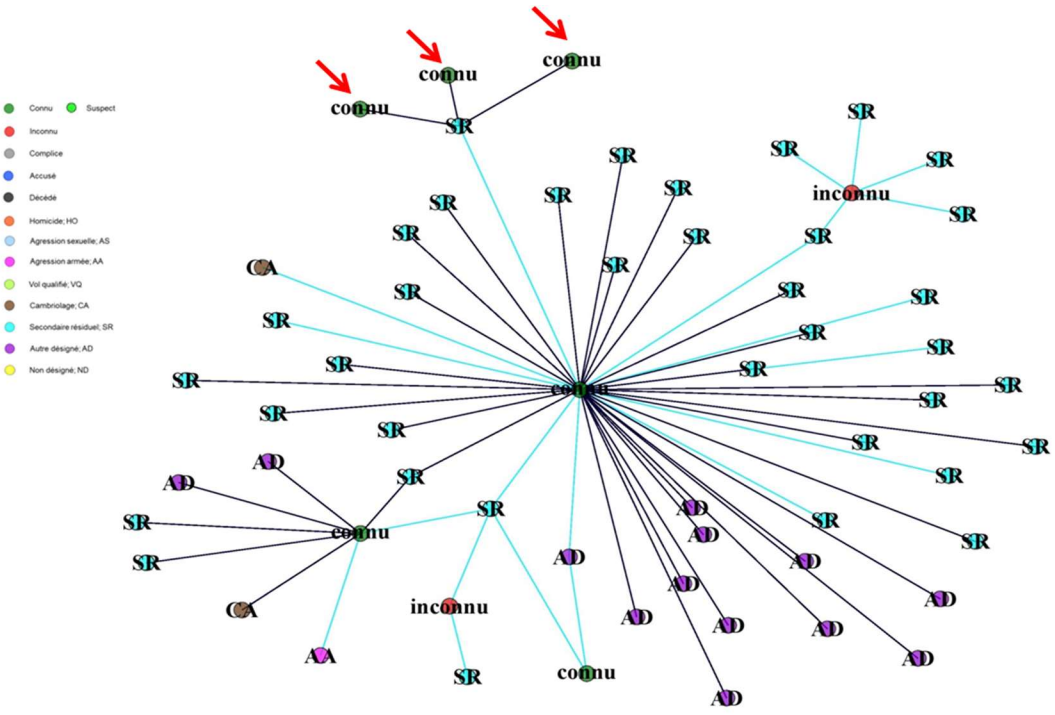


Figure 5B : Le composant 44, en format sociogramme bimodal après le jumelage avec les données policières (étapes B et C de la Fig. 2). Les individus ajoutés sont indiqués par une flèche rouge.

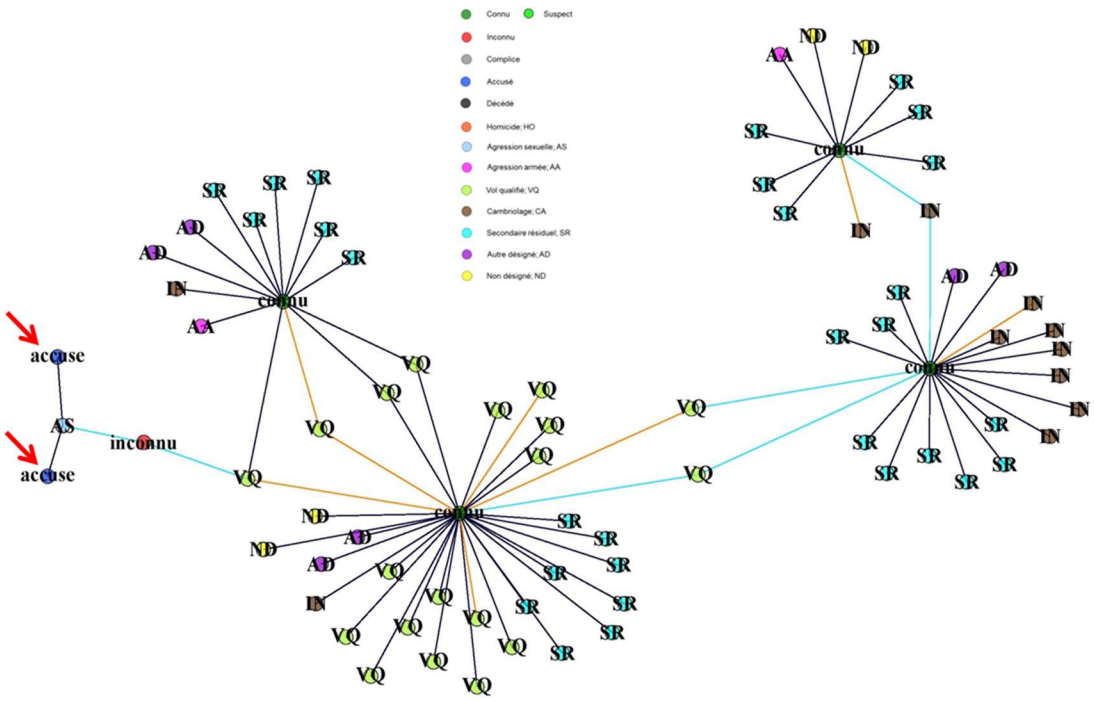


Figure 5C : Le composant 87 en format sociogramme bimodal après le jumelage avec les données policières (étapes B et C de la Fig. 2). Les individus ajoutés sont indiqués par une flèche rouge.

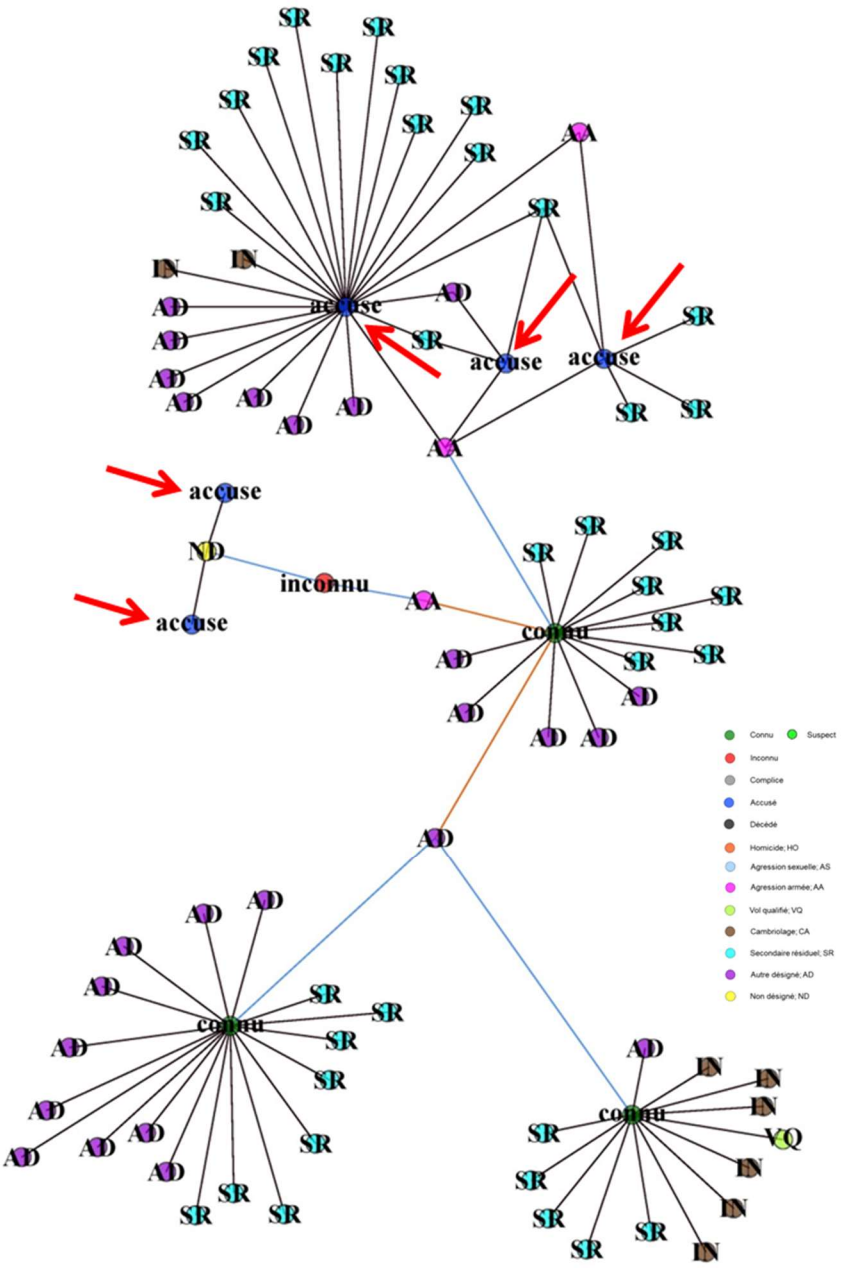


Figure 5D : Le Composant 366 en format sociogramme bimodal après le jumelage avec les données policières (étapes B et C de la Fig. 2). Les individus ajoutés sont indiqués par une flèche rouge.

4.3.1.1 Le composant 36

Ce composant est constitué d'individus principalement actifs en agressions sexuelles, dont deux se sont vus attribuer un délit supplémentaire, c'est-à-dire non inscrit dans les données du LSJML. Un troisième (individu au bas du schéma, à gauche; Fig. 5A) s'est vu attribuer un homicide, trois agressions armées et un autre délit non-désigné soit cinq délits non liés à d'autres individus. Aucun autre individu que ceux déjà connus n'ont été retrouvés dans les dossiers criminels de ce composant. Ce qui pourrait s'expliquer par la tendance des agresseurs sexuels à agir plus souvent de manière solitaire, sans montrer de spécialisation marquée dans ce type de délit (Bijleveld et Hendriks 2003). De plus, à la deuxième étape d'ajout d'information, aucun complice n'a été identifié dans les événements du composant 36, une observation qui tend à soutenir cette image de l'agresseur sexuel solitaire.

4.3.1.2 Le composant 44

Contrairement au composant 36, les données policières permettent l'ajout au composant 44 de nombreux délits de type secondaire résiduels (SR) et autre désigné (AD), ainsi que trois individus connus de la police, tous co-délinquants sur une SR (Fig. 5B). La structure de ce composant peut rappeler celle d'un large réseau de trafiquants de stupéfiants (Malm et Bichler 2011). Pour en décortiquer la complexité, il sera ici examiné à partir du composant initial basé sur les concordances ADN (Fig. 1 et 6A). Dans ce dernier on retrouve plusieurs délits de type SR attribués à l'individu connu du centre (43899). Il s'agit de l'individu sur lequel, au départ l'attention des enquêteurs s'est portée, puisque ce dernier est le plus actif en trafic de stupéfiants.

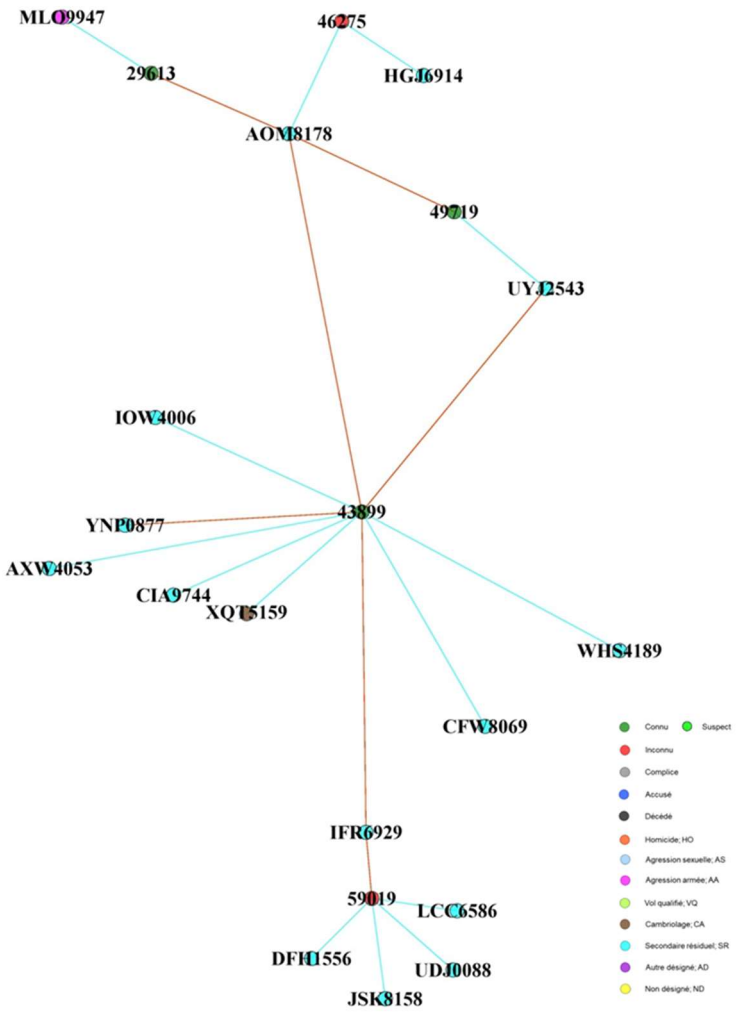


Figure 6A : Le composant 44 et ses liens connus grâce aux concordances ADN du LSJML (en bleu) où certains dossiers ont aussi fait l'objet d'une enquête par la police (en orange). Les numéros anonymisés de dossiers du LSJML sont précédés de trois lettres (p. ex. : IOW4006) et ceux des individus sont composés de 5 chiffres (p. ex. : 43899). Pour les types de délits, se référer au code de couleur de la Fig. 3.

Le croisement avec les données policières a permis d'ajouter 27 individus dont plusieurs sont en lien avec le délit AOM8178 où les ADN de deux individus connus et un autre d'identité inconnue (46275) avaient été mis en évidence au LSJML (Fig. 6B). On poursuit en comparant l'identification des nouveaux individus à ce qui figure dans les fiches du LSJML afin de vérifier s'il n'y aurait pas d'autres délits où leur ADN aurait été détecté. Cette vérification s'est avérée fructueuse pour les deux individus C_86 et C_42 qui, respectivement, se retrouvent au LSJML dans les fiches d'identification 23439 et 63916. (Voir les flèches rouges de la Fig. 6B)

L'exercice démontre qu'il y a au LSJML des composants ADN vus comme isolés, qui sont dans les faits très proches d'autres composants, mais dont les liens ne sont détectés qu'avec l'ajout des informations policières. En effet, l'ADN de l'individu C_42 (fiche 63916) n'est pas retrouvé dans le délit IFR6929, mais ailleurs dans UEN6851 et ZUW4346 qui sont inscrits dans une autre fiche correspondant au composant de l'individu C_42 du LSJML. Le délit UEN6851 est aussi lié à un troisième individu, le 29229 qui sera identifié uniquement par l'enquête policière et auquel on aura donné le numéro de C_227. Sur C_86 (fiche 23439) deux autres délits ADN du LJML sont reliés et il s'agit ici aussi d'un second composant qui s'intègre au composant 44 (Fig. 6B).

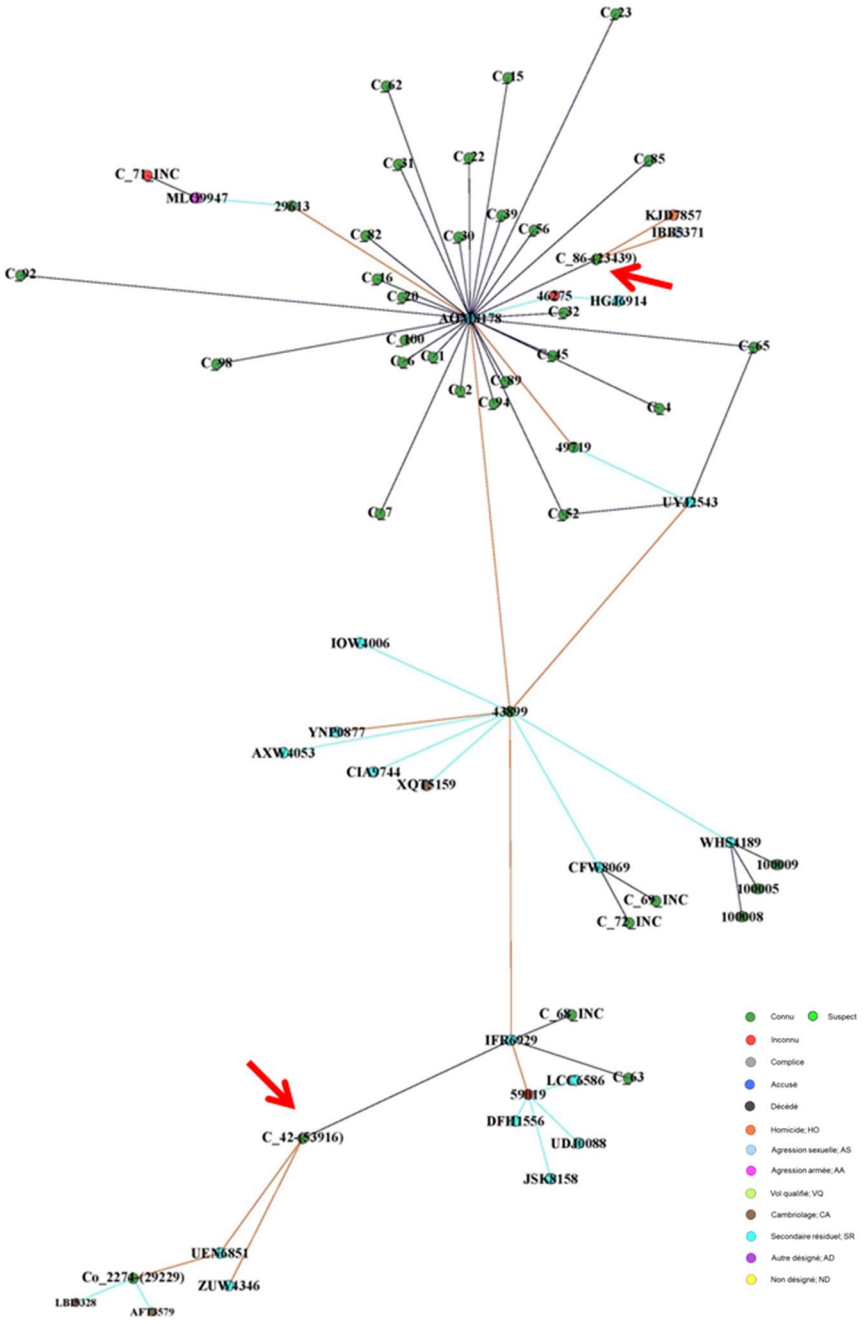


Figure 6B : Le composant 44 de la Figure 6A, auquel est ajouté un ensemble d'individus connus grâce aux enquêtes policières (liens noirs) et pour la plupart accusés (sauf ceux se terminant par « INC » pour inconnu) (étapes B et C, Fig.2). Pour ce qui est des informations policières, les numéros d'individus varient (ex : C_23, 100008). Les deux flèches rouges pointent vers les individus vus dans deux autres composants du LSJML n'ayant qu'un lien « policier » (sans ADN) avec le composant à l'étude.

La Figure 6C présente le résultat du second ajout d'informations qui concernent tous des délits associés aux individus ajoutés à l'étape précédente. On prend la mesure de l'ampleur du volume des affaires policières où peu d'entre elles se retrouvent au laboratoire pour une analyse génétique.

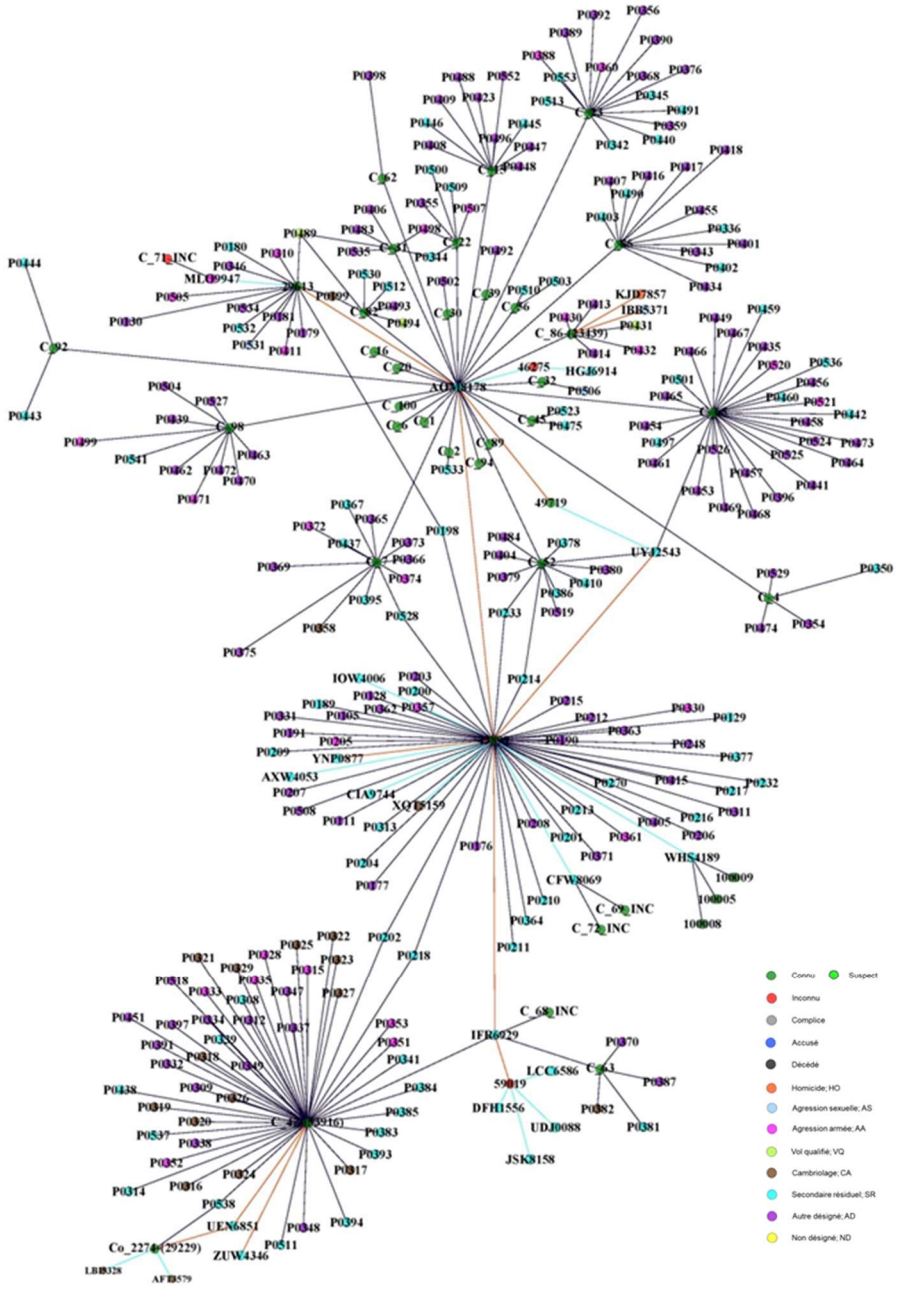


Figure 6C : Le composant 44 de la Figure 6B auquel on a ajouté tous les dossiers d'enquêtes policières (suffixe PO) associés aux individus ajoutés à la première étape. (étape E, Fig.2)

On constatera qu'après ce deuxième ajout d'informations, la structure du composant est grandement alourdie et qu'il y aurait un potentiel de délits suffisant autour des inconnus pour y chercher des informations pertinentes à leur identification. Est-il besoin de chercher les individus associés à ces délits supplémentaires ? La Figure 6D permet de visualiser l'ensemble des complices associés à ces dossiers de police. On comprendra que les enquêtes entourant les délits de trafic de stupéfiants ont énormément de ramifications dans la société et que la mise en réseau des individus impliqués de près ou de loin à ces activités devient rapidement très complexe. Toutefois la mise en réseau des informations permet de mettre en évidence certains aspects, comme les individus C_62, C_92 et C_4 qui sont reliés à quatre délits ayant chacun de très nombreux complices. Ces trois individus font partie de la première cohorte d'ajouts d'informations et sont tous reliés au délit AOM8178 où de nombreux autres individus avaient été ajoutés (Fig. 6B). Un tel ajout d'informations peut apporter de nouvelles pistes pour le développement d'une enquête.

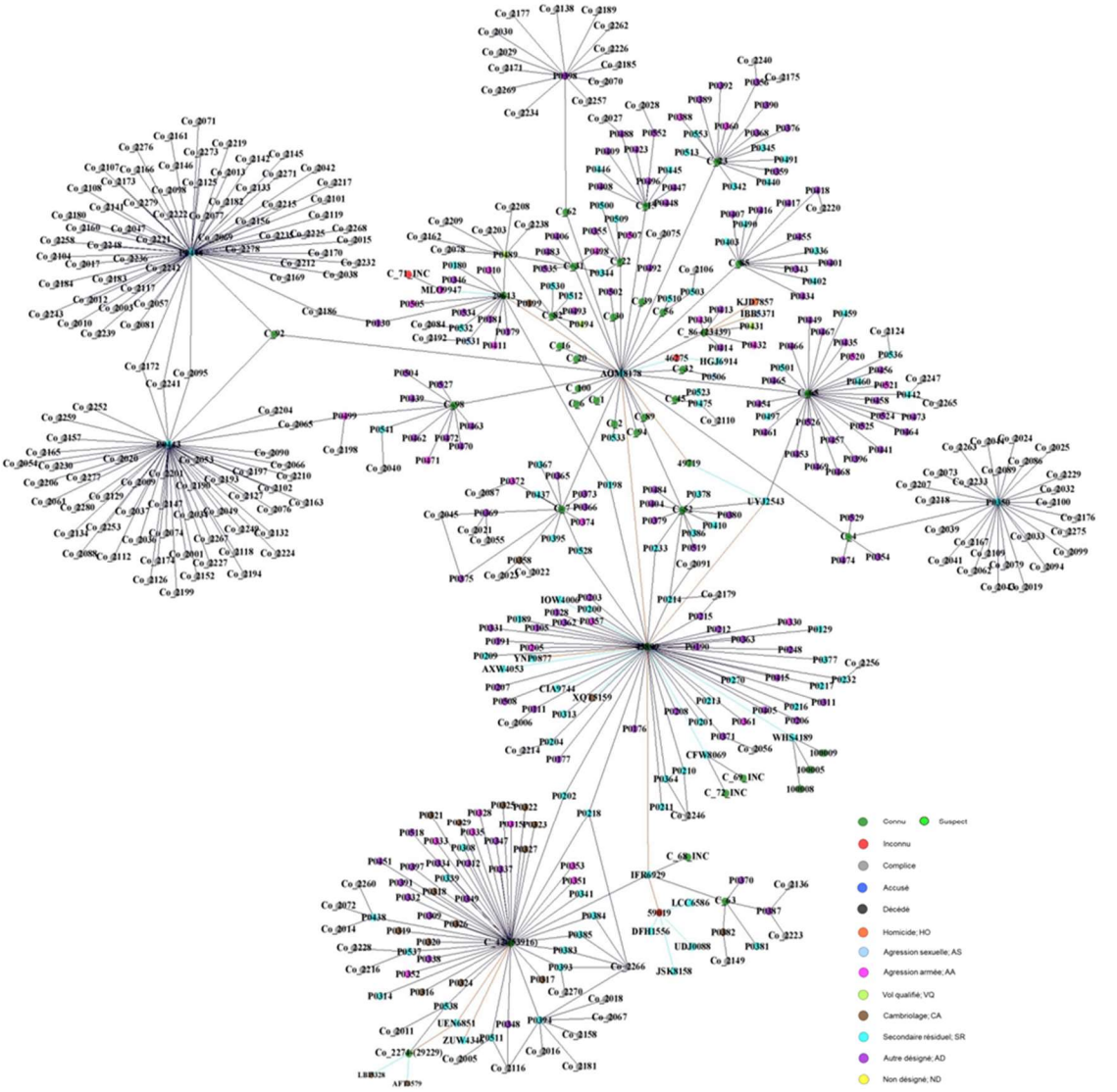


Figure 6D : Le composant 44 de la Figure 6C auquel on a ajouté tous les complices (Co) identifiés dans les délits ajoutés à la deuxième étape (étape F, Fig. 2).

4.3.1.3 Le composant 87

Les Figures 7A à 7C reprennent la structure de base (ADN) du composant 87. Les informations, obtenues grâce à l'ajout fait à la première étape, y sont très importantes (Fig. 7B). Les nombreux dossiers supplémentaires se retrouvent dans les catégories de cambriolages (CA), secondaires résiduels (SR), autres désignés (AD) et vols qualifiés (VQ) dont plusieurs en co-délinquance. Deux accusés (100001, 100007), en lien avec le dossier d'agression sexuelle où se trouve l'inconnu (52752), se sont ajoutés au

composant (Fig. 7B). Les analyses ADN ont aussi permis de relier cet inconnu à un délit de vol qualifié (LQX5830,) via une co-délinquance avec les individus 20946 et 52230, qui eux partageaient un deuxième vol qualifié (DEN6151). À cela s'ajoutent, grâce aux données policières, deux liens supplémentaires de co-délinquance de ces deux derniers individus sur des vols qualifiés. Ceux-ci et l'inconnu 52752 partagent un type d'activité commun, ainsi, la probabilité qu'ils soient co-délinquants sur ces vols qualifiés doit être prise en considération (Fig. 7B), tout en étant absents des données ADN. Ce regroupement de vols qualifiés autour de trois individus est une caractéristique du composant qui attire l'attention. Ce genre d'observation est de celles qui demandent d'être approfondies et c'est ce qui sera fait un peu plus loin dans l'analyse dynamique.

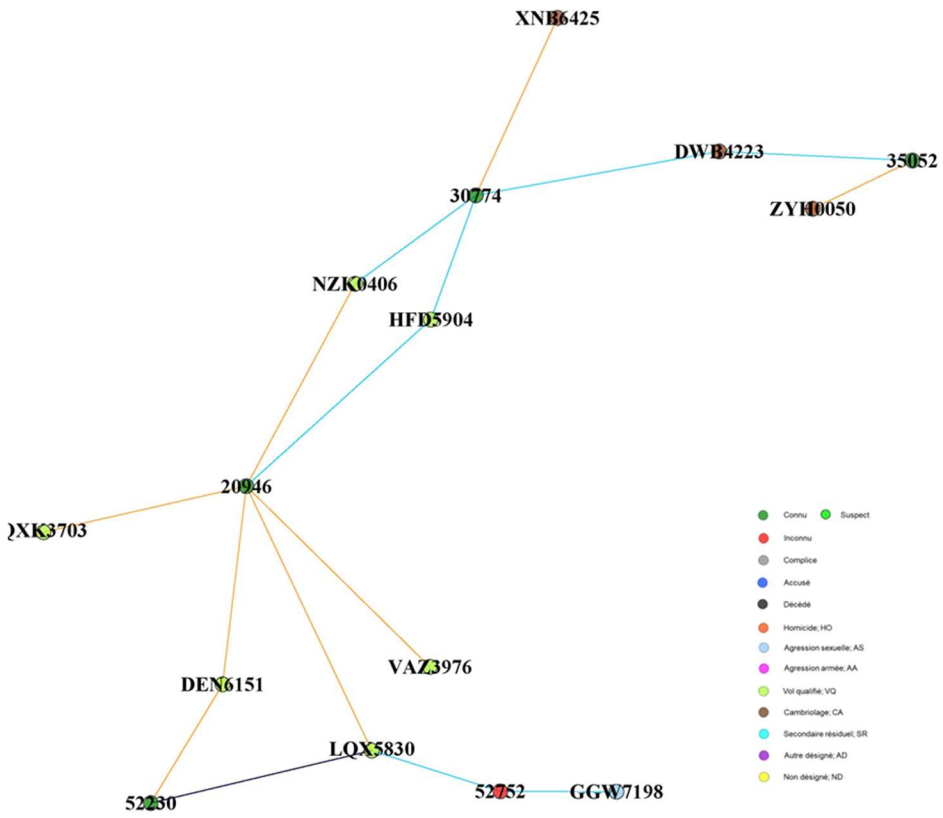


Figure 7A : Le composant 87 version ADN de base avec les numéros anonymisés pour le suivi.

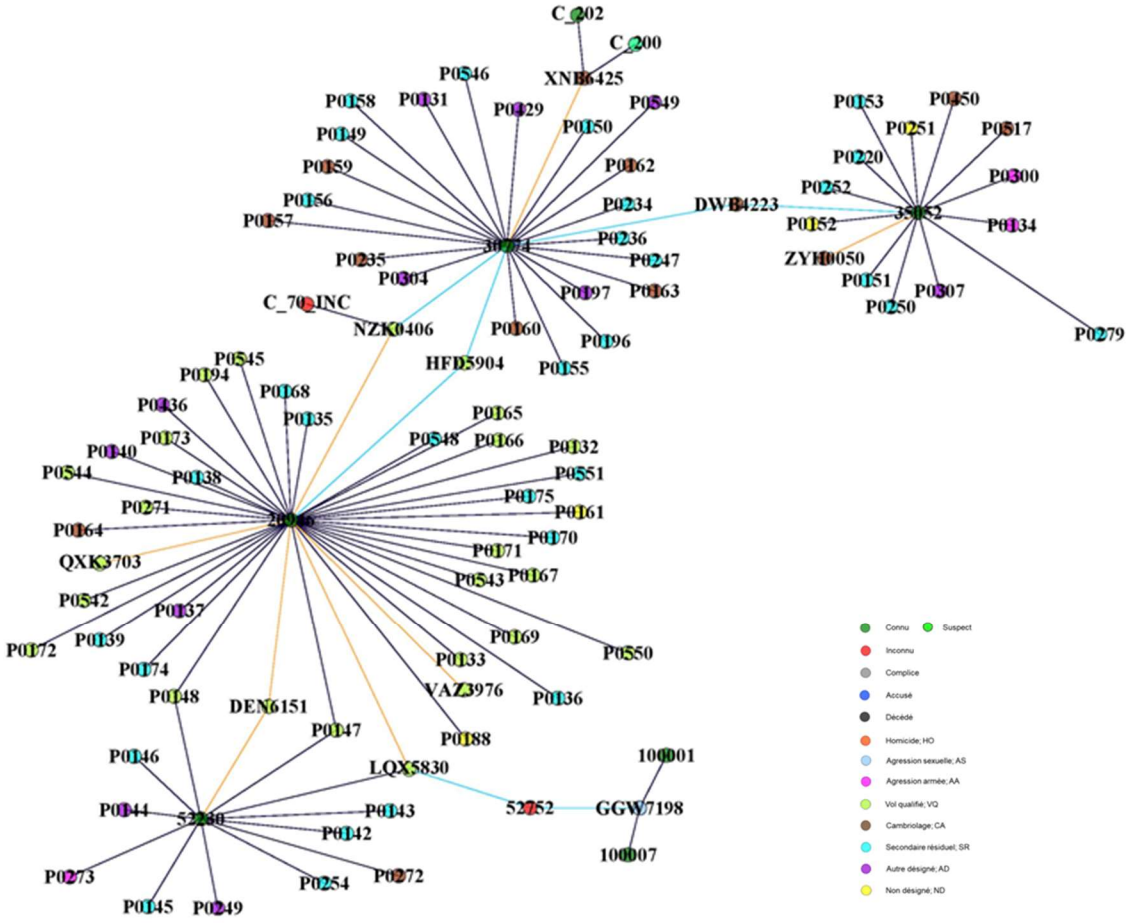


Figure 7B : Le composant 87 après la première étape d'ajout d'informations; délits criminels et individus. (Étape B et C, Fig. 2)

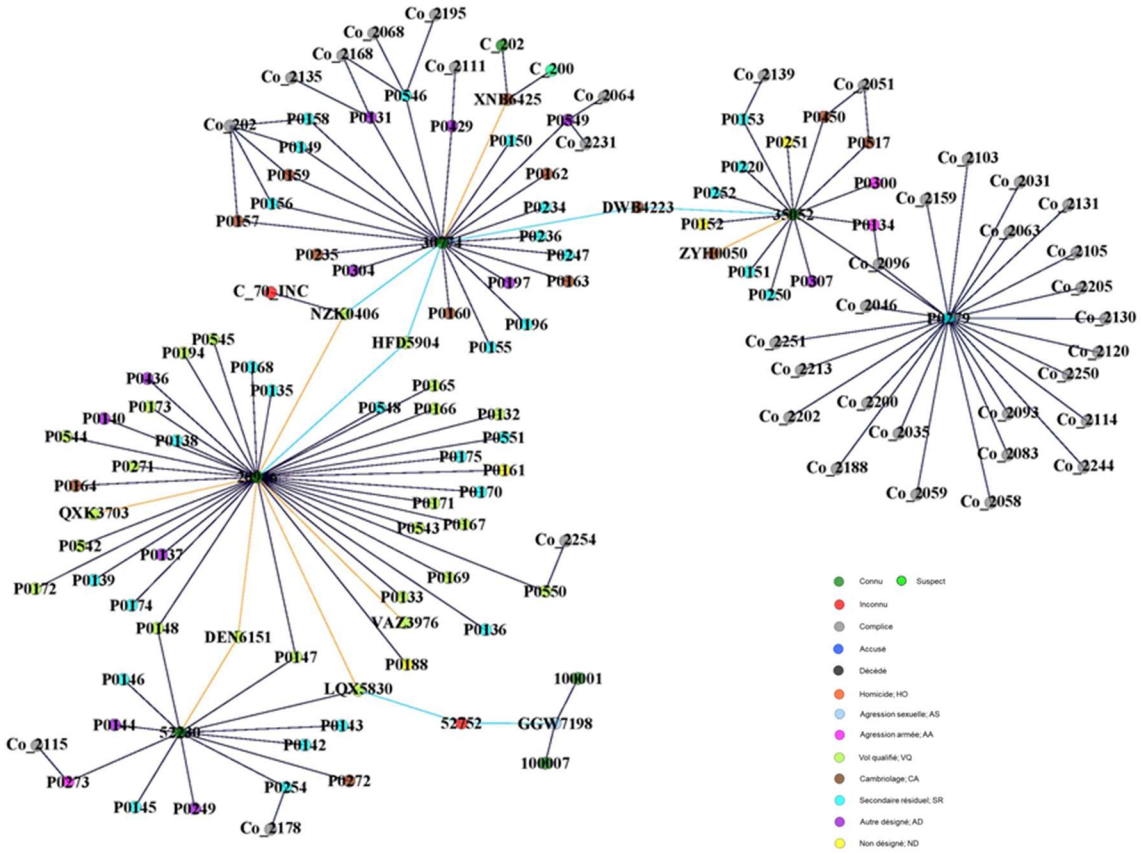


Figure 7C : Le composant 87 après la dernière étape d'ajout d'informations, les complices. (étape F, Fig. 2)

À la dernière étape de production de renseignement (Fig. 7C), l'ajout des complices permet de positionner un grand nombre d'entre eux sur le délit PO279 où l'un des complices est aussi en lien avec le délit PO134. Ailleurs dans le composant, on distingue quelques liens multiples entre plusieurs délits et quelques complices, comme on peut le constater avec les délits PO131 et PO546. Seuls trois complices ont été identifiés dans les nombreux délits associés aux individus 20946 et 52230 plus spécialisés en vols qualifiés. Nous verrons plus loin le renseignement que l'on pourra avancer avec cette structure d'informations.

4.3.1.4 Le composant 366

Les Figures 8A à 8C montrent les versions obtenues à la suite de l'ajout d'informations policières au composant ADN de base. À la première étape d'ajout

d'informations, cinq nouveaux individus s'intègrent dans ce composant (voir les flèches rouges de la Fig. 8B). À noter que trois de ces derniers accusés (100002, 100003 et 100004) sont associés à plusieurs délits, dont certains en co-délinquance, mais reliés à un seul individu connu (49890) par l'agression armée YAZ8242. C'est aussi un dossier du même type, le ZSB3023 qui relie l'inconnu 53388 à l'individu connu qui est en lien avec ces trois accusés. Ainsi, de même qu'avec l'exemple du composant 87, on retrouve une série de délits de même nature chez des individus connus qui sont en co-délinquance de proximité avec un inconnu.

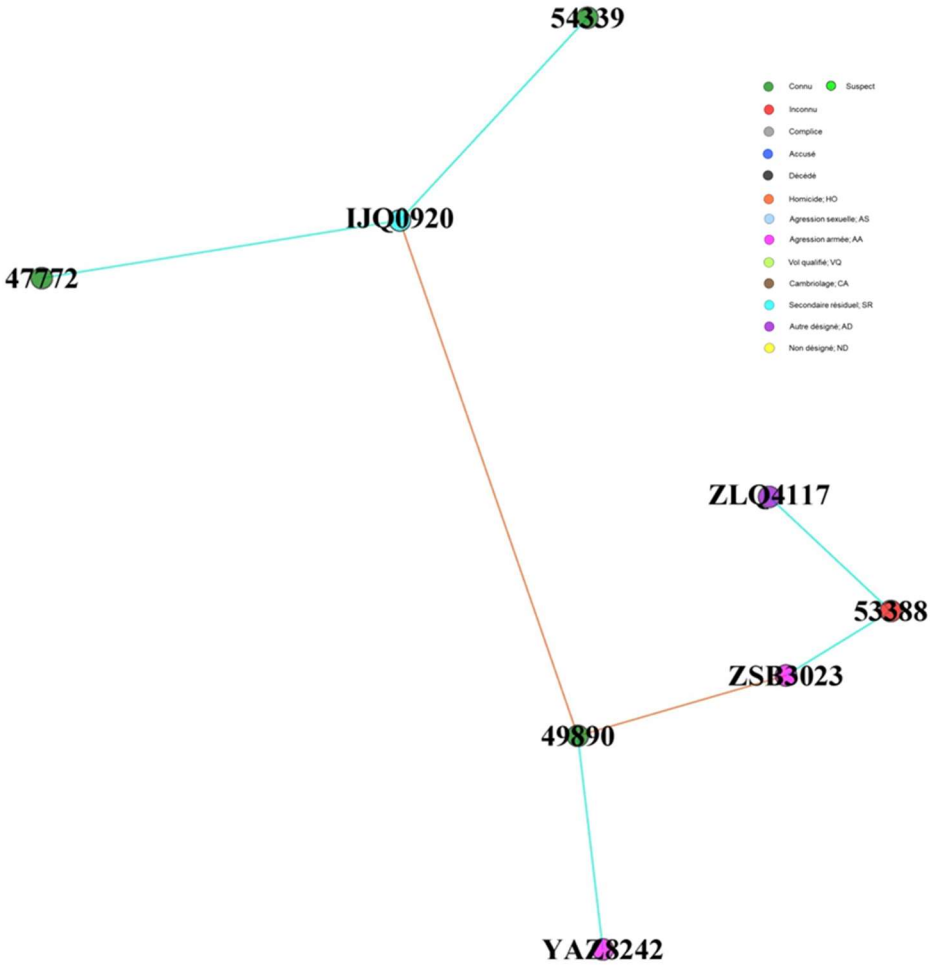


Figure 8A : Le composant 366 : version ADN de base.

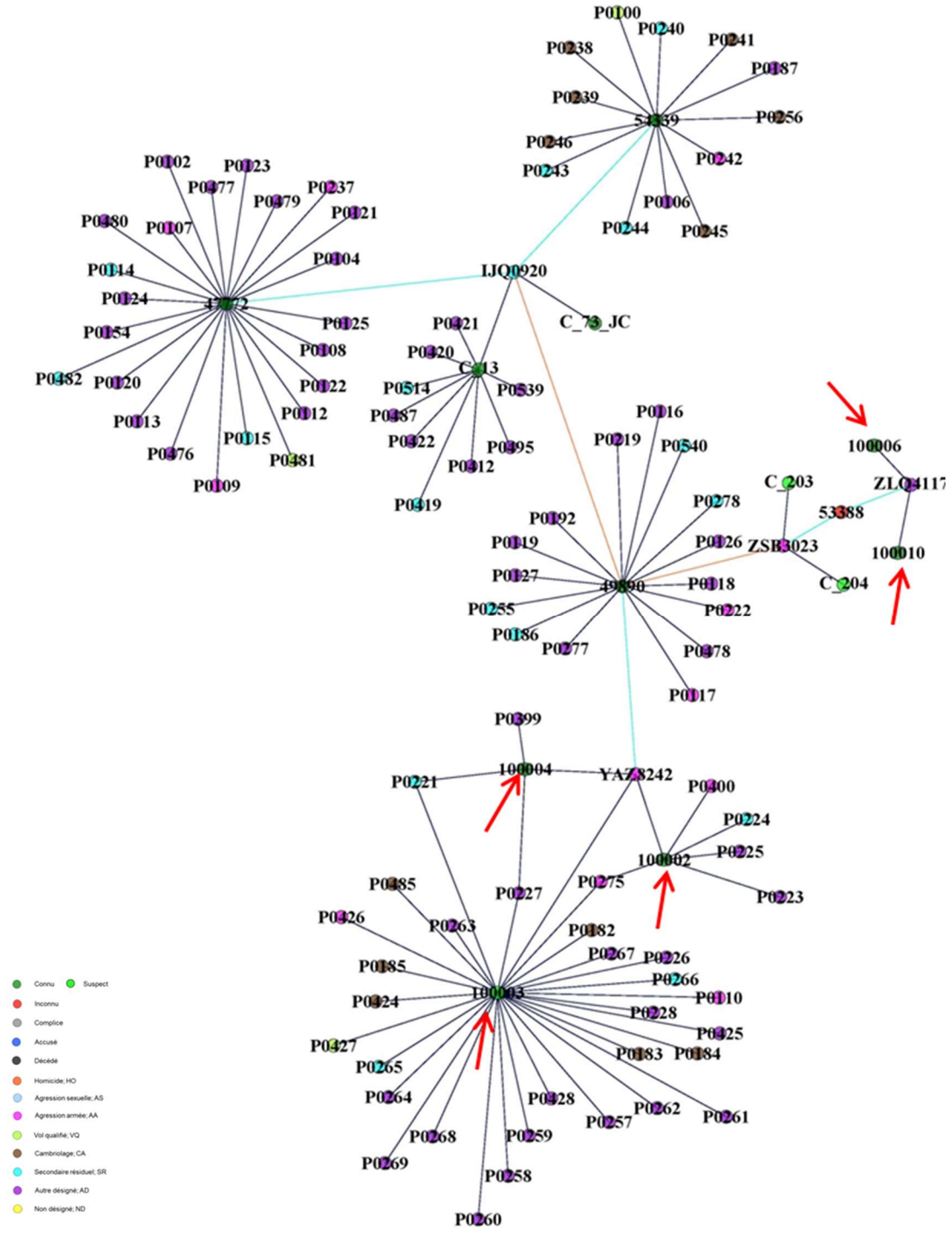


Figure 8B : Le composant 366 : la première étape d'ajout d'informations policières. (Étape B et C, Fig.2)

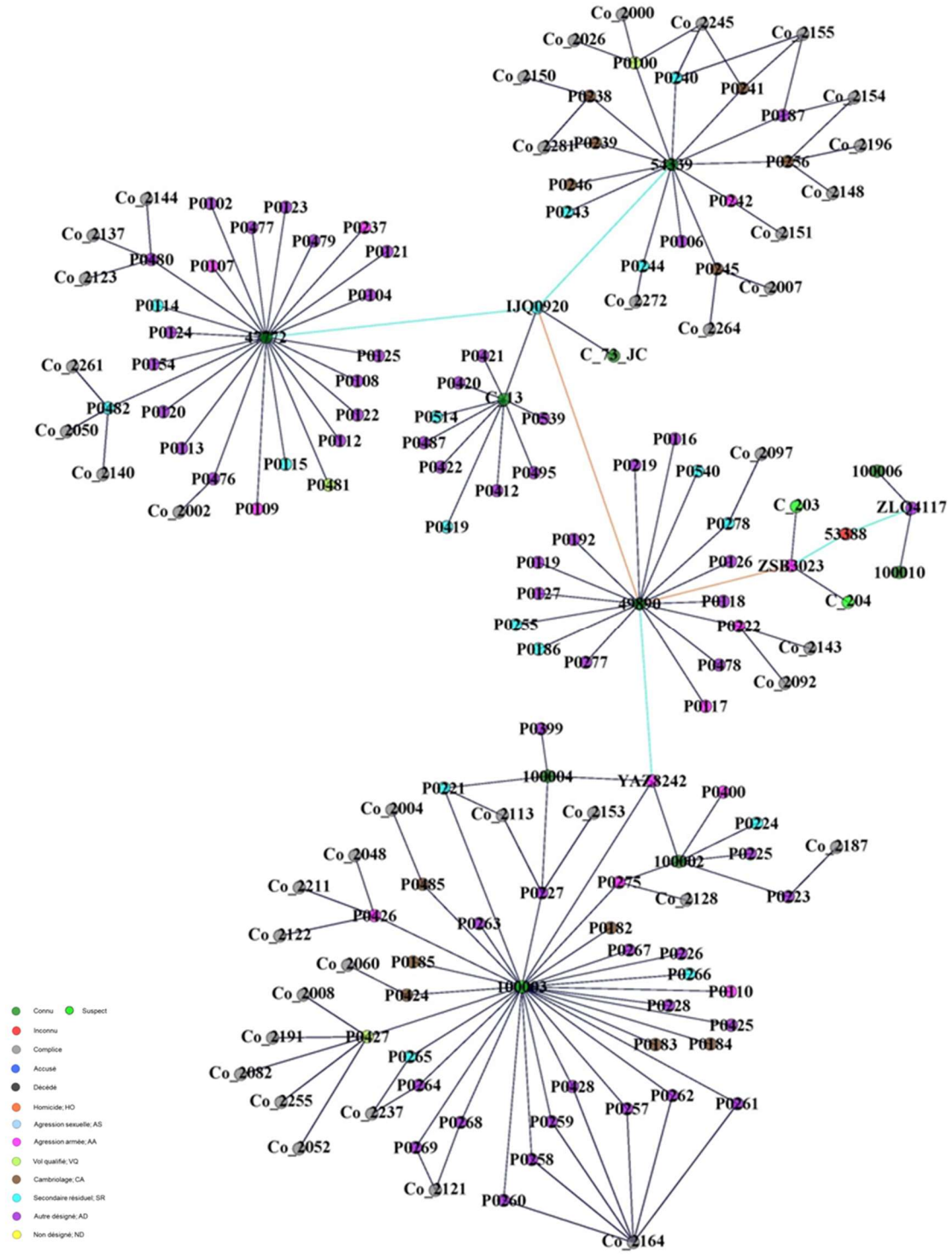


Figure 8C : Le composant 366 : la dernière étape d'ajouts d'informations : les complices. (Étape F, Fig. 2)

4.3.1.5 Recentrage sur les délits présentant le plus d'intérêt

Devant la complexité de l'assemblage de ces informations dans les composants 87 et 366, il est nécessaire de se concentrer sur les co-délinquances liées aux inconnus (Fig. 9 et 10). Ainsi, suite à l'ajout global d'informations, les regroupements de délits proches des inconnus doivent être traités de manière plus approfondie. Pour apporter plus de précision à ces co-délinquances de proximité, il faut détailler la dynamique du composant. L'utilisation de la date d'événement permet d'introduire une structure dynamique aux composants, et les résultats obtenus par cette approche temporelle sont présentés dans la section suivante.

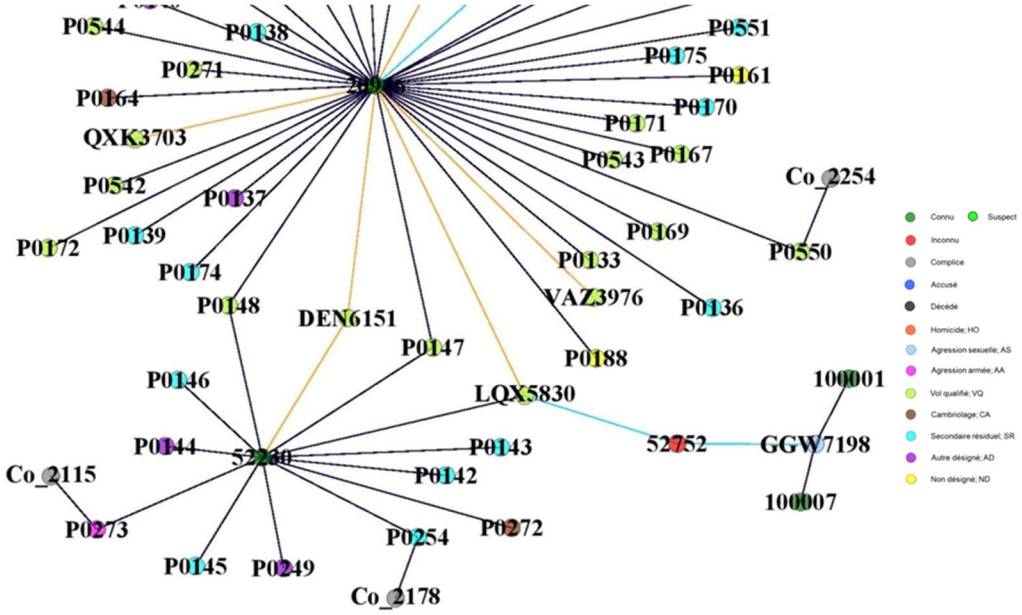


Figure 9 : Détail du composant 87 autour de l'inconnu co-délinquant sur un vol qualifié avec deux individus qui eux en partagent 3 autres.

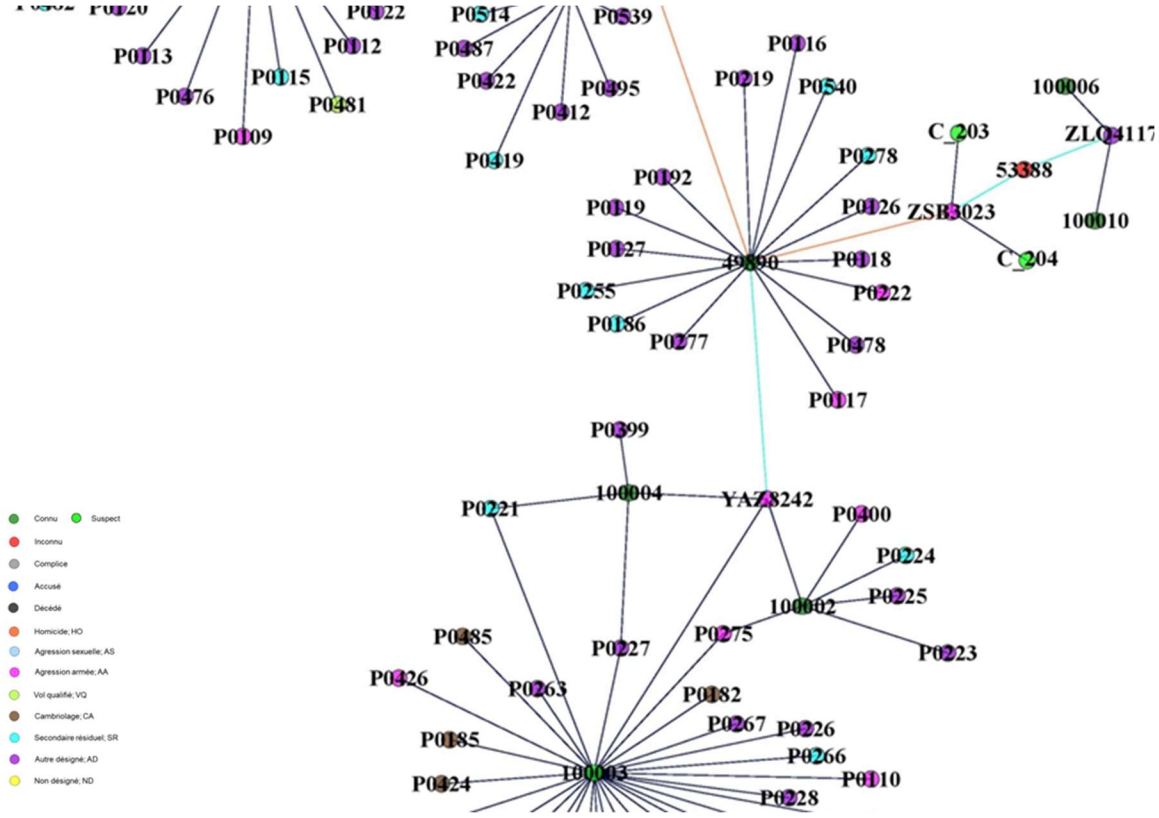


Figure 10 : Détail du composant 366 autour de l'inconnu ayant commis une agression armée en co-délinquance avec un individu qui est aussi co-délinquant dans le même type de délit impliquant trois autres contrevenants.

4.3.2 Dynamique des composants

Trois des quatre composants choisis, pour étudier notre approche de production de renseignement, présentent des situations de co-délinquance avec un inconnu, pertinentes pour une analyse dynamique, chacune apportant un aspect différent. Les deux derniers composants exposés ci-dessus, soit le 87 et le 366, présentent une structure commune de co-délinquance sur des délits de même type, associant l'inconnu et des individus avoisinants. À ces derniers, s'ajoute le composant 44 qui, malgré sa grande complexité, présente un aspect différent qui sera approfondi en utilisant une approche unimodale associée à un paramètre d'analyse de réseaux sociaux.

4.3.2.1 La dynamique du composant 87

En utilisant toutes les informations concernant le composant 87 que l'on retrouve à la Figure 7C et en y intégrant les dates d'événements, on arrive à reconstruire le composant en suivant l'ordre d'apparition des délits[18]. Les Figures 11A à 11E présentent le composant 87 sur 8 périodes entre l'an 1 et l'an 1 +14 (15 ans). On constate à première vue que le composant est plus éclaté dans les premières années où les dossiers s'accumulent lentement. Au-delà de la cinquième année, (AN 1 +4) et jusqu'à la fin de septième, (AN 1 +6), les individus 20946 et 30774 accumulent déjà un bon nombre de délits.

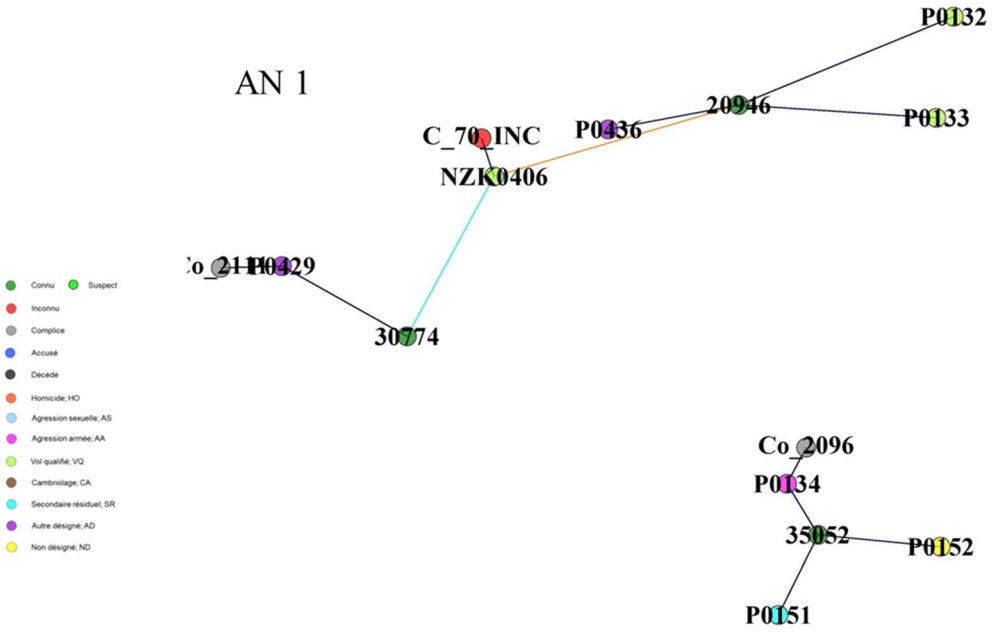


Figure 11A : La première des 15 années d'évolution du composant 87. Les groupes de délits ou d'individus sont ceux observés jusqu'à la fin de l'année identifiée.

[18] Dans cet exercice, les années seront codées pour conserver l'anonymat des données.

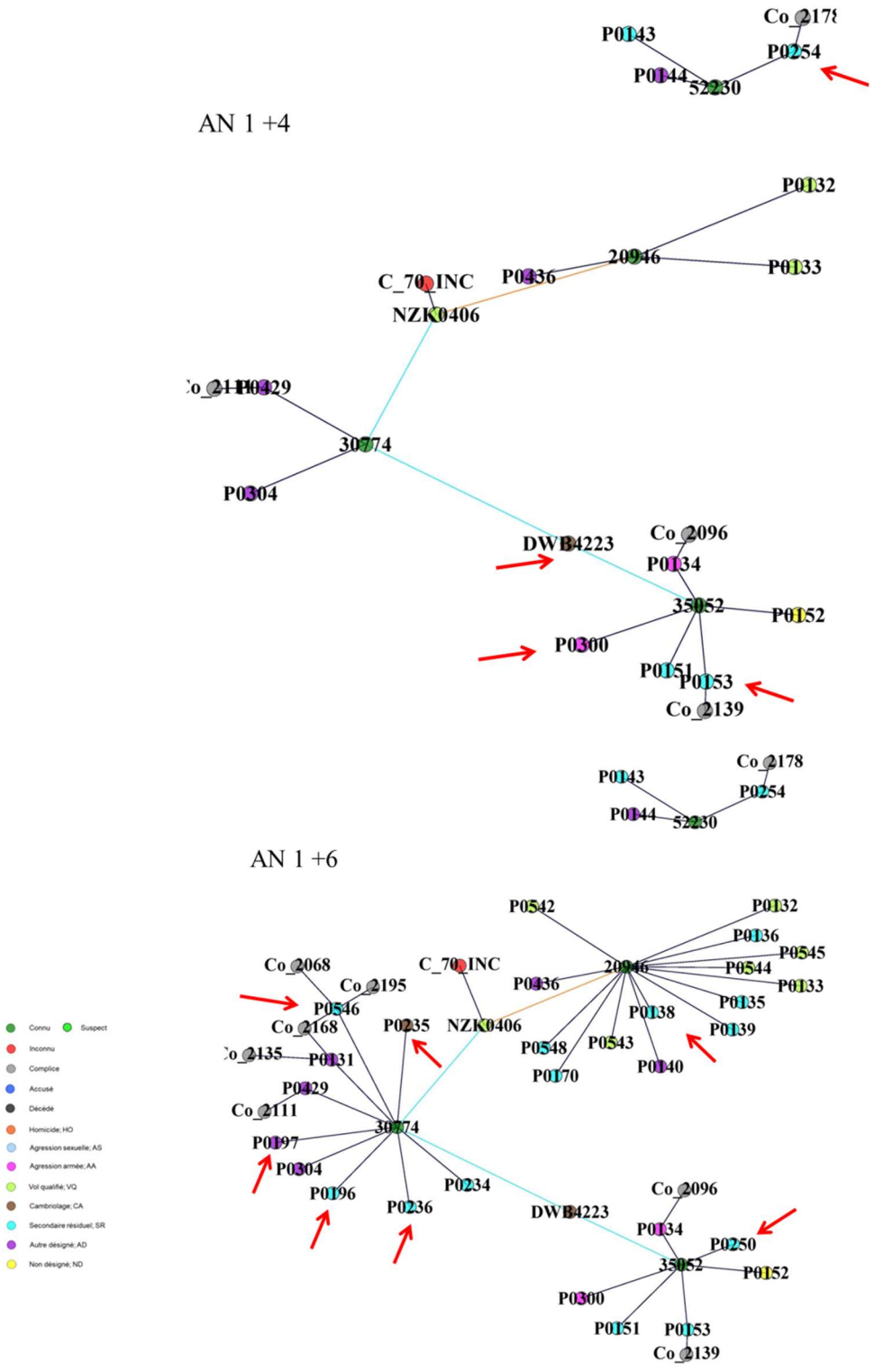


Figure 11B : Les cinquième et septième années sur les 15 ans d'évolution du composant 87. Les flèches rouges indiquent des ajouts de délits, de groupe de délits ou d'individus jusqu'à la fin de l'année identifiée.

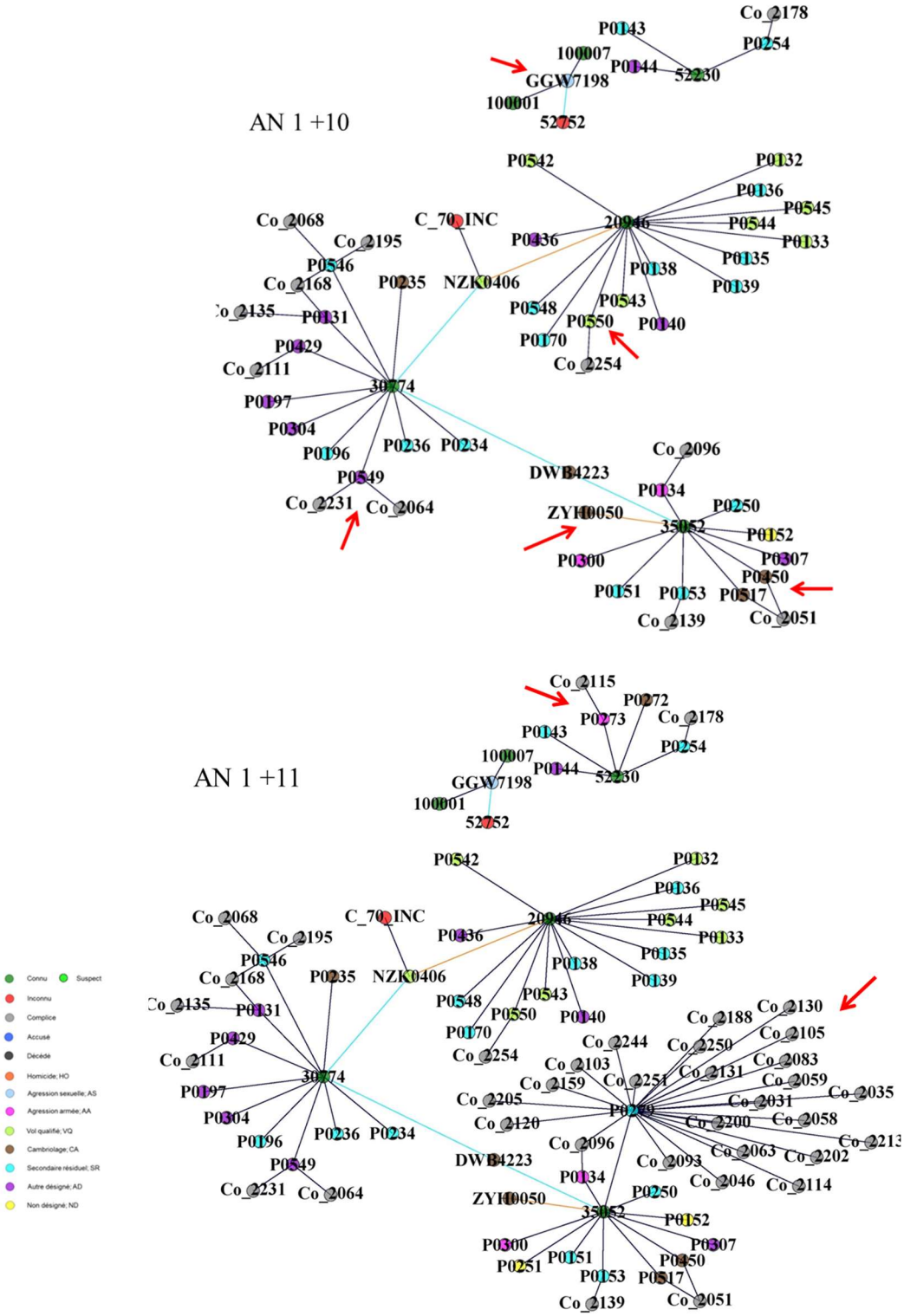


Figure 11C : Les onzième et douzième années sur les 15 ans d'évolution du composant 87. Les flèches rouges indiquent des ajouts de délits, de groupe de délits ou d'individus jusqu'à la fin de l'année identifiée.

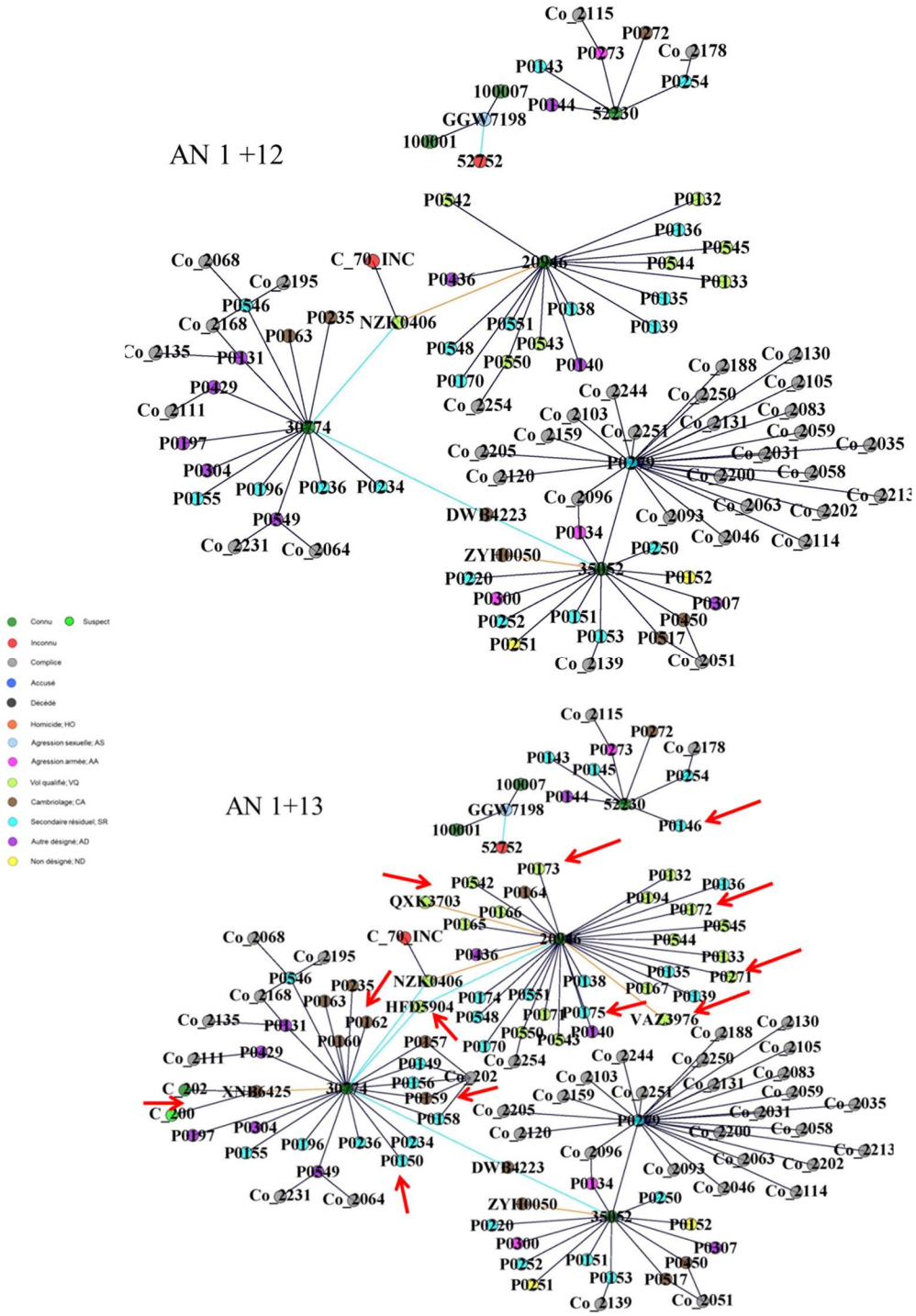


Figure 11D : Les treizième et quatorzième années sur les 15 ans d'évolution du composant 87. Les flèches rouges indiquent des ajouts de délits, de groupe de délits ou d'individus jusqu'à la fin de l'année identifiée.

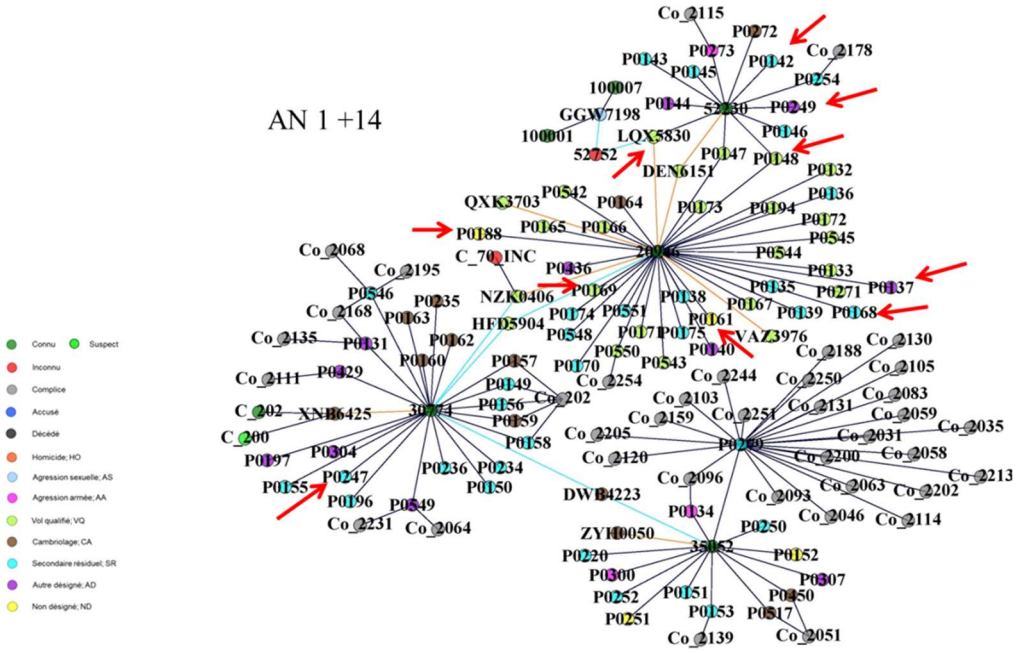


Figure 11E : La dernière année des 15 ans d'évolution du composant 87. Les flèches rouges indiquent des ajouts de délits, de groupe de délits ou d'individus jusqu'à la fin de l'année identifiée. La zone d'intérêt entre l'an1 +13 et +14 se trouve dans l'espace libre du haut où 4 vols qualifiés en co-délinquance (deux sont pointés) s'ajoutent dans un laps de temps très court.

À l'an 1, l'individu 20946 compte, à son actif, trois vols qualifiés, à la fin l'an 1 +6 il en est à sept (Fig. 11B). Dans toute cette évolution sur 15 ans, l'inconnu 52752 n'apparait qu'en l'an 1 +10 en lien avec un délit d'agression sexuelle (qui s'avérera être aussi un vol qualifié) avec deux individus connus des policiers (Fig. 11C). Ce petit groupe ne s'intégrera au composant 87 que lors de la quinzième année (an 1 +14) via un vol qualifié (LQX5830) où l'on retrouve un résultat ADN de cet inconnu en co-délinquance avec l'individu 20946 (Fig. 11E). Dans les données policières, les résultats de l'enquête associent un autre individu connu, le 52230, au délit LQX5830. Ce dernier individu est connu des policiers et aussi par son ADN dans un autre vol qualifié, le DEN6151, qu'il partage avec l'individu 20946. Entre les cas de l'an 1 +13 et +14, on remarque la zone libre dans le haut qui est comblée par quatre vols qualifiés (Fig. 11E), dont les deux présentant des résultats d'ADN qui ont été nommés dans les phrases précédentes. Les deux autres qui s'ajoutent (PO147 et PO148) apparaissent tous approximativement en même temps que les deux autres ayant les données ADN.

Un regroupement de vols qualifiés dans un laps de temps court attire l'attention et, pour mieux évaluer la situation, il est alors intéressant de procéder à une évaluation de cette portion des VQ du composant 87, comme ils sont présentés dans la Figure 12 avec les dates d'événement et leur localisation[19].

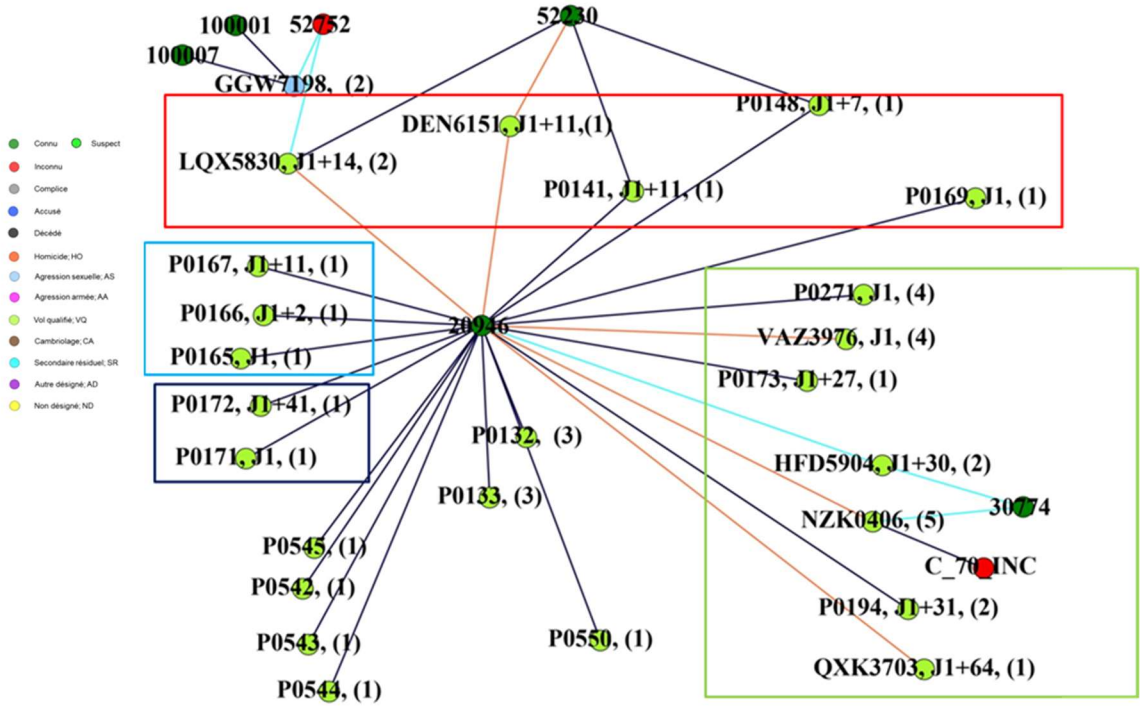


Figure 12 : Détail sur les vols qualifiés du composant 87, avec des dates d'événement en jour J (J +1 etc.) et des localisations rapprochées décrites par numéros ((1), (2) etc.) Les encadrements regroupent des délits à des dates rapprochées et sont décrits dans le texte.

Sur cet ensemble, on constate que les 4 VQ, avec co-délinquance, mentionnés au paragraphe précédent, ont eu lieu à des dates rapprochées, du J1 au J1 +14, au cours de l'an 1 +14 dans les deux régions limitrophes (1) et (2) (Fig. 12, encadrement rouge). De plus, un autre VQ solo (PO169), le premier de la période, a aussi eu lieu dans la même région. En y regardant de plus près, on remarque que trois autres délits (PO165, PO166 et PO167) ont été commis un mois plus tôt, ceux-là aussi dans un période de 12 jours (Fig. 12, encadrement bleu). De surcroît, ces derniers ont aussi été perpétrés dans la

[19] Toujours pour conserver l'anonymat des données, les dates sont codées pour chaque bloc d'information sans mentionner les mois. Il en est de même pour les lieux qui sont codés.

région (1). Si la nécessité de pousser un peu plus loin s'avérait nécessaire, on constate que la série des VQ s'étire aussi dans les mois qui suivent en s'étalant sur une plus longue période (Fig. 12, encadrement vert) et de même pour une autre petite série (Fig. 12, encadrement noir). Avec l'ajout d'informations de localisation, quoique grossières, on constate que l'ensemble de cette longue série de VQ qui, lorsqu'elle n'a pas lieu dans la région (1), se concentre dans trois secteurs des environs, soit (2), (3) et (4). Face à une concentration de délits si forte, un enquêteur pourrait présumer que l'inconnu 52752 aurait très bien pu participé à quelques autres vols qualifiés de cette série de l'an 1 +14, voire de l'an 1 +13, en se joignant aux mêmes complices et que la présence de son ADN sur un seul délit n'aurait été pour lui que malencontreuse. C'est, à cette étape qu'entre en jeu la recherche d'informations policières du deuxième type. Que pourrait-on trouver dans ces dossiers qui pourraient établir des rapprochements ? Des traces de pas, des témoignages? Les dossiers de ces délits ont été fouillés plus à fond pour tenter d'y trouver des éléments circonstanciels qui pourraient aider à solutionner l'identification de cet inconnu. Après avoir traité de la dynamique des deux autres composants, nous y reviendrons.

4.3.2.2 La dynamique du composant 366

La reconstruction temporelle du composant 366 présenté aux Figures 13A à 13G est étalée sur huit étapes de 2003 à 2019. Ici aussi, on constatera un morcellement important du composant dans les premières années. L'inconnu 53388 apparaît en 2011 et c'est en 2016 qu'on le voit relié en co-délinquance avec l'individu 49890 dans une agression armée, la ZSB3023. L'individu 49890, connu des policiers, a, à son actif, deux autres agressions armées, la PO222 et la YAZ8224, et cette dernière est en co-délinquance avec trois autres individus (100002, 100003 et 100004).

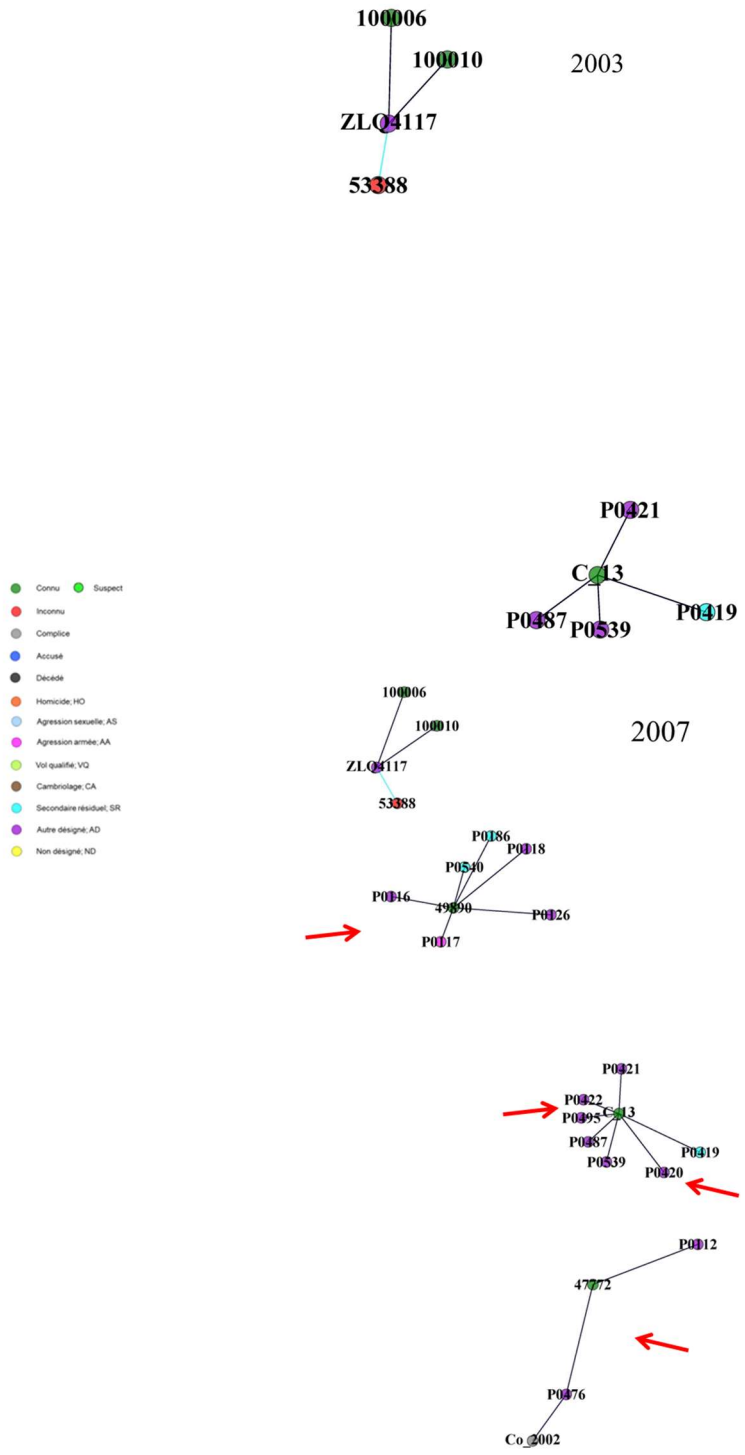


Figure 13A : Les années 2003 et 2007 des 16 années d'évolution du composant 366. Les flèches rouges indiquent des ajouts de délits, de groupes de délits ou d'individus jusqu'à la fin de l'année identifiée.

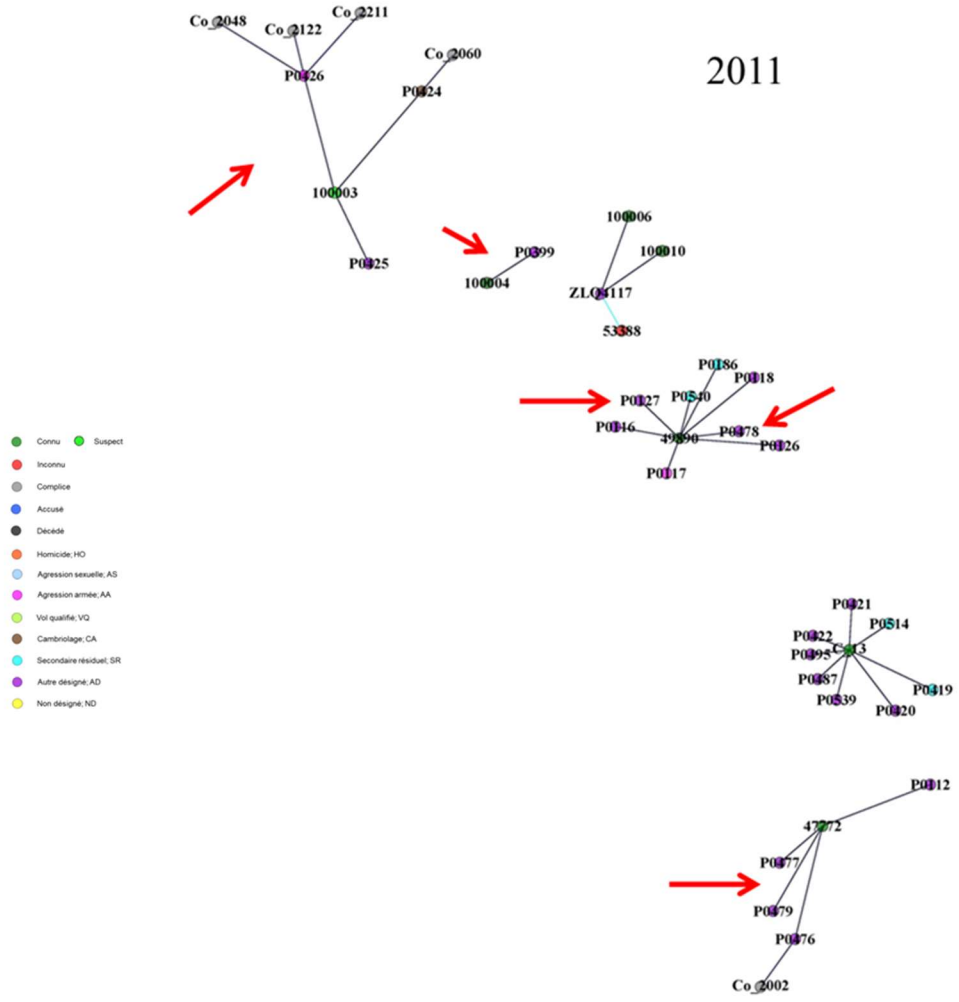


Figure 13B : L'année 2011 des 16 années d'évolution du composant 366. Les flèches rouges indiquent des ajouts de délits, de groupe de délits ou d'individus jusqu'à la fin de l'année identifiée.

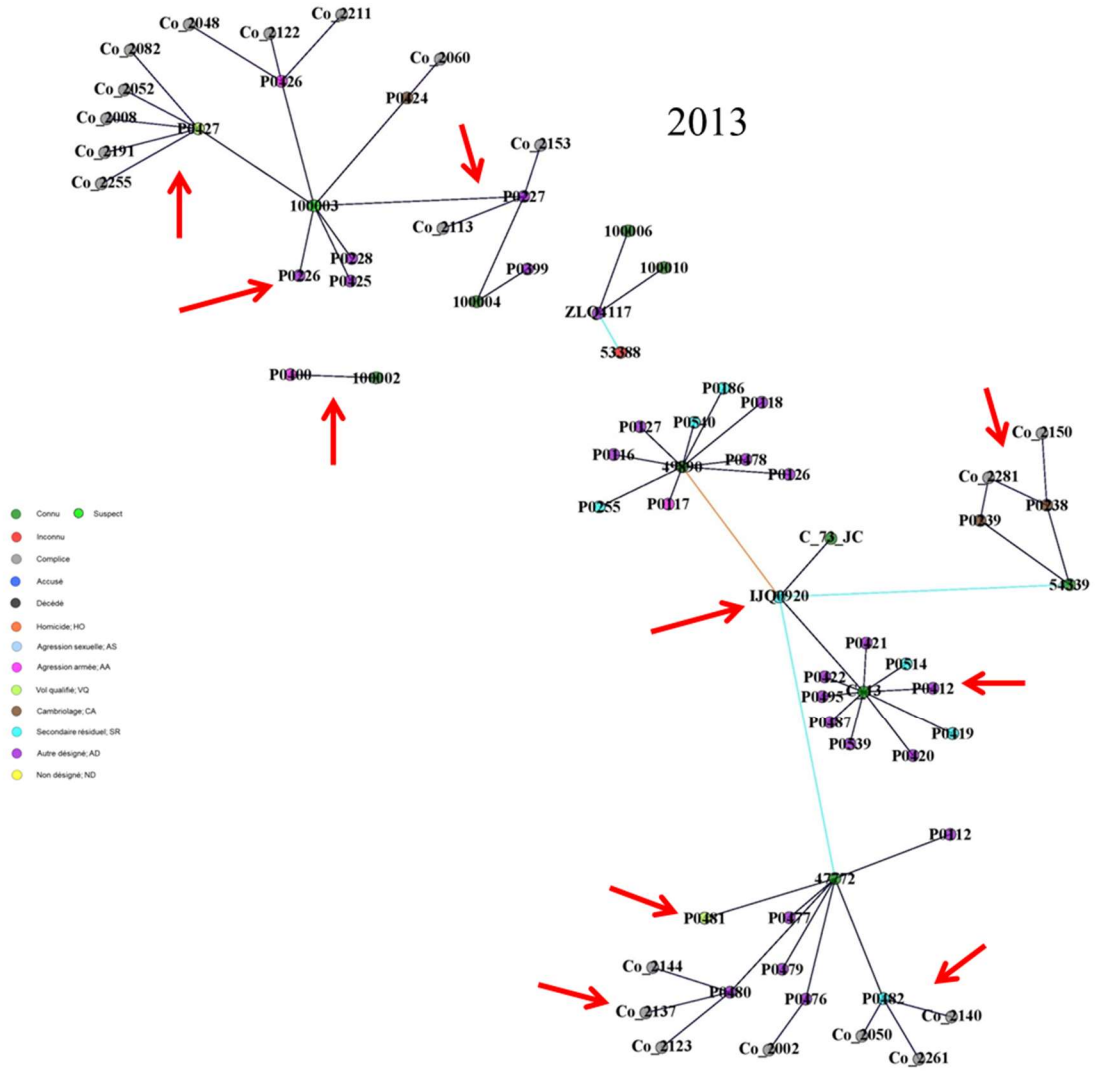


Figure 13C : L'année 2013 des 16 années d'évolution du composant 366. Les flèches rouges indiquent des ajouts de délits, de groupe de délits ou d'individus jusqu'à la fin de l'année identifiée.

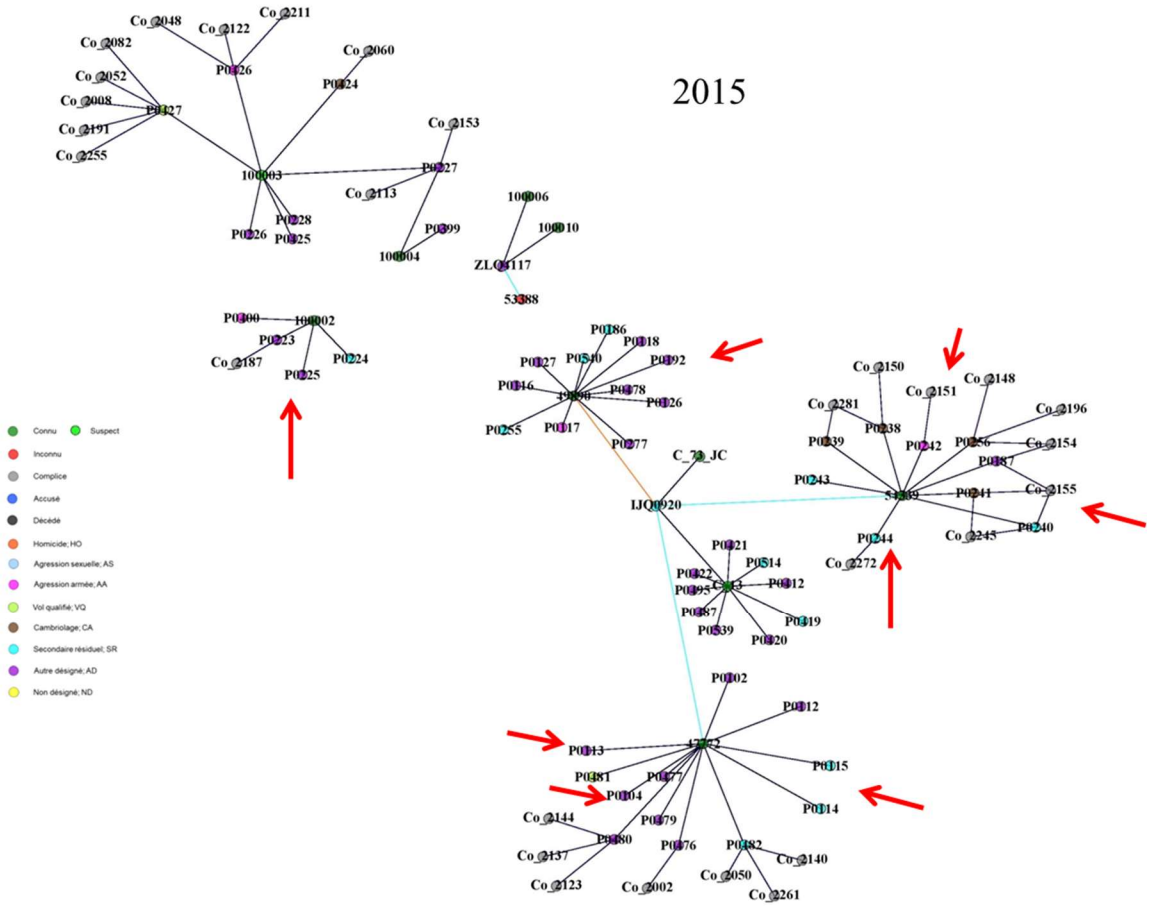


Figure 13D : L'année 2015 des 16 années d'évolution du composant 366. Les flèches rouges indiquent des ajouts de délits, de groupe de délits ou d'individus jusqu'à la fin de l'année identifiée. La zone d'intérêt se situe entre 2015 et 2016, où l'inconnu est relié à une agression armée en co-délinquance.

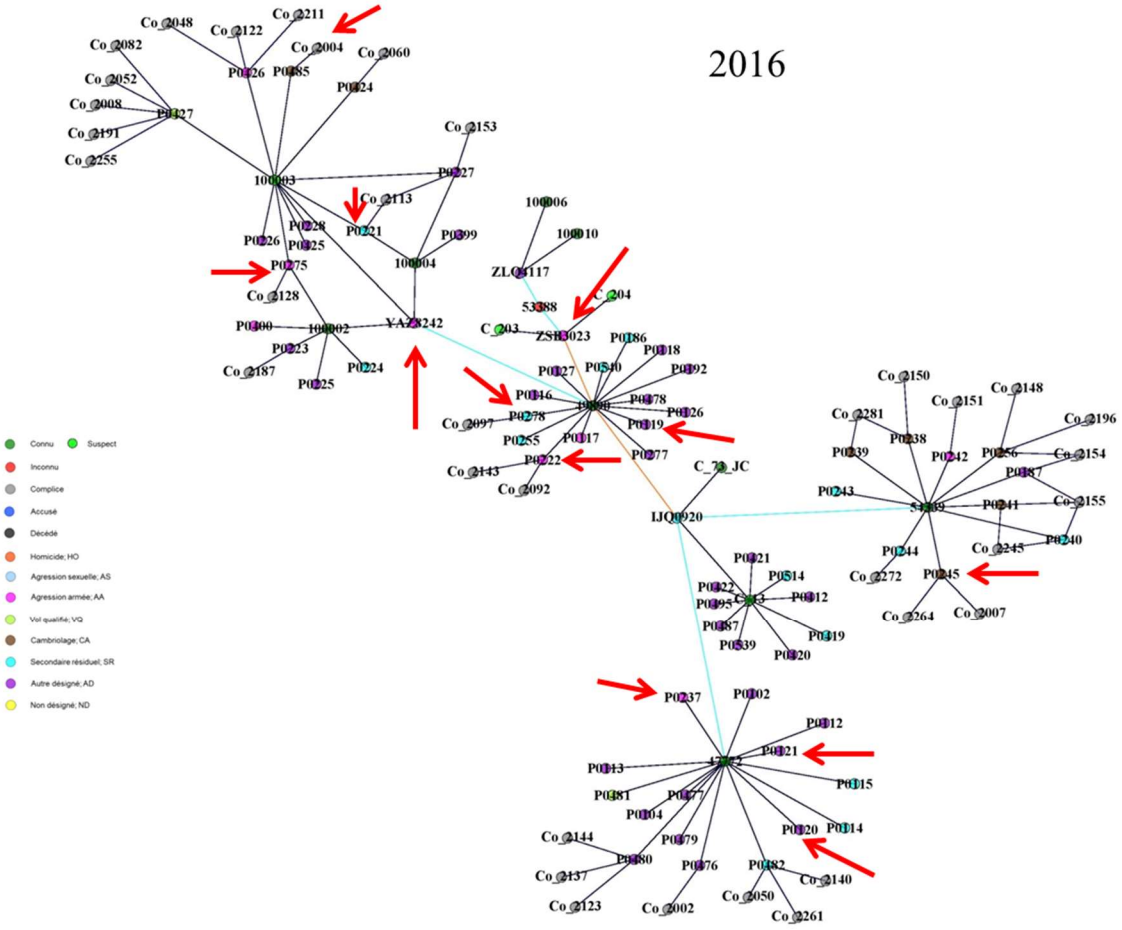


Figure 13E : L'année 2016 des 16 années d'évolution du composant 366. Les flèches rouges indiquent des ajouts de délits, de groupe de délits ou d'individus jusqu'à la fin de l'année identifiée. La zone d'intérêt se situe entre 2015 et 2016, où l'inconnu est relié à une agression armée en co-délinquance.

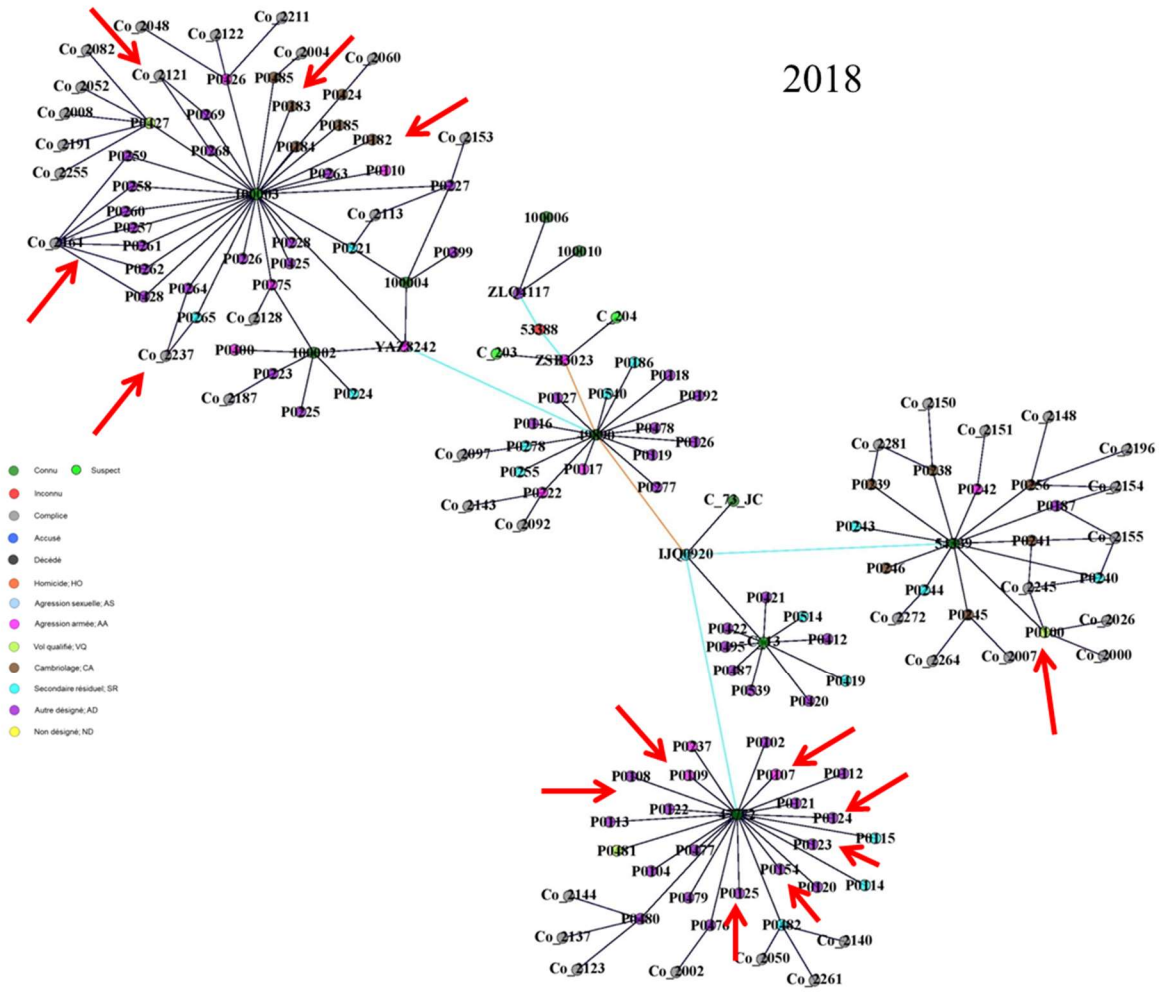


Figure 13F : L'année 2018 des 16 années d'évolution du composant 366. Les flèches rouges indiquent des ajouts de délits, de groupe de délits ou d'individus jusqu'à la fin de l'année identifiée.

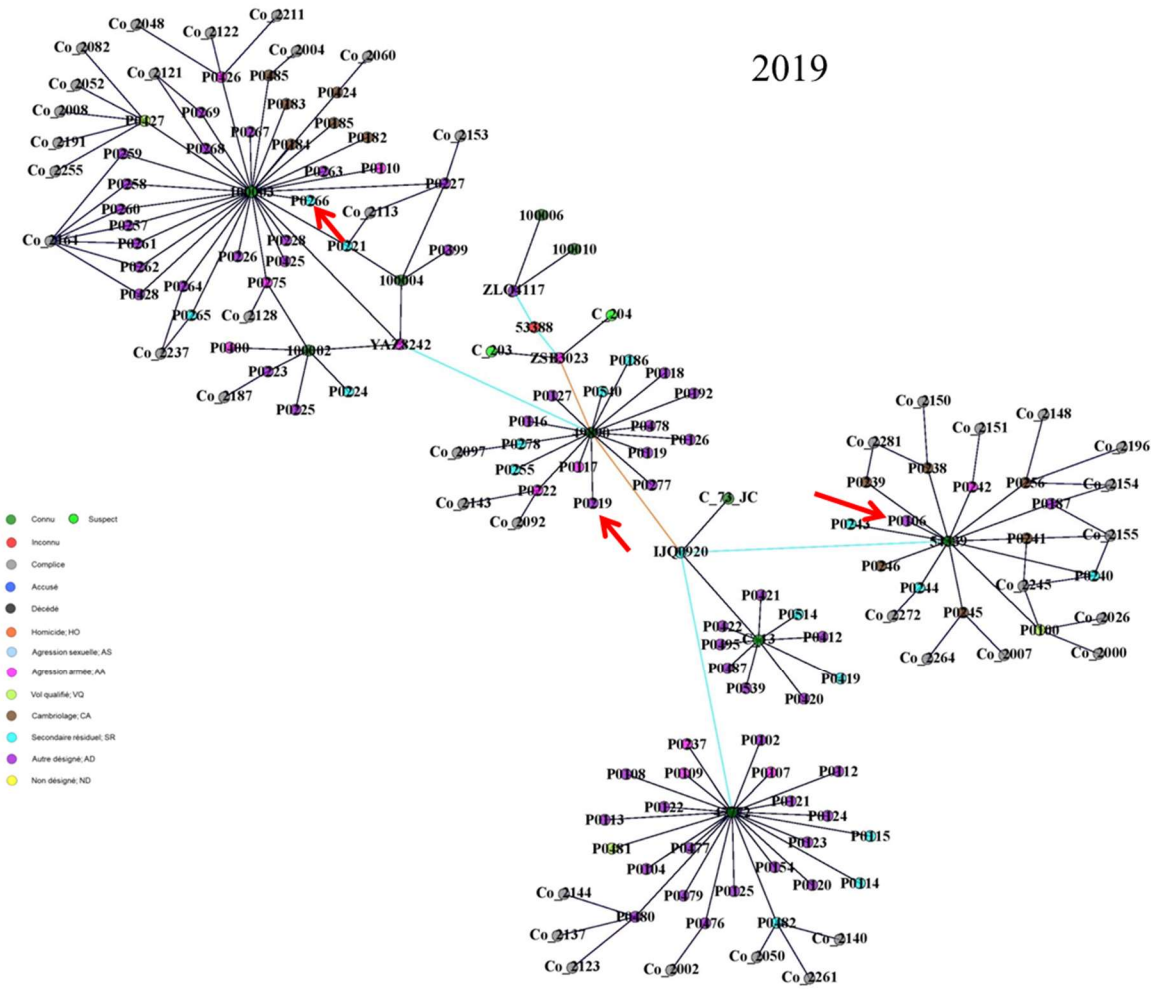


Figure 13G : L'année 2019 des 16 années d'évolution du composant 366. Les flèches rouges indiquent des ajouts de délits, de groupe de délits ou d'individus jusqu'à la fin de l'année identifiée.

Ces trois agressions armées sont toutes apparues sur 7 jours en 2016 dans des régions rapprochées. Ici encore, la proximité temporelle ainsi que géographique permet d'envisager que l'inconnu 53338 ayant participé au délit ZSB3023 puisse avoir participé aux autres délits. Il devient alors pertinent de trouver d'autres éléments circonstanciels communs à ces délits qui pourraient aider à identifier cet inconnu. En poussant les recherches sur les agressions armées un peu plus loin dans le temps, quelques autres cas, du même type ou en régions rapprochées, pourraient être considérés afin d'y rechercher

un fil conducteur pour soutenir le développement de l'enquête. En effet, les délits PO426, PO400 et PO117 sont respectivement de 2011, 2013 et 2007.

4.3.2.3 La dynamique du composant 44

Comme mentionné un peu plus haut, lors de l'ajout d'informations policières, le composant 44 se démarque par les nombreux individus qui s'y greffent. En tenant compte de cet aspect, la dynamique de ce composant sera abordée sous un angle d'analyse de réseau social, en format unimodal axé sur les individus, plutôt que sous l'angle des délits, comme dans les exemples précédents. Notre attention se portera sur les trois inconnus de ce composant que l'on retrouve à la Figure 14, dans les données combinées, LSJML et policières

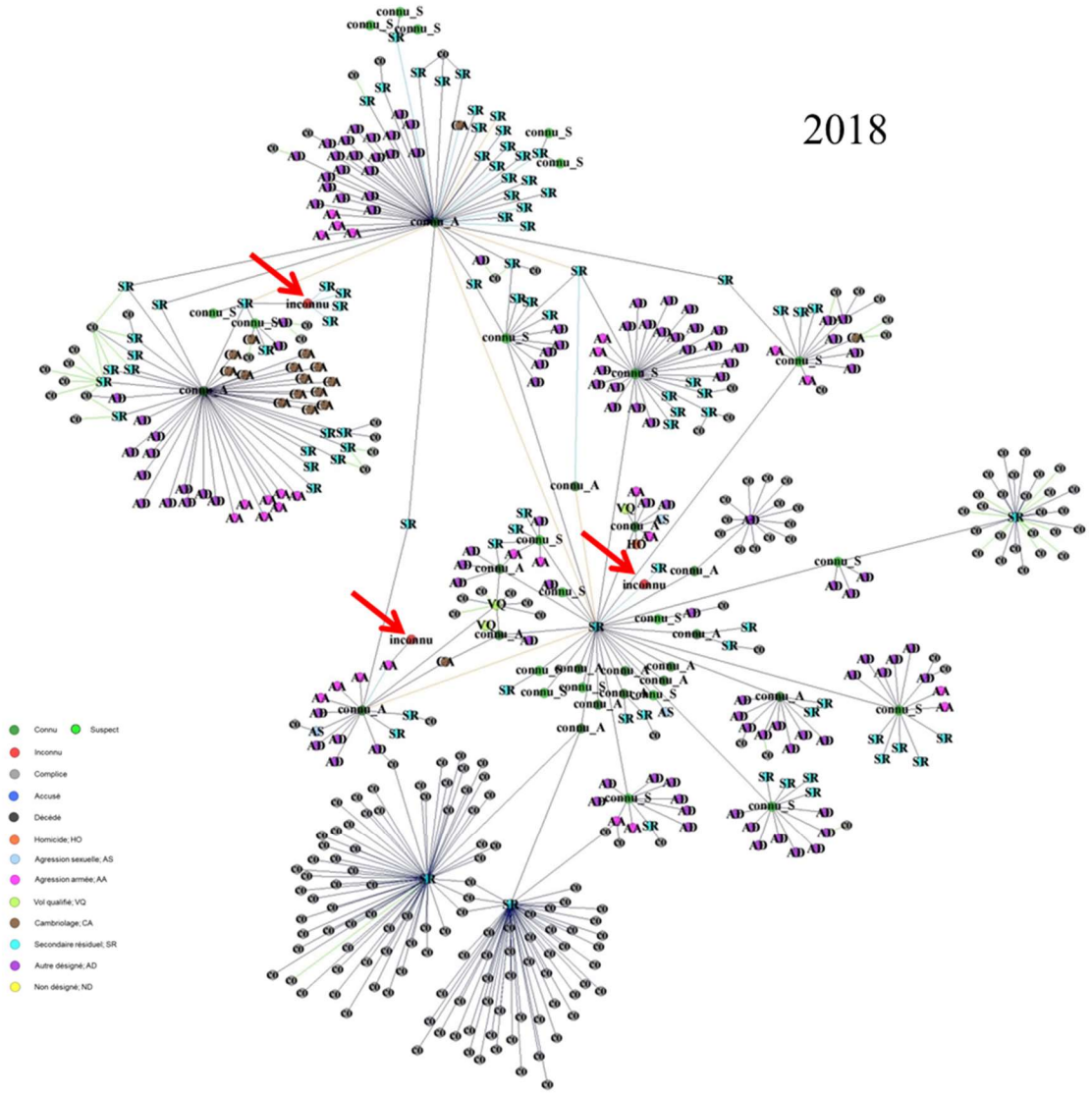


Figure 14 L'aboutissement de l'évolution du composant 44 en 2018; les individus et leurs délits criminels. Les flèches rouges indiquent la position des inconnus.

Ce type de structure complexe ayant de nombreux liens, délits et individus, peut se visualiser différemment en utilisant une version unimodale n'ayant que les individus comme seul type de nœud du schéma relationnel. Cette approche, puisqu'il ne reste que des individus, permet de faire une analyse des réseaux sociaux (ARS; voir le chapitre II). L'évaluation et l'évolution des valeurs du coefficient d'agglomération (*clustering*) utilisé au deuxième chapitre, sont présentées dans les Figures 15A et 15B dans les trois

clichés du composant 44, couvrant les années de 2012 à 2018. À première vue, les Figures 15A et B semblent tout aussi complexes, mais c'est sur les résultats d'ARS des inconnus que l'attention doit se concentrer.

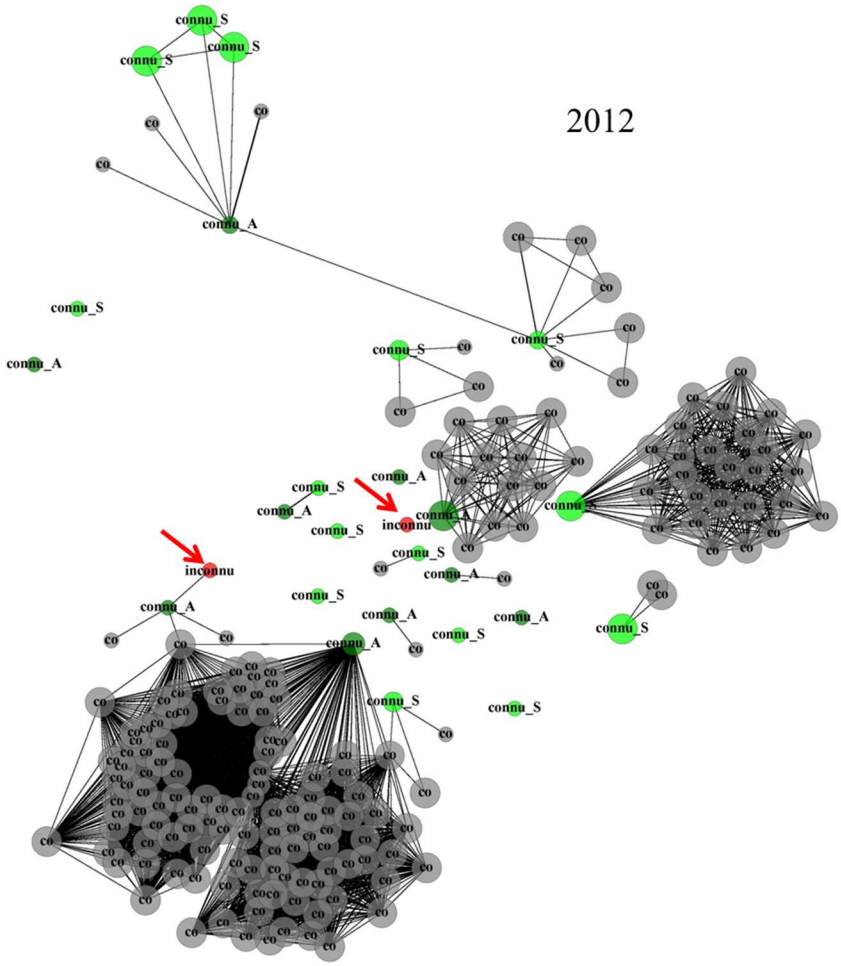


Figure 15A : Sociogramme de 2012, première année sur six ans d'évolution du composant 44, présenté dans un format unimodal ne montrant que les individus interreliés. La grosseur des nœuds est proportionnelle à la valeur du coefficient d'agglomération (*clustering*). Les flèches rouges indiquent les inconnus.

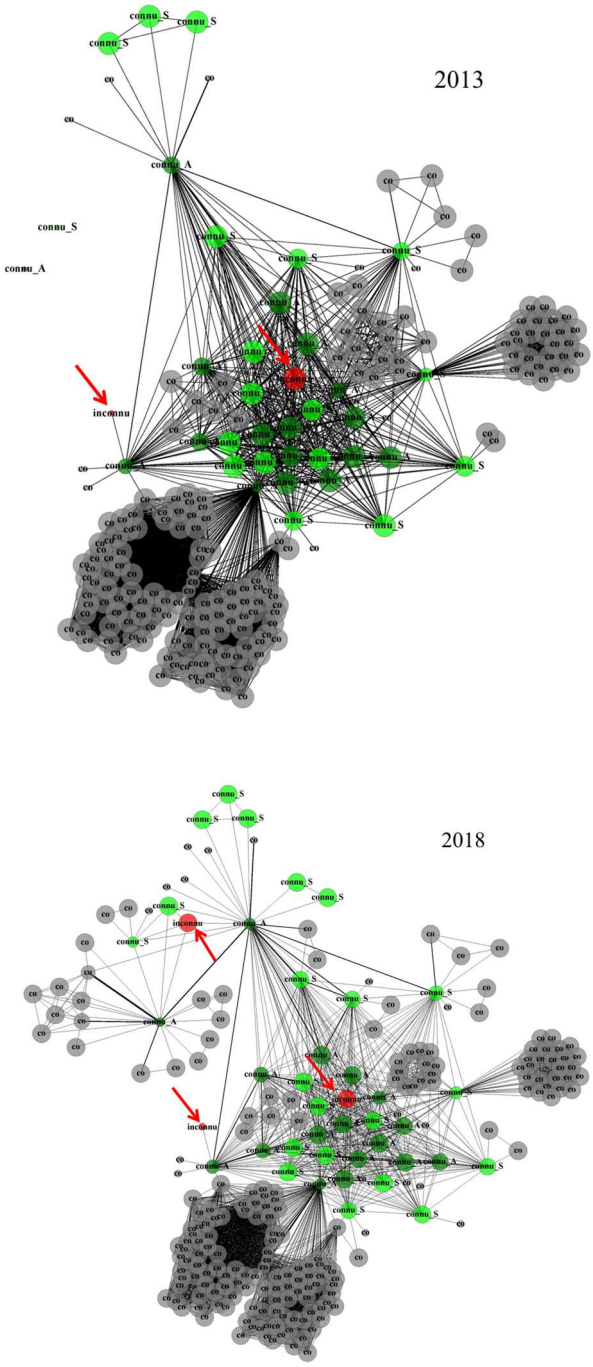


Figure 15B : Les sociogrammes des années 2013 et 2018 dans les six ans d'évolution du composant 44, présenté dans un format unimodal ne montrant que les individus interreliés. La grosseur des nœuds est proportionnelle à la valeur du coefficient d'agglomération (*clustering*). Les flèches rouges indiquent les inconnus.

Dans les clichés des Figures 15A et 15B, la valeur du coefficient d'agglomération[20], qui était l'un des paramètres ayant permis de démontrer l'intégration des inconnus dans les composants (voir chapitre II sections 3.4.1 et 3.4.2), est maintenant présentée de manière visuelle sur un ensemble d'individus. Le diamètre des nœuds (individus) est proportionnel à la valeur du coefficient d'agglomération. Ce paramètre permet de présenter les résultats sous un jour nouveau, montrant qu'en 2012 les deux inconnus qui sont positionnés au centre, malgré leurs différences dans le nombre de complices, sont très bien intégrés au composant, et ce, avec une valeur identique à celle que l'on retrouve chez la plupart des individus qui les entourent. En 2013, par contre, l'ajout de nombreux complices apportera une valeur d'agglomération beaucoup plus importante à l'inconnu positionné au centre (Fig. 15B, 2013) L'inconnu du haut, apparu en 2018, présente aussi une valeur d'agglomération importante alors que celui du bas à gauche reste toujours isolé. Notre exemple ne présente pas de changements notoires pour plusieurs des individus, comme ceux entourant l'inconnu central, et il en est de même pour l'intermédiarité[21] (valeurs non présentées). Toutefois, il s'agit ici d'une analyse bien spécifique au coefficient d'agglomération, qui donne un résultat qui est fonction de la structure du composant. Un ensemble de calcul d'ARS donnerait des informations variées qui pourraient répondre à diverses questions de recherche concernant le positionnement relatif des individus dans le composant. Dans notre exemple, notre inconnu (celui du centre) est lié à un délit rassemblant une trentaine de personnes, ce qui est en soi un élément majeur à intégrer dans une réflexion d'enquête, et l'évaluation du coefficient d'agglomération se présente comme une approche stratégique permettant de pointer vers cet individu. L'ajout de d'autres paramètres d'ARS pourrait supporter l'attention que le coefficient d'agglomération apporte à cet inconnu central, tout en mettant au jour une information supplémentaire qui pourrait pointer d'autres individus du composant. Le renseignement obtenu d'un

[20] Parfois aussi appelé coefficient d'agrégation (en anglais *clustering*). Il s'agit de la proportion des triangles (3 individus reliés fermés) par rapport au total des triangles possibles d'un composant

[21] L'intermédiarité est un paramètre d'ARS qui correspond au nombre de fois qu'un individu se retrouve sur le plus court chemin reliant toutes les paires d'individus présents dans un composant.

ensemble de paramètres d'ARS devient une étape stratégique dans le développement d'une enquête.

4.3.3 Recherche d'éléments circonstanciels

Au départ, l'idée conductrice autour du rassemblement d'informations dans le but de créer du renseignement associé à des composants ADN, est de retourner aux notes des enquêteurs pour mieux les comprendre et les étudier jusque dans leurs moindres détails. Dans ces notes, on pourrait retrouver des informations supplémentaires, plus précises qui s'ajouteront à la structure du composant pour mieux comprendre l'événement comme tel et son contexte. En effet, on serait en mesure de s'attendre à faire de meilleurs rapprochements en intégrant davantage d'informations autour des dates, des distances, sur la nature des prélèvements ou toute autre information pertinente obtenue de témoins ou de caméras de surveillance, pour arriver à mieux cerner le ou les inconnus. Il s'agit là d'un exercice qui compte énormément sur la chance et la collaboration des intervenants associés à ces délits. En effet, il est tout à fait possible que les détails que l'on recherche n'aient pas été à la portée de l'enquêteur au moment de son intervention ou que sa curiosité ait été fixée sur d'autres détails que ce qui serait intéressant dans le contexte d'aujourd'hui. D'autant plus que, comme le montre la dynamique des composants, l'inconnu qui attire notre attention aujourd'hui n'était pas nécessairement intégré aux informations du moment, alors qu'aujourd'hui nous espérons retrouver des éléments intéressants pouvant s'y rapporter. De plus, les informations recherchées pourraient être absentes des dossiers de police pour des raisons de prescription ou de gestion interne[22] et il faudrait alors se contenter d'un minimum qui au pire sera insuffisant pour compléter le renseignement nécessaire à l'avancement de l'enquête. Aussi, les dossiers de police n'étant pas nécessairement centralisés ou numérisés, les recherches d'informations précises peuvent s'avérer laborieuses dans les nombreux cas où les délits ont été accomplis dans des secteurs administratifs différents.

[22] Une prescription est un délai au-delà duquel les poursuites ne sont plus possibles, et dans ces cas les dossiers sont souvent détruits, ce qui se produit aussi pour les délits résolus qui s'étendent au-delà d'une limite de temps fixée administrativement et qui varie selon la gravité des délits.

À titre d'exemple, nous allons approfondir les délits entourant les inconnus, les éléments du composant 87 (sections 4.3.1.3 et 4.3.2.1) et, plus précisément, la série des quatre vols qualifiés effectués conjointement par les individus 52230 et 20946 dont un est aussi en co-délinquance avec l'inconnu 52752 (Fig. 12), ce dernier ayant aussi été détecté par son ADN dans un délit d'agression sexuelle. Avant d'aller plus loin avec cet exemple, il convient de préciser qu'il a été choisi en premier lieu pour sa structure de plusieurs co-délinquances sur quatre vols qualifiés où, selon des données du LSJML, un inconnu, comme troisième individu, est présent dans l'un d'eux. Au fil de nos recherches avec la Sûreté du Québec, il est apparu que pour les services de police il n'y avait pas d'inconnu dans ce composant. En effet, au fil des étapes de l'enquête, les trois individus impliqués dans l'agression sexuelle « GGW7198 » avait tous été identifiés. Cette connaissance des faits enlève un peu de mystère à notre exemple, mais non pas l'intérêt que celui-ci présente pour illustrer le potentiel du recoupement des données ADN et policières pour la génération de renseignement. C'est pourquoi l'exemple sera traité en incluant l'individu 52752 comme s'il était toujours inconnu.

Notre collaboration avec la Sûreté du Québec et le LSJML a permis d'avoir accès à un peu plus d'informations sur certains dossiers dans cette série de co-délinquances, notamment pour le vol qualifié « LQX5830 » et l'agression sexuelle « GGW7198 » en ce qui a trait à la nature des prélèvements et la localisation précise des délits. La Figure 16 présente sur une carte des lieux et les objets saisis sur lesquels l'analyse d'ADN a été effectuée pour ces deux délits où l'inconnu a été mis en évidence. Les noms réels de lieux ont été changés pour garder l'information confidentielle. Voyons ces cas un par un.

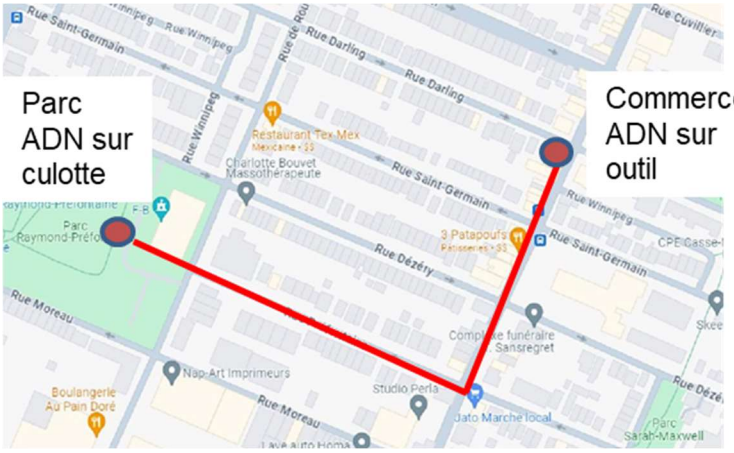


Figure 16 : Plan de la zone d'activités des deux délits (GGW7198 et LQX5830) incluant l'inconnu 52752 du composant 87. Le tracé en rouge représente les 650 m entre les deux endroits.

4.3.3.1 L'agression sexuelle commise au parc

Dans un parc, deux individus en agressent sexuellement un troisième; il serait aussi question d'extorsion, si bien que le délit est aussi classé comme vol qualifié. On comprendra ici que la situation présente une structure complexe, car on ne sait pas vraiment de qui partait, et vers qui était dirigé, le vol qualifié. Par l'enquête, on en arrive à identifier, du moins comme agresseurs, deux travailleurs étrangers qui auraient été en relation avec un troisième, la victime présumée, dont on a retrouvé le caleçon. Tous trois, en provenance d'Amérique latine, étaient des employés saisonniers et travaillaient dans la région. Peu de temps après, il n'a plus été possible de retracer les deux agresseurs, ces derniers étant retournés dans leur pays d'origine, et aucune trace ne permet de les revoir au pays depuis l'événement. Par contre, on ne sait rien de ce qu'est advenu de la présumée victime. C'est pourquoi le profil génétique masculin de la fraction épithéliale,[23] mis en évidence dans le caleçon, a été déposé en banque par le LSJML. Ce profil a conservé son statut inconnu sans concordance sur plusieurs années

[23] La fraction épithéliale d'une analyse correspond aux cellules de type épithélial qui proviennent des muqueuses. Les cellules qui se détachent de la muqueuse viennent se coller aux vêtements. Dans un cas d'agression sexuelle homme/femme, la fraction épithéliale d'une culotte ou d'un prélèvement vaginal donnera le profil génétique féminin de la victime et la fraction spermatozoïde, le profil masculin de l'agresseur. Dans notre cas, les cellules épithéliales correspondent plutôt au porteur du vêtement.

jusqu'au moment où le même profil génétique refait surface dans une concordance avec un délit de vol qualifié dans un commerce perpétré sept ans plus tard.

4.3.3.2 Le vol qualifié perpétré au commerce

Ainsi, sept ans plus tard, à 650 m de distance du parc, on retrouve, dans un délit de vol qualifié dans un commerce (LQX5830), le profil génétique de notre inconnu sur un outil. Avec près de sept ans d'écart entre les deux délits, on s'éloigne d'un contexte d'une série à enchaînement rapide de délits comme on en retrouve parfois. La serveuse, témoin de la scène, raconte qu'elle a vu deux individus : un sur les lieux qui procédait au délit et un autre qui attendait dehors. Ces deux individus identifiés dans le processus d'enquête sont «52230 » et « 20946 » (Fig. 12). De prime abord, il ne semble donc pas qu'un troisième individu soit sur place. Le délit semble avoir été perpétré entre individus du milieu criminel puisque deux des trois victimes dans le commerce sont aussi connues des policiers pour des vols qualifiés. Jusqu'ici, les premières informations reçues de la Sûreté du Québec ne permettent pas de faire la lumière sur le comment et le pourquoi de la présence de l'outil sur la scène, et sa saisie. C'est à cette étape de l'assemblage des informations qu'il faut retourner aux notes précises de l'enquêteur pour établir les motivations qui ont suggéré la collecte ou la saisie de cet outil. Ainsi, aujourd'hui il faudrait être en mesure de répondre à plusieurs questions pour structurer du renseignement informatif autour de ce profil génétique d'origine inconnue retrouvé sur l'outil : à savoir où était cet outil ou d'où vient-il ? Est-ce le suspect qui l'avait avec lui à son arrivée ou était-il déjà sur les lieux ? S'il était sur les lieux, est-ce que l'on peut savoir d'où il vient et depuis quand est-il dans le commerce ? On chercherait aussi des indices qui permettraient de savoir si cet outil est pertinent à la cause et quelle est la raison qui a motivé la saisie de cet objet par les services de police ? Il apparaît ici que l'absence de précision entourant l'outil nous démontre qu'il n'est pas assuré de trouver des éléments de réponses pertinents quand on reprend les dossiers d'enquêtes après de nombreuses années. C'est un peu ce qui montre notre exemple, et l'on reste avec des questions en tête au lieu de découvrir des indices pouvant mener à la production de renseignement.

4.3.4 Production de renseignement; un processus par étapes

On voit ici tout le chemin parcouru : au départ, deux délits séparés de près de sept ans impliquant un profil génétique inconnu, auxquelles deux individus connus s'ajoutent, eux-mêmes récidivistes et co-délinquants sur plusieurs autres délits de même nature. Quand on utilise une approche de montage des concordances en réseau pour visualiser les liens de co-délinquances, les relations entre ces individus, étalées dans le temps, deviennent évidentes, et dès lors, qu'on se penche plus à fond sur un délit comme celui qui a eu lieu dans le commerce, impliquant un outil associé à un profil inconnu, un questionnement naît de la nécessité d'éclaircir les raisons qui ont amené cet outil dans ce commerce, d'autant plus qu'il porte le profil génétique d'un individu dont on a perdu la trace sept ans plus tôt. Ici la production de renseignement n'est pas complète, mais son processus est entamé et déjà porte fruit. C'est alors que les questions qui se posent changent et de nouvelles hypothèses émergent, qui devront être vérifiées à la lumière des nouvelles informations qui pourront s'ajouter. L'utilisation des données structurées en réseaux permet aux enquêteurs de positionner un inconnu complice dans un dossier où les suspects n'en ont jamais fait mention alors qu'il semble encore être présent sept ans plus tard. C'est ainsi qu'à la suite d'une petite trace de renseignement, une multitude de questions peuvent être orientées de façon à faire avancer l'enquête. L'inconnu du parc sept ans auparavant est-il toujours présent au pays au moment du vol dans le commerce? Travaille-t-il alors comme sans papier dans ce commerce ? Ce qui expliquerait la présence de son profil à cet endroit sept ans après l'événement au parc.

Dans la série de VQ commune aux individus «52230 » et « 20946 », Il y a peu d'informations à trouver à propos du délit PO147. Aussi inscrit dans la région (1), ce VQ s'est effectué dans un commerce de restauration où les deux serveuses n'ont été en mesure de détecter que deux individus présents lors de l'événement. Donc, ici aussi aucun indice particulier ne peut indiquer que notre inconnu ait été présent. Pour ce qui est des deux autres VQ, nous n'avons pas pu retrouver d'informations supplémentaires, mais les liens entre les individus sont en soi déjà un début de production de renseignement.

4.3.5 Recherche de potentiel

À la lumière de l'exercice précédent sur nos trois exemples, il est maintenant possible d'effectuer un retour en arrière pour voir comment on pourrait déceler les individus inconnus ayant le meilleur potentiel d'être identifiés. Une première stratégie serait de cibler les inconnus en lien avec un maximum d'individus connus. C'est en ajoutant les informations policières sur ces derniers que l'on pourrait augmenter les chances de trouver des délits lors desquels les individus connus ont pu être complices d'un inconnu recherché. Cette première stratégie de recherche ouvre la porte à l'établissement de nouveaux liens avec les informations policières, mais la finalité d'obtenir des liens informatifs n'est absolument pas garantie puisqu'à cette étape l'analyste n'a aucun contrôle sur les types de délits en cause et les informations présentes dans les enquêtes. Toutefois, ce premier pas est un élément qui tend vers la recherche d'interrelations où il y aurait un maximum d'informations à soutirer des banques de données policières.

La Figure 17 montre deux composants fictifs A et B qui incluent respectivement 3 et 6 inconnus. Le positionnement de ces individus et les liens avec leurs complices connus ne les placent pas tous sur le même pied d'égalité. On y remarquera que le positionnement de l'individu est un élément clé qui n'est pas nécessairement en lien avec la taille du composant. En effet, tous les inconnus du composant A ont deux ou trois individus connus comme complices immédiats alors que dans le composant B, les inconnus 1, 5 et 6 ne sont en relation qu'avec un seul connu et le 2 n'est en lien qu'avec un inconnu. Ici, dans le composant B, l'inconnu 3 est en co-délinquance avec trois individus connus, comme les inconnus 2 et 3 du composant A. La proximité des interrelations est aussi à considérer, comme celle que l'on observe entre ces deux derniers inconnus qui ont en commun un complice connu en lien avec un délit d'homicide. Devant ces observations, la production de renseignement doit se concentrer sur les activités criminelles des connus, autour des inconnus 2 et 3 (composant A). Il en serait de même pour les connus qui entourent l'inconnu 3 du composant B.

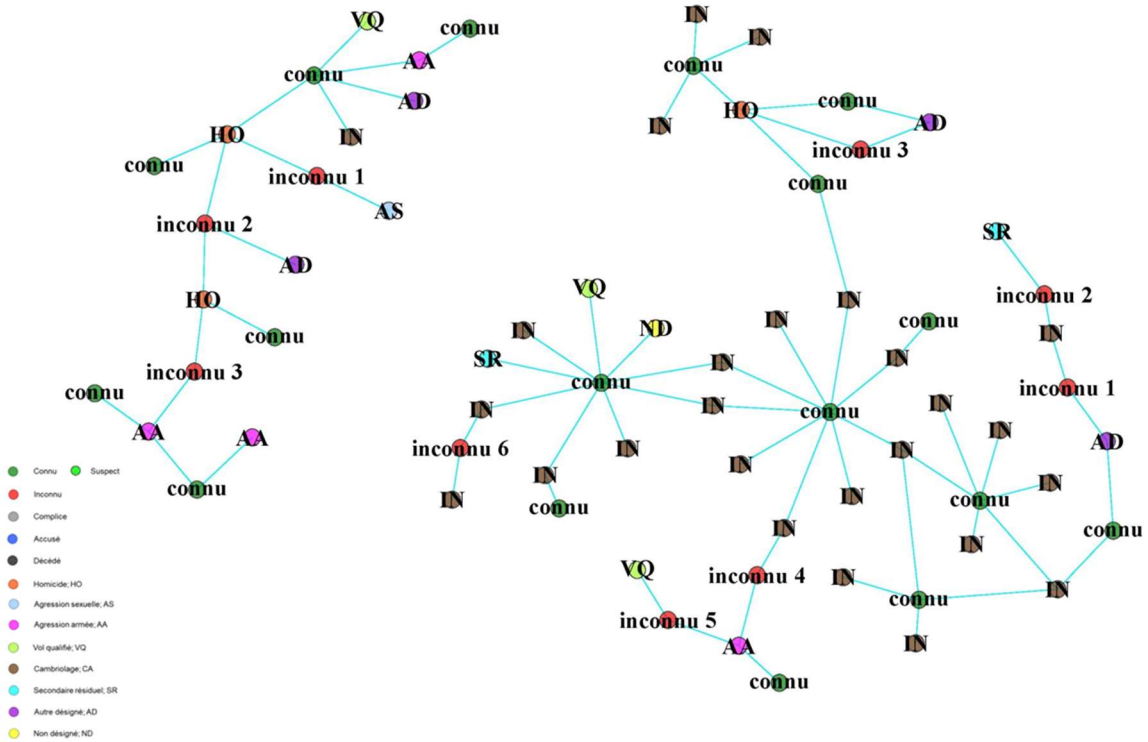


Figure 17 : Deux composants fictifs ayant de nombreux inconnus et un nombre de liens variable

4.4 Discussion

Ce chapitre explore une approche de production de renseignement en utilisant quatre exemples de composants avec peu d'individus ayant à leur actif des délits pouvant être associés à des carrières criminelles substantielles. En permettant d'ajouter un plus grand nombre de délits criminels et éventuellement de co-délinquants aux réseaux basés sur les concordances ADN (ou à l'inverse sur les données policières), cette approche ouvre la porte à plus de comparaisons et de recoupements. Cette production de renseignement utilisant des composants ADN ne doit toutefois pas faire oublier pour autant qu'il restera quand même beaucoup d'activités criminelles et de délinquants qui resteront inconnus des services de police. Mais l'approche proposée ici a l'avantage d'inclure des individus inconnus sur lesquels des efforts d'enquêtes doivent être déployés.

Inspirés par des travaux sur l'intégration des informations d'enquêtes (Rossy 2016), nous avons utilisé le schéma de réseau auquel nous avons greffé une dimension temporelle, inspirée du schéma de série. Cette visualisation, bien adaptée aux données ADN, fournit un portrait d'ensemble des liens entre des individus et leurs délits. Dans les schémas relationnels, des couleurs ont été ajoutées afin de distinguer les individus connus des inconnus ainsi que les huit types de délits présents dans nos données. À cela s'ajoute la couleur des liens qui renseigne l'observateur sur la source des données, soit par l'analyse de l'ADN au laboratoire ou les enquêtes des services de police. Ce codage par couleur rend plus efficace la visualisation des liens les plus pertinents autour des individus, qu'ils soient : des inconnus à identifier, de nouveaux perpétrateurs à interroger ou certains déjà connus qui prendraient de l'importance en fonction de la position qu'ils occupent dans la structure du composant dans lequel ils s'insèrent. Nous avons fait ces choix pour limiter les difficultés d'ordre technique ou sémantique qui peuvent découler de l'utilisation des schémas relationnels. En effet, des auteurs comme Schroder *et al.* (2007) discutent de la question de surcharge du schéma tandis que Peterson *et al.* (2000) soulève la question de la capacité de l'analyste-enquêteur à modéliser avec précision le problème à exposer. Même soutenue par une expérience solide, l'approche est sujette à interprétation ou à ambiguïtés, conséquence de la structure du schéma (Rossy et Ribaux 2012, Sparrow 1991). Dans leurs études utilisant des panels variés d'experts et d'étudiants, Rossy et Ribaux (2012) ont démontré la vaste étendue d'interprétations (et de compréhension) que l'on pouvait obtenir même à partir d'une présentation d'énoncés simples et structurés. Heureusement, il existe des recommandations générales mises en place par des instances judiciaires et académiques (interpol 1997, fedpol 2010, Rossy 2011) qui sont résumées dans Rossy (2016).

L'analyse des activités criminelles exposée au troisième chapitre, nous a permis de constater que les inconnus sont, en général, moins actifs que ceux qui sont identifiés par les services de police. Nous devrions donc nous attendre à ce que ces inconnus soient souvent associés à moins de délits et de co-délinquances en comparaison aux individus connus. Par conséquent, une production de renseignement qui n'utiliserait que des données de concordances ADN ne représentant qu'une bien petite partie des activités

criminelles ne saurait être suffisante, sauf si on les associe à des données complémentaires provenant des services de police. Ces dernières permettent d'ajouter beaucoup plus de délits criminels et d'individus connus, ce qui augmente les chances d'observer des caractéristiques propres aux divers événements, lesquelles sont collectées à l'enquête et retrouvées dans les notes des dossiers.

Dans notre exercice effectué à partir des données policières, deux étapes d'ajouts d'information –du type individus et dossiers– ont été réalisées : 1) des délits sans analyse d'ADN ont été ajoutés aux individus connus et 2) des individus connus des policiers ont été ajoutés aux délits ayant un lien ADN. Ce faisant, on obtient un tableau beaucoup plus complet des activités pouvant être associées aux individus présents dans le composant. Pour finaliser, nous avons aussi procédé à un ajout des complices connus des policiers pour chacun des délits. Ces ajouts peuvent permettre une intégration plus ou moins grande d'individus qui peut varier en fonction des types de délits. Comme le démontrent nos exemples, le composant ayant des individus associés à de nombreuses agressions sexuelles présente beaucoup moins de co-délinquance, tout comme une absence de complices, alors que c'est l'opposé pour les individus actifs en recel de stupéfiants et autres substances illégales. Dans cette situation, le foisonnement d'individus supplémentaires pourrait complexifier un composant jusqu'à compromettre la simplicité que l'on tente de conserver dans la production de renseignement aux fins de diffusion de l'information. De plus, le modèle est adaptable à une approche axée sur l'individu et l'ARS, qui sera entreprise si l'on cherche à répondre à des questions d'enquête qui relèvent des relations entre individus plutôt que de la structure des délits. De plus, à cette étape, le contexte de l'enquête et le positionnement du ou des inconnus pourrait dicter le volume des informations à recueillir en vue d'un développement du renseignement plus adapté.

Dans l'utilisation des concordances ADN, il est aussi très important de garder à l'esprit que le composant est une structure qui résulte d'une collection « historique » de co-délinquances qui s'étalent sur de longues périodes se chiffrant en années. L'utilisation d'une telle structure d'information temporelle intégrée au composant

possède l'avantage de faire ressortir les périodes d'activités les plus marquantes autour des inconnus et de leurs co-délinquants voisins. Toutefois, un individu ayant plusieurs liens avec de nombreux autres individus n'est pas obligatoirement encore en lien, disons de manière contemporaine, avec les co-délinquants de ces délits. Malgré cette limitation, il n'est pas exclu que de nombreux délits retrouvés dans des périodes relativement courtes et localisées dans un environnement proche puissent avoir été effectués par plus d'individus qu'il n'y paraît selon les résultats de l'analyse d'ADN ou selon les informations recueillies pendant l'enquête. Déjà, il ressort de nos exemples que deux composants présentent des situations de ce genre où l'individu inconnu pourrait avoir été présent dans une série de délits perpétrés par des co-délinquants connus de ce dernier. C'est ici que le renseignement peut efficacement s'intégrer à la structure du composant en ajoutant à celui-ci toutes les informations circonstancielles que l'on peut trouver dans les délits communs aux individus visés dans la recherche. Il convient de prendre note du nombre relativement élevé de composants qui, à l'analyse, offrent ce type de structure ayant potentiellement un inconnu relié à de nombreuses co-délinquances alors que seul, une ou deux avaient été vues par ADN. Dans nos exemples, 50% des quatre cas présentent ce potentiel. Il faut rappeler ici que ces cas ont été choisis uniquement en fonction de la nature de leurs délits et du petit nombre d'individus le constituant, ce dernier point étant nécessaire pour ne pas grossir inutilement les exemples. Ces quatre cas, choisis selon les critères susmentionnés, proviennent de l'ensemble des composants à l'étude et non pas d'un ensemble de composants, dont le potentiel de renseignement aurait été vérifié au préalable pour y déceler délibérément les meilleurs exemples. Au vu des résultats observés relativement aux composants retenus, il devient possible d'avancer l'hypothèse que les possibilités de production de renseignement sont réelles pour de très nombreux composants, d'autant plus qu'un des exemples montre qu'avec les liens ajoutés grâce aux données policières, il est possible d'assembler des composants ADN qui, autrement, apparaîtraient isolés. Ce seul fait milite en faveur d'une intégration à grande échelle de la production de renseignement, proposition qui est au centre de la présente dissertation. À la lumière des exemples utilisés, on pourrait dès lors se pencher davantage sur des composants qui contiendraient, possiblement, plusieurs inconnus et davantage de co-délinquances associées de près à ces inconnus,

comme ceux présentés dans la Figure 17. Il devrait en résulter un potentiel accru de développement du renseignement.

Toutefois, quoique le modèle tienne bien la route et ses promesses, des limitations importantes peuvent surgir dans l'élaboration d'un tel plan pour d'anciens délits entourant des inconnus. Dans une perspective d'une production de renseignement plus contemporaine aux événements, disons dans l'année qui suit les dates d'un groupe de délits, on pourrait espérer être en mesure de rassembler un bon nombre d'informations. En effet, les enquêtes seraient pour ainsi dire encore « ouvertes », et la mémoire des intervenants serait aussi plus accessible, et ce, même s'il y avait des lacunes dans les notes au dossier. À l'opposé, nos exemples démontrent que les anciens dossiers, après une période de latence, présentent une difficulté de collecte d'informations pour ce qui est de la production de renseignement. L'exemple du composant 87 nous monte une série de vols qualifiés qui remontent à moins de 10 ans. On serait porté à croire qu'un si faible recul n'est pas très grand, mais dans les faits il l'est. Les dossiers où des individus ont été identifiés ont suivi leur cours dans le processus judiciaire et sont classés. Les enquêteurs, s'ils n'ont pas changé d'affectation ou été promus, sont parfois déjà à la retraite. Les notes consignées dans ces « anciens » dossiers, si elles ne sont pas détaillées, présentent peu d'espoir pour tenter de donner une direction à la production de renseignement. Dans nos exemples, puisque la numérisation et la centralisation des dossiers d'enquêtes n'existent pas, il n'a pas été possible d'accéder aux notes concernant tous les délits intéressants, dans la mesure où elles n'existaient plus, qu'elles eurent été peu nombreuses ou pires, détruites.

Dans cette optique, nos exemples montrent le potentiel de production de renseignement et du même coup la nécessité de structurer un processus d'analyse dynamique qui suivrait les cas au jour le jour, sans grand décalage temporel, pour agir directement sur l'enquête afin de l'orienter le plus possible dans une collecte d'informations détaillées qui servirait à faire des recoupements efficaces. Ce genre d'approche permettant de concentrer les efforts d'enquêtes dans la période autour des délits observés, dans une région donnée, est reconnu pour son efficacité à identifier les

contrevenants (Johnson 2009). Encore de nos jours, la structure du travail policier n'est pas intégrée et centralisée pour ce qui est de l'ensemble de la gestion des délits. Comme il a été mentionné précédemment pour la gestion des concordances, beaucoup d'enquêtes fonctionnent selon une approche de gestion au cas par cas qui est le contraire de ce que l'on pourrait imaginer pour la mise en place d'une gestion globale et intégrée de toutes les données pertinentes aux enquêtes (Ribaux 2017). Les diverses administrations policières, les subdivisions territoriales sont des freins connus à l'échange et à la mise en commun des ressources et des données policières (Egger, 1984). Nous avons été à même de constater cette difficulté d'échange dans notre processus de recherche d'informations.

4.5 Conclusion

Au fil de nos travaux, il a été démontré que les composants ADN possèdent une valeur ajoutée en intégrant les individus inconnus, mais dans la plupart des cas, il n'y a que trop peu de co-délinquants autour d'eux pour ajouter suffisamment de renseignements. Par contre, la masse importante de connaissances des activités criminelles que l'on retrouve dans les données d'enquêtes policières a le pouvoir de changer la donne. Par cet ajout d'information individus et dossiers, il est démontré qu'il est non seulement possible, mais souhaitable de procéder à un assemblage de données policières et de concordances ADN pour créer un potentiel de renseignement. Ce potentiel réside dans l'étude de la dynamique des délits d'un réseau en portant attention aux périodes d'activités intenses où de nouveaux liens pourraient être créés grâce à l'ajout des individus connus autour des inconnus. N'utilisant qu'un choix de quatre petits composants, sans autre biais connu que d'y retrouver des délits souvent associés à des criminels de carrière, il a été observé que, dans la structure des composants obtenus par l'ajout d'informations policières, plusieurs situations permettaient de cibler les délits à étudier en profondeur, afin d'y chercher des informations circonstancielles pouvant être en lien avec les activités criminelles autour des inconnus. De plus, il a été démontré qu'un grand assemblage de données criminelles permet d'intégrer divers composants apparaissant à l'origine isolés dans les données de concordances ADN.

Les difficultés rencontrées dans les exemples, comme le peu ou l'absence d'informations circonstancielles, relèvent plus de l'étalement des délits dans le temps que de l'acte de production de renseignement lui-même. Comme telle, l'intégration de l'ensemble des informations de concordances ADN et policières pourrait permettre d'atteindre un meilleur taux d'identification des individus inconnus.

CHAPITRE V

DISCUSSION ET CONCLUSION

5. Discussion

5.1 Généralités (vision globale et perspective)

Le développement de cette thèse est né d'un désir d'améliorer le taux d'identification des individus inconnus détectés sur les scènes de crime par leur seul ADN. Le fichier des condamnés de la Banque nationale des données génétiques est très efficace pour fournir des identifications dans de très nombreux cas, et la seule véritable option qui reste pour les cas non résolus est de retourner à l'enquête sur le terrain. Mais dans de nombreux délits, l'enquêteur ne dispose que de très peu d'informations pour générer de nouvelles pistes. On pourrait citer en exemple les nombreux cambriolages et vols qualifiés, qui ne laissent souvent que très peu d'indices et, en l'absence d'identification génétique, l'enquêteur doit composer avec une faible marge de manœuvre, pour ajuster sa stratégie. Pour lui, le problème se décline en plusieurs questions. Quelles nouvelles informations faut-il chercher et où ? Comment les organiser pour qu'elles soient utiles à son enquête ?

Un des objectifs de cette thèse était d'explorer une approche différente pour générer du renseignement, qui fournirait aux enquêteurs un nouvel outil pour les aider à identifier l'individu qui n'est connu que par la trace d'ADN qu'il a laissée sur la scène de crime. La première étape de cette approche consiste à structurer en réseau les concordances ADN du fichier de criminalistique local de la banque de données génétiques. Cela est possible grâce aux nombreux délits effectués en co-délinquance, que l'on retrouve dans ce genre de données. Ce faisant, bon nombre d'individus inconnus ne sont plus considérés comme des acteurs isolés avec leurs délits individuels, mais comme étant intégrés à un réseau de co-délinquance (Jeuniaux *et al.* 2016). Déjà ces "réseaux sociaux ADN" se prêtent à l'analyse quantitative (SNDNA), faisant ainsi émerger de nouvelles connaissances de nature criminologique (chapitres II et III).

On peut pousser l'analyse plus loin en ajoutant des données policières à ces réseaux reconstruits à partir de données criminalistiques (chapitre IV), en vue de la production de renseignement forensique. Ce concept est mis de l'avant par les chercheurs de l'université de Lausanne (UNIL), dont les travaux offrent de nombreux exemples de renseignement généré à partir de traces de pas laissées par des cambrioleurs (Rossy *et al.* 2013), de la composition chimique de drogues illicites (Esseiva *et al.*2007; Morelato *et al.* 2013), de faux documents (Morelato *et al.*2014; Baechler *et al.* 2015), de photographies (Milliet *et al.* 2014), d'analyses d'accélérants de scènes d'incendies criminels (Bruenisholz *et al.* 2019A) et même de profilage de montres de contrefaçon (Hochholdinger *et al.* 2019). Outre ces études de nature empirique, d'autres publiées par les chercheurs de l'UNIL abordent les aspects conceptuels de la production de renseignement à partir de données forensiques (Ribaux et Margot 1999; Ribaux *et al.* 2006; Ribaux et Wright 2014; Ribaux et Caneppele 2017). Par conséquent, l'approche réseau intègre dès le départ les scènes de crime dans le contexte large des interrelations entre les agents et facteurs de la criminalité sur un territoire et dans leur temporalité (p. ex. co-délinquance et récidive) (Ribaux *et al.* 2010A, 2010B; Delémont *et al.* 2017). Comme l'expliquent Rossy et Ribaux (2014) et Rossy (2016), l'aspect visuel de la présentation des résultats s'avère très important pour le partage des connaissances et la qualité des communications, dans un but d'amélioration des méthodes d'enquête. En reliant des délits très séparés dans le temps, l'ADN peut y contribuer (Rossy *et al.* 2013).

L'utilisation des concordances ADN incluant des individus inconnus a aussi été faite par De Moor *et al.* (2017), à la suite des travaux pionniers en Belgique. En effet, les données de concordance de la banque de données génétiques belge ont été mises à profit pour démontrer l'intérêt et la faisabilité de développer le montage en réseau des données de concordances ADN, une approche des plus prometteuses puisqu'elle valorise les informations concernant les individus connus que par leurs traces ADN (Jeuniaux *et al.* 2015, 2016, 2017). En plus du montage des données en réseau, Jeuniaux *et al.* (2016) apportent déjà un volet plus criminologique en intégrant une analyse des délits par types et par régions.

Ce concept de renseignement forensique (*forensic intelligence)*, c.-à-d. la production du renseignement utilisant, de manière plus systématique, les données d'un laboratoire de sciences judiciaires, et en particulier celles génétiques, a incontestablement inspiré de nombreux travaux de recherche, dont cette thèse qui impliquait de structurer les concordances ADN en réseau pour en arriver à créer une nouvelle approche de production de renseignement en soutien aux enquêtes policières.

Rappelons en dernier lieu que d'autres études ont aussi utilisé des données ADN en portant une réelle attention à la présence des inconnus et à certains aspects de leur comportement. On cherchait alors surtout à quantifier la période de temps dans laquelle ils restaient inconnus ou si l'étalement de leurs délits sur un territoire donné avait un effet sur leur probabilité d'arrestation (Lammers *et al.* 2012; Lammers et Bernasco 2013; Lammers 2014; Bernasco *et al.* 2016). Ces études à visées surtout criminologiques n'abordent toutefois pas l'analyse des données en réseau ou la production de renseignement comme nous le faisons ici.

5.2 Précisions et développement

Notre étude se démarque par son ampleur. À ce jour, elle est la seule qui couvre une si grande période, dépassant les douze années de l'étude belge (Jeuniaux *et al.* 2016), les chapitres II et III portant respectivement sur 18 et 19 années de concordances ADN. Par conséquent, le potentiel de retrouver des composants dans lesquels on décèle un plus grand nombre de co-délinquances et de délits augmente. Par contre, la catégorisation des délits, dans les données du LSJML utilisées ici, est moins précise (huit types) que celle qu'on retrouve dans les données belges (12 types). Ce point est particulièrement important pour qui voudrait se servir des données de la criminalistique pour étudier certaines questions en criminologie (Bright *et al.* 2022). En effet, des études portant sur la nature et l'évolution des activités criminelles ne pourraient pas nécessairement se contenter d'une classification trop sommaire des types de délits (De

Moor *et al.* 2018). Ce point sera repris un peu plus loin, à la lecture de nos résultats concernant les délits associés aux individus connus et inconnus.

Dans cette thèse, trois aspects de l'analyse des concordances ADN en réseau et des inconnus qui s'y trouvent vont être développés. Premièrement, il fallait démontrer que le positionnement des inconnus dans les réseaux de co-délinquance était tel que nous pouvions potentiellement en tirer du renseignement. En effet, dans le cas où les inconnus seraient tous en périphérie des composants du réseau social criminel, et possiblement en lien avec peu de complices, il deviendrait beaucoup plus difficile de procéder à une production de renseignement pour appuyer une enquête visant une identification.

Deuxièmement, et malgré la limitation précitée quant à la catégorisation des délits en seulement huit types, il nous a été possible d'explorer les différences entre les activités criminelles attribuables aux individus connus et inconnus. Cet aspect de notre analyse fournit un bel exemple de complémentarité entre deux domaines savants, la criminalistique et la criminologie, comme le montre d'ailleurs la collaboration essentielle du Centre international de criminologie comparée (CICC) à ces travaux.

Enfin, dans le chapitre IV sont mis de l'avant des exemples de production de renseignement à partir du jumelage des données forensiques (ADN) et policières, démontrant les possibilités tactiques et logistiques qu'offre cette approche en réseau "intégrée et dynamique", notamment pour aider à cerner les inconnus et leurs complices. L'approche fait appel à un schéma relationnel où la collaboration ainsi que la qualité des informations recueillies et jumelées déterminent les possibilités et limitations. Voyons ces divers aspects plus en détail.

5.3 Les individus en réseau

5.3.1 La distribution des inconnus

Avant de nous concentrer sur le positionnement des individus à l'intérieur d'un composant, il est intéressant d'examiner la distribution des inconnus dans les composants de diverses tailles, question de voir si ceux-ci y sont présents de manière plutôt uniforme. Rappelons ici que les plus grands composants contiennent davantage de liens entre les individus (co-délinquances) pour la production de renseignement.

La distribution des individus inconnus dans un ensemble de données complexes, structurées en réseau, apporte un niveau de difficulté d'analyse statistique qui a été pris en compte en procédant par des permutations aléatoires. De plus, l'ensemble des composants a été subdivisé en trois groupes distincts en fonction du nombre d'individus dans ces composants, soit les petits composants (un ou deux individus), les moyens (trois ou quatre) et les grands (cinq et plus). Ce faisant, nous avons démontré que pour ces trois groupes de composants, la distribution des individus n'était pas aléatoire, et ce, plus fortement pour les petits et moyens composants. Les valeurs réelles du nombre d'inconnus dans les petits composants sont en deçà de celles attendues par hasard, et la situation est inversée pour les composants moyens (Fig. S2, chapitre II). Cette observation s'explique peut-être par un attrait particulier des individus destinés à rester inconnus, pour la co-délinquance en compagnie de peu de complices. Le chapitre III fournit plus d'informations sur cet aspect des délits solos et ceux en co-délinquane.

Toutefois, pour ce qui est des grands composants, la valeur observée du nombre d'inconnus, qui s'élève à 84, est équivalente à la valeur de 85 attendue par hasard. Cependant, si l'on voulait appliquer la proportion globale de 21% d'inconnus dans les données, la valeur obtenue pour les inconnus, à partir du nombre d'individus présents dans les grands composants, serait 32 % plus basse. Ce faisant, on pourrait expliquer cette observation en proposant l'hypothèse qu'un individu puisse plus difficilement passer inaperçu lorsqu'il est présent dans un plus grand nombre de délits commis avec de nombreux complices (Lammers *et al.* 2012) ou qu'on observe peut-être un plus haut

taux de dénonciation dans les grands composants. Toutefois, nos données ne permettent pas de soutenir cette hypothèse qui nécessiterait une approche différente, et des données supplémentaires, ce qui pourrait être abordé dans une étude future.

La structure complexe des données rend la distribution réelle des inconnus dans les composants difficile à saisir, comme le montrent les résultats présentés à la Figure S2 (chapitre II) : de très nombreux individus agissent en solitaire et parfois sans récidiver, entraînant un biais dans les données de concordances ADN.

Puisque la distribution des inconnus n'était pas un point critique pour structurer la production de renseignement, nous n'avons pas poussé plus à fond la compréhension de ce qui pourrait influencer cette distribution générale des inconnus dans les composants. Le point à retenir concerne la complexité des données de concordances ADN qui se rapportent à des récidivistes, des individus solitaires, des co-délinquants; identifiés ou non. Cette complexité résulte de l'assemblage des concordances obtenues par comparaison des profils génétiques provenant des scènes de crime auxquelles s'ajoutent les résultats d'identification obtenus de la BNDG (fichier des condamnés). Il en résulte que cet amalgame de sources diverses présenterait des biais si les données étaient toutes comparées sans distinction.

5.3.2 Le positionnement des individus inconnus dans les composants

Au moment de finaliser ce texte, les concordances ADN en un format réseau avaient été utilisées dans très peu de travaux pour étudier le positionnement des individus inconnus. Outre les études belges de Jeuniaux *et al.* (2015, 2016) mentionnées précédemment, une autre étude belge s'est intéressée à cette approche d'analyse des données ADN en réseau. Cette étude de De Moor *et al*. (2020) compare des résultats de paramètres d'analyse des réseaux sociaux (ARS) obtenus des données ADN seules, avec ceux obtenues en agrégeant les données ADN à des données policières et finalement en les comparants à une distribution aléatoire. C'est une approche qui a le mérite d'être complète et très structurée. Toutefois, les conclusions obtenues doivent être bien

interprétées en fonction du fait que les moyennes obtenues le sont pour l'ensemble des données. Voyons les conséquences de ce choix, et par la suite il sera question de l'agrégation des données policières.

À propos de l'utilisation globale des données, il faut garder à l'esprit que de nombreuses variables associées à des activités criminelles ne sont presque jamais observées selon une distribution dite normale. À ce titre, les données de De Moor *et al.* (2020), tout comme celles provenant du LSJML, ou celles que l'on peut rencontrer dans de nombreuses publications (Gründ et Morselli 2017; Jeuniaux *et al.* 2016; Morselli et Boivin 2016) sont toutes distribuées de manière exponentielle, comme le démontre notre Fig. 3 du chapitre II. Pour comparer des résultats de centralité, pour ne citer que cet exemple de paramètre ARS, il n'est pas des plus adéquat d'utiliser une moyenne arithmétique calculée à partir d'un ensemble de données, où une vaste portion de celles-ci ne peut fournir qu'une centralité de zéro. Cette portion des données correspond aux individus isolés et sont à hauteur de plus de 40 % chez De Moor *et al.* (2020) (Cette valeur est de 88 % dans nos données). De plus, les individus solitaires ne sont pas les seuls qui ont une centralité nulle; il en est de même pour les individus en duo, le deuxième groupe d'individus en importance (8,5 % dans nos données). Ce n'est qu'à partir des petits réseaux composés de trois individus que, dans certaines situations, un seul des trois individus possède une valeur de centralité. De la même façon, dans les composants intégrant quatre individus, ces derniers présenteront davantage de valeurs de centralité non nulle. Ainsi notre choix final pour calculer des moyennes de paramètres d'ARS s'est fixé sur les composants de cinq individus et plus, et ce, bien avant la publication de De Moor *et al.* (2020) qui recommandait d'ailleurs, pour des analyses futures, de n'utiliser que des composants de plus grande taille.

La seconde approche mise en pratique par De Moor *et al.* (2020), qui peut aussi apporter des difficultés, est l'agrégation des données ADN et policières. Cette approche a certes l'avantage de recourir à un ensemble de données plus complet et proche de la réalité, où le volume important des données totales pourrait permettre d'observer plus d'individus inconnus ou autres dans des situations de centralités importantes. Toutefois,

malgré cette agrégation de données plus complète, les inconnus n'ont pas été caractérisés comme étant aussi centraux que les individus connus et on conclut donc qu'ils sont plus périphériques (De Moor *et al.* 2020). Comme précédemment mentionné au chapitre II, ce résultat s'expliquerait par le biais de l'utilisation des moyennes obtenues des données globales. L'agrégation des données qui se veut un rapprochement au plus près de la réalité ne corrige en rien l'effet des nombreux isolats sur les moyennes des paramètres d'ARS. Au pire, c'est une approche complexe qui nécessite beaucoup d'interventions et de vérifications puisqu'il faut absolument éviter de dédoubler les individus qui, s'ils sont connus dans les données ADN, le sont obligatoirement dans les données policières. Pour les novices, il ne semblerait pas y avoir là sujet à problème, mais il existe de nombreux écueils qu'il faut contrer quand on s'attaque à un assemblage de données provenant de deux sources, et structurées différemment. Dans le cas présent, les nombreux et différents « alias » que l'on retrouve, autant dans une banque de données que dans l'autre, ainsi que d'autres types d'erreurs de saisies de données peuvent induire un important lot de faux positifs ou négatifs dans les liens observés entre individus, et ce, sans compter que la somme de travail pour vérifier l'ensemble des données peut-être phénoménale. Dans la présente étude, il a été décidé d'opter pour une simplification de la gestion des données en mimant les données policières à partir des données ADN. Par cette approche, les données ADN sont considérées comme complètes, alors que, sans les inconnus, elles représentent les données policières. Du coup, cette simplification évacue les problèmes d'agrégation de données. Par la simple soustraction des individus inconnus, les individus connus qui restent représentent, quoique fort réduite par rapport à la réalité, une portion véritable des données policières. Le but étant de mieux comprendre où et comment les inconnus s'insèrent dans les structures en réseau. Cette utilisation ciblée des données permet une analyse plus fine alors qu'un groupe massif de données ne contribuerait qu'à masquer ce que l'on cherche à détecter.

Ainsi, les valeurs moyennes obtenues de centralités sont représentatives de ce que l'on peut observer dans des composants qui présentent de nombreux liens entre des individus, et le calcul n'est pas faussé par des problèmes externes provenant

d'agrégation complexe de données ou de la présence d'un grand nombre de valeurs nulles. Le groupe de cinquante et un composants étudiés a permis de démontrer que les inconnus sont bien intégrés dans la structure de co-délinquance, et ce, même s'ils sont parfois situés en périphérie. En effet, 19 % de la centralité d'intermédiarité observée chez les individus connus est conséquente de la présence des inconnus qui sont présents dans l'ensemble dans une proportion de 21 %. C'est dire que ces derniers soutiennent l'intermédiarité presque au niveau de leur présence. Aussi, lorsque les inconnus sont retirés, l'augmentation de 46 % de la proportion des valeurs de coefficient d'agglomération (clustering) de zéro chez les connus démontre bien que les inconnus sont très bien intégrés dans les composants. Les individus connus auraient conservé une proportion bien plus grande de valeur positive s'ils n'étaient pas interreliés aux inconnus. En d'autres mots, l'absence des inconnus diminue de beaucoup le coefficient d'agglomération; c'est donc dire que ces derniers participent à la cohésion de nombreux liens de co-délinquances et ce, qu'ils soient centraux ou non. En lien avec cette dernière valeur, la densité des relations avoisinant un individu (egonet density) est similaire, que ce dernier soit connu ou non. De plus, en compilant les distributions de trois paramètres d'ARS, il est démontré que plus la valeur de ces paramètres est élevée pour les inconnus plus ces derniers ont une influence marquante sur la diminution des moyennes observées de ces mêmes paramètres, pour les individus connus. Certains exemples de composants montrent aussi que dans une approche d'enquête ciblée l'absence des inconnus peut avoir un effet majeur sur l'interprétation des liens observables entre les individus, allant jusqu'à diviser le composant d'origine en deux plus petits composants (Chap. 2, Fig. 2A, 2B, 6).

Il est maintenant plus facile de comprendre les différences entre les travaux de De Moor *et al.* (2020) et la présente dissertation. Dans les deux cas, il y a un biais important dans les données utilisées. D'une part, les variations de valeurs des paramètres d'ARS, des inconnus, sont bien réelles et peuvent être évaluées si on se concentre sur les composants qui sont adéquats pour soutenir de telles évaluations. D'autre part, cette réalité est aussi bien présente dans les données de Belgique (De Moor *et al.* 2018; 2020), mais noyée dans une masse de données qui pèse lourd sur l'ARS. Dans notre approche,

en omettant les données sans valeurs d'ARS, on s'éloigne de l'aspect général des données, mais cela permet de faire ressortir ce qu'il y a de structurant autour des inconnus dans des composants plus complexes.

5.3.3 Mise en perspective

Les données de concordances ADN d'un laboratoire de sciences judiciaires ne représentent que la pointe de l'iceberg des délits que l'on peut attribuer aux individus criminalisés connus. Il s'agit là d'une limitation importante qui sera reprise plus loin. Toutefois, il faut garder à l'avant-plan ce que ces données apportent de substantiel en la présence des inconnus. Il s'agit là de l'élément le plus important à considérer dans l'utilisation des données de concordances. Et puisque ces individus sont aussi bien intégrés dans leurs réseaux de co-délinquances que ceux connus des services de police, il serait adéquat de poser l'hypothèse que l'intégration des données policières permettra d'ajouter davantage de liens autour des inconnus. Ce qui pourrait faire une énorme différence dans l'approche stratégique d'une enquête. En effet, au travers des exemples décelés dans les divers composants étudiés, il est évident que plusieurs inconnus sont positionnés avec suffisamment de centralité pour soutenir une réorientation dans l'évolution de l'enquête. C'est ainsi que l'ajout d'informations policières pourrait aider à identifier de nouveaux témoins à questionner ou à pointer ceux qui, déjà connus, pourraient être remis à contribution à la lueur de nouvelles informations mises au jour par la structure en réseau des co-délinquances. Aussi, la mise en commun d'informations imposerait une collaboration plus serrée entre diverses régions administratives quand les délits sont répartis sur de grands territoires et ferait tomber les barrières de l'analyse au cas par cas. Finalement, l'insertion de données policières permet un assemblage plus intégré des petits composants qui apparaissent isolés lorsqu'ils ne sont vus que par l'analyse d'ADN. Dans ce contexte bien précis, l'ajout d'une information policière pourrait «relier» un petit composant dans lequel on retrouve un inconnu et très peu de co-délinquance, à un autre composant beaucoup plus informatif et interreliant des individus connus.

5.3.4 Comprendre et intégrer la dynamique des composants

L'utilisation des données ADN en réseaux ouvre la voie à des études tenant compte de la structure dynamique de ces derniers. En effet, les composants ADN sont le résultat d'une collection de délits en co-délinquance « bâtis » au fil du temps, parfois sur plus de dix ou douze ans comme dans certains de nos exemples. Pour vérifier si la complexité de la structure des composants avait une quelconque influence sur les valeurs d'ARS observées, une étude de sensibilité a été mise de l'avant.

Profitant de l'avantage d'un éventail de données qui s'étalait sur 18 ans, nous avons choisi d'utiliser les paramètres de réseaux établis pour l'ensemble des données et de les comparer aux données scindées en deux groupes de sept ans pour les quatorze dernières années (2005 à 2011, 2012 à 2018). Cette approche a permis de démontrer que le positionnement des inconnus ne subit pas de changement important au cours des diverses périodes. Évidemment, la dynamique d'un composant est entièrement dépendante des comportements des individus qui le composent, selon leurs choix de vie et/ou des événements qui se présentent à eux. Sans aucune information vraiment pertinente pour comprendre la dynamique réelle d'un composant, notre choix de structurer notre analyse sur deux périodes précises a été réfléchi en fonction de l'évolution de la banque. En choisissant deux périodes assez longues et en éliminant les premières années, il en résultait l'obtention de deux périodes plus homogènes qui permettraient à de nombreuses co-délinquances de se produire. Si deux périodes de neuf ans avaient été choisies, les quatre premières années de la banque à plus faible volume auraient induit un biais entre le groupe le plus ancien et le plus récent. Évidemment, une connaissance approfondie de l'évolution d'un composant, qui serait obtenue en utilisant des données plus complètes et personnalisées (en considérant les individus impliqués) serait des plus intéressantes pour des études futures sur la dynamique de ces structures sociales.

Au final, le développement du premier grand volet de cette dissertation montre que les inconnus sont bien intégrés dans les composants et que leur présence fournit la base

structurante nécessaire à la production du renseignement. Par contre, il devient aussi évident que les données de concordances ADN doivent être jumelées aux données policières afin d'intégrer les très nombreux liens de co-délinquances supplémentaires, qui n'ont pas fait l'objet d'une analyse génétique. Ces nouveaux liens de co-délinquances vont permettre l'ajout d'informations, issues des enquêtes, qui sont absentes des données ADN. C'est à cette étape de notre étude, présentée au chapitre IV, que le tableau général prend forme.

Avant d'en arriver là, il fallait tirer profit d'avoir accès aux données relatives aux activités criminelles des individus inconnus des services de police pour explorer cette avenue qui reste inaccessible aux connaissances criminologiques du fait que les données proviennent habituellement des dossiers de police ou judiciaires dans lesquels les délinquants sont connus des services de l'ordre.

5.4 Divers aspects de l'activité criminelle des inconnus

5.4.1 Généralité

La comparaison des activités criminelles entre des individus inconnus et connus restait à faire puisque les études antérieures utilisant des données ADN ont plutôt analysé, l'étendue de la période d'activité dans laquelle les individus étaient inconnus, les déplacements de ces derniers et une évaluation générale du niveau de violence de leurs délits (Lammers *et al.*2012, 2014). Jeuniaux *et al.* (2016) ont utilisé des informations relatives aux types de délits commis par les individus en réseau, mais en s'intéressant à la distribution géographique des délits plutôt qu'aux différences entre les individus connus et inconnus.

Pour aborder l'étude des activités criminelles des individus inconnus, notre approche s'est construite autour de l'utilisation de plusieurs paramètres permettant de caractériser et, pour finir, de comparer l'activité criminelle de ces derniers à celle des individus connus. Les types et la gravité des délits, la diversification des activités criminelles chez un individu et l'étalement de celle-ci dans le temps, sont les principaux

paramètres qui ont été utilisés pour les deux groupes d'individus en tenant compte de leurs dynamiques à savoir s'ils œuvraient en solitaires ou en co-délinquances. Puisqu'à peu près rien n'a été fait à ce jour sur l'activité criminelle d'individus inconnus, nous avons choisi de privilégier une approche usant de multiples paramètres afin d'augmenter les chances de mettre en évidence des observations pertinentes. Il s'agit donc d'un travail exploratoire et descriptif, mais ayant suffisamment de données pour faire état de la valeur probante des découvertes. L'ensemble des paramètres analysés ont permis d'évaluer que les activités criminelles favorisées par les inconnus différaient significativement de celles pratiquées par les individus connus et qu'elles étaient en accord avec les théories connues de l'exposition et de la compétence (Blumstein *et al.* 2010; Clare 2011; Ouellet et Bouchard 2017).

5.4.2 Observations relatives aux comportements solitaires ou de co-délinquances

Pour ce volet, il fallait classifier les individus connus et inconnus en fonction de leur comportement solitaire ou de co-délinquance, à dossier unique ou multiple, sans égard aux types de délits. Ces trois subdivisions « comportementales » sont mutuellement exclusives. Un individu agissant en solitaire n'est pas intégré à un quelconque composant. S'il l'était, il serait alors dans l'une ou l'autre des deux catégories de co-délinquance qui sont fonction du nombre de délits observé pour l'individu. On comprendra que l'individu ayant un seul délit en co-délinquance se situe en périphérie de la structure d'un composant avec un lien le rattachant à ce dernier, ou avec plusieurs liens si l'individu est actif avec de nombreux complices. Dans l'autre regroupement de co-délinquance, les individus sont récidivistes et parfois en position de co-délinquance multiple. La classification tient compte de ces trois aspects sans prendre en compte le nombre précis de délits en solitaires que les individus en co-délinquances multiples ont peut-être (et souvent) à leur actif.

Cette classification de la distribution des individus permet déjà de constater que les individus inconnus ont des comportements différents de ceux des connus. Premièrement, ils sont très portés sur les activités criminelles en solitaire, plus des deux tiers d'entre

eux étant des récidivistes solos. Sous l'aspect de la co-délinquance, on observe une nette préférence pour les activités multiples. Les valeurs observées pour les connus sont plus également distribuées, montrant une plus faible proportion d'individus pour les activités criminelles solitaires et plus de similitudes entre les deux catégories de co-délinquance.

Dans notre analyse, les conclusions élaborées pour les inconnus tiennent principalement compte des activités solitaires de ceux-ci et, dans une moindre mesure, des activités en co-délinquances multiples. Les données liées aux activités de co-délinquance unique ont sciemment été mises de côté puisqu'elles souffrent d'un biais imposé par la détection des inconnus qui doivent, eux, présenter un minimum de deux délits pour être détectés, ce qui n'est pas le cas pour les connus. (Lavergne *et al.* 2022). Par conséquent, la catégorie de co-délinquance unique est surreprésentée chez les individus connus et sous-représentée chez les inconnus. En effet, ces derniers sont uniquement observés lorsque leur co-délinquance unique est reliée à un autre délit qui permet la détection de la concordance[24].

Tous les individus inconnus présents dans la banque de données sous un seul exemplaire, qu'ils soient en co-délinquance unique ou solitaire, sont obligatoirement absents du tableau. Cette absence des données des inconnus en co-délinquance unique fausse partiellement la comparaison avec les connus de la même catégorie qu'il fallait conserver. En effet, il aurait été malvenu de mettre de côté ce grand volume d'activités criminelles de co-délinquance des connus alors que notre étude porte sur ce sujet. C'est pourquoi, compte tenu de la complexité des données, il était préférable de ne pas commenter sur les différences entourant les co-délinquances uniques. D'autres études pourraient reprendre ce volet des données et raffiner l'approche. Toutefois, étant donné que l'utilisation du fichier des condamnés est très efficace pour fournir des identifications[25], on est en mesure de s'attendre à ce que les individus, à dossier unique, qui gardent leur statut d'inconnus, ne soient pas aussi nombreux que les connus. Il

[24] Voir section 1.1.4 la fiche de concordance et les réseaux.
[25] Le volume des données sur les connus à dossier unique en est un exemple flagrant. Voir chap. II, Tab. 1, les connus uniques (n= 6645)

apparaît donc que le nombre des 25 inconnus en co-délinquances uniques est effectivement sous-estimé, mais il serait inapproprié d'imaginer qu'il puisse s'approcher de celui des connus de cette catégorie, qui avoisine les 1900.

En ce qui concerne les inconnus agissant en solitaire, ici aussi les données souffrent de l'absence des inconnus ayant à leur actif un seul dossier. Mais la connaissance de leur nombre ne ferait qu'augmenter la proportion déjà importante observée dans cette catégorie. Il en résulterait que les inconnus demeurent les champions de l'activité criminelle en solitaire et que notre conclusion, uniquement basée sur les récidivistes agissant en solos, n'est que légèrement sous-estimée.

Comme pour de nombreux aspects relatifs à l'analyse de ces données complexes, la valeur probante des observations demeure difficile à obtenir. L'utilisation d'un modèle statistique plus conventionnel comme le chi^2 n'est pas applicable puisqu'il présuppose toujours que les données sont indépendantes, ce qui est tout sauf vrai pour les données de concordances ADN. En effet, ces données sont la résultante de nombreux choix effectués par des individus en fonction de leur expérience de vie (Blomberg *et al.* 2012; Charrette et Papachristos 2017; Menting *et al.* 2016; Pyroz *et al.* 2017), des succès et erreurs du passé (Clare 2011), des opportunités qui se présentent à eux (Mc Gloin et Piquero 2010; Bright *et al.* 2022) et de l'inter influence que peuvent avoir sur eux les divers autres délinquants qui les entourent (Haynie 2001, 2002; Gründ et Morselli 2017). De plus, les interrelations et les divers types de délits, propre aux spécialités et talents de chacun, sont à la base structurante des composants observés dans les données (Gründ et Morselli 2017). Il est possible d'affronter cette complexité, en utilisant des permutations aléatoires du statut connu et inconnu des individus dans la structure intacte des données. Cette approche a permis de démontrer que les inconnus présentaient un comportement différent de celui des individus connus.

Pour finir, il faut préciser que le biais associé aux données de co-délinquances uniques n'a toutefois pas d'effet sur la catégorie des co-délinquances multiples, car ici,en raison des récidives, toutes les données pour les deux types d'individus sont

présentes. De plus, soutenue par l'analyse aléatoire, l'observation de la prépondérance des inconnus dans les co-délinquances à deux individus plutôt que dans les grands composants est tout à fait pertinente. Par leur présence accrue dans ce sous-groupe, les inconnus ont l'avantage supplémentaire de demeurer le plus discret possible, comme ils le sont dans leurs activités solitaires. Toutefois, il faut garder à l'esprit que le biais policier est toujours présent dans ce type de données. Plus spécifiquement, ce biais indique qu'il faut naturellement s'attendre à ce qu'il y ait systématiquement moins d'individus connus, et par opposition, plus d'inconnus parmi les délits secondaires, et l'inverse dans le groupe des délits violents, ces derniers délits attirant davantage l'attention des policiers.

5.4.3 Diverses perspectives sur les activités criminelles

L'ensemble des paramètres utilisés pour qualifier, quantifier et comparer les activités criminelles entre les individus connus et inconnus ont permis de démontrer des différences notables entre les deux groupes d'individus. Globalement, les inconnus sont plus spécialisés lorsqu'ils sont plus actifs, se concentrent davantage sur les crimes secondaires que sur les plus violents et étalent leurs délits sur de plus longues périodes, comme le démontre la densité de leurs délits. Les inconnus ne sont toutefois pas exempts de sensibilité humaine, si l'on peut dire, car pour eux aussi, à l'instar des individus connus, la violence de leurs délits présente une corrélation positive avec le nombre de complices présents à l'événement. Comme quoi l'influence des pairs demeure un élément universel chez l'humain (Haynie 2001, 2002). Un autre aspect observable chez l'humain en situation de délinquance est cette propension à essayer de passer inaperçu, qui semble une tendance marquée pour les individus inconnus qui tentent de garder un profil bas, ce qui les aide à éviter la pression des enquêtes policières plus soutenues que l'on associe aux délits plus violents. Finalement, la démonstration intéressante et innovatrice qui ressort de ces analyses est que le comportement des individus qui évitent d'être identifiés supporte le modèle de la théorie d'exposition (délits de moindre violence et étalés) et de compétence (délits plus spécialisés)(Blumstein *et al.* 2010; Ouellet et Bouchard 2017).

Nos analyses ont permis de mettre au jour des observations inédites en phase avec les théories que proposent les connaissances générales de la criminologie, mais certains aspects restent plus obscurs, ou du moins nécessiteraient plus d'approfondissement. À titre d'exemple, dans nos résultats relatifs aux inconnus, l'indice de diversification propose une spécialisation qui n'est pas en corrélation positive avec la densité des composants, comme on peut l'observer chez nos individus connus et comme présenté dans McGloin et Piquero (2010). Par contre, on constate que les inconnus sont plus spécialisés (Di=0) lorsqu'ils sont plus actifs en co-délinquance, donc dans une structure sociale où l'on serait en mesure de s'attendre à trouver parfois des composants de plus grande densité. Ce type de corrélation semble toutefois très sensible à la nature ou aux regroupements de données comme on peut le constater dans les travaux de McGloin et Piquero (2010) dans lesquels la corrélation entre la densité et la diversification n'est pas observable dans les données globales ou dans le sous-groupe des délits solos; une catégorie de données que l'on sait prépondérante chez les inconnus. Ainsi, l'absence de corrélation observée chez les inconnus est possiblement une confirmation des observations de McGloin et Piquero (2010). Les autres observations utilisant les ARS, et plus spécifiquement l'intermédiarité, présentent aussi un certain intérêt, mais il faudrait davantage de recherche pour améliorer la connaissance. Ces deux exemples à approfondir et à développer permettent de prendre la mesure de la complexité de la dynamique de ces variables et leur interdépendance.

5.4.4 Un potentiel à développer

D'autres possibilités ont aussi été examinées. Un exemple intéressant de comparaison des activités entre les individus connus et inconnus a été de considérer les activités des individus connus, dont les dates d'événements seraient antérieures à la première date d'identification, c'est-à-dire pour la période où ces individus connus étaient encore inconnus. Le volume de données à notre disposition nous permettait d'espérer trouver des observations intéressantes pour les comparer aux observations faites sur les individus qui demeurent toujours inconnus. En réalité, la séparation des

données, en période pré- et post-identification, apporte un volume de données très réduit pour de nombreux d'individus, qui de plus sont apparus plus actifs dans leur période connu qu'auparavant. Cette observation a ceci de bon, qu'elle confirme l'idée que les individus inconnus sont moins actifs que les connus; mais est-ce que ces derniers sont plus actifs une fois connus du fait qu'ils n'ont plus rien à perdre ? Pour ceux qui ont vécu un séjour en prison, est-ce que cette expérience a raffermi leur profil de carrière criminelle ?

À cela s'ajoute aussi la possibilité d'utiliser l'âge des individus tout autant que la croissance en gravité des derniers délits au moment de leur transition d'inconnu à connu. Sur ce dernier point, une exploration des données a permis de constater qu'il semblait bien y avoir une augmentation en gravité dans les quelques délits précédant l'arrestation comparativement aux délits antérieurs. Il s'agit là d'une piste des plus intéressantes à suivre pour préciser le comportement des inconnus d'autant plus qu'elle est aussi en accord avec l'hypothèse d'exposition.

Voilà autant de questions qui mériteraient d'être approfondies, mais qui nécessitent assurément un ajout substantiel de données sur les activités des individus criminalisés. Il existe un réel potentiel de poursuivre l'analyse des individus juste avant leur identification, en considérant aussi la présence de ceux qui sont encore inconnus. Dans notre situation, compte tenu du volume de données ADN disponible par rapport à celui des valeurs connues dans le monde des études sur les carrières criminelles, il n'était plus pertinent de pousser plus à fond ces questions, et bien qu'elles aient été partiellement explorées, elles restent assurément à développer dans de prochaines analyses sachant qu'on y détecte déjà un réel potentiel d'enrichissement des connaissances. La simple utilisation des données policières, incluant celles concernant les délits des individus connus, pourrait soutenir davantage une étude sur les activités criminelles dans cette transition entre les périodes inconnues et connues, puisque ces données représentent déjà un avantage pour la production de renseignement, comme nous le verrons plus loin.

5.6 La production de renseignement

Dans ce dernier volet de notre recherche, la production de renseignement prend place, non pas comme un exercice d'hypothèses à soutenir par des modèles statistiques, mais plutôt comme une proposition d'assemblage intégré et visuel d'informations génératrices de liens pour alimenter des réflexions d'enquêtes. Il s'agit de proposer un modèle, où serait adapté au cas par cas une mise en forme des informations, obtenues des enquêtes, afin de faire ressortir des possibilités de liens qui, autrement seraient inaccessibles. De manière habituelle, une enquête procède afin d'identifier des individus inconnus liés à une ou des scènes de crime. C'est ainsi que les concordances ADN structurées en réseau et incluant les inconnus seraient une base adéquate pour la production de renseignement.

Le modèle de schéma retenu se veut une simplification de la présentation des données, ce qui est toujours un atout dans une optique de synthèse (Rossy 2016). Quoique la structure des composants soit dynamique, cet aspect de la temporalité des événements criminels sera relégué à une deuxième phase de l'analyse. Le premier effort de structuration est axé sur la mise en place du maximum possible de délits effectués par les individus connus, complices des inconnus. On garde ici à l'esprit que les liens ADN ne représentent qu'une petite partie des activités criminelles d'un groupe d'individus, et qu'ainsi l'assemblage des autres délits connus associés à ce noyau d'individus pourrait mettre en évidence des liens sur des délits où les inconnus seraient potentiellement présents sans avoir laissé pour autant une trace d'ADN. C'est dans ce sous-groupe structuré que la notion de temporalité pourra être intégrée afin de mieux choisir les dossiers qui ont le plus de potentiel d'être étudiés en profondeur.

Les exemples utilisés montrent que la production de renseignement, qui utilise les dossiers et complices entourant les inconnus et les quelques individus connus qui y sont associés par co-délinquance peut rapidement devenir très complexe. D'où l'idée de garder l'attention sur les individus qui sont le plus près des inconnus. Il serait inutile d'essayer de mettre en place des schémas organisationnels de grande envergure intégrant

des informations très élaborées, le facteur limitant étant que chez tout interlocuteur, la limite de saturation de ce que les yeux et l'esprit peuvent emmagasiner et intégrer est rapidement atteinte. Pour être efficace, surtout dans un groupe de personnes qui se penchent sur une enquête, il faut porter l'attention sur ce qui est évident, efficace et porteur de nouvelles pistes à développer (Rossy et Ribaux 2012).

La grande difficulté d'une telle approche repose sur l'espoir de trouver des éléments pertinents qui permettraient de générer ces nouvelles pistes tant attendues. En effet, dans l'imaginaire populaire, les dossiers de police sont truffés de détails, l'enquêteur a accès à tout ce qu'il veut et la plupart du temps un « geek » d'informatique, présent au poste et entièrement disponible, trouve en un tour de main, dans les bases de données intégrées, tout ce qui est nécessaire à la poursuite de l'enquête dans l'heure qui suit! La réalité est tout autre. S'imaginer qu'il suffirait de bien s'attarder à l'intégration de ces multiples détails dans le schéma relationnel pour y déceler la voie à suivre relève de la fiction télévisuelle où tout tombe à point nommé. Dans la réalité, beaucoup d'enquêtes sont résolues grâce à un minimum d'informations ou piétinent devant un manque d'éléments de preuves. Et c'est sans compter que certaines informations, comme celles obtenues d'écoutes électroniques, qui pourraient être des plus intéressantes pour créer des liens entre de nombreux individus (Soudijn *et al*. 2022), ne sont nullement accessibles pour les enquêtes régulières. En effet, les résultats des mandats d'écoute sont sous contrôle judiciaire et, dans ce cas, la procédure stipule que les informations recueillies ne sont disponibles que pour les seuls dossiers pour lesquels les mandats ont été émis.

La production de renseignement se veut une réponse moderne à l'enquête moderne. L'assemblage d'informations, comme ces concordances ADN qui peuvent être structurées en composants, est la prémisse d'un monde s'apparentant au « big data ». Pour l'instant, l'expérience mise en place, par le truchement des quatre exemples, démontre qu'il n'est pas facile de rassembler des informations sur plusieurs dossiers criminels fermés et disséminés dans divers postes de police depuis des années. Outre l'absence possible des informations recherchées, dont il a été fait mention

précédemment, la centralisation incluant toutes les informations des dossiers de police est inexistante. Dans un tel contexte, la mise en place de schémas relationnels, pour appuyer les enquêtes, procède d'un parcours plus contraignant. L'exercice pourrait se faire au cas par cas pour des enquêtes qui demandent d'être approfondies, portant sur des délits graves. Toutefois, l'intégration globale de toutes les données policières aujourd'hui accessibles et ADN serait une prémisse des plus souhaitables pour simplifier l'implantation d'une approche générale de laquelle ressortirait, non pas seulement les dossiers prioritaires qui pourraient souffrir d'un manque d'informations, mais aussi tous les dossiers pour lesquels la production de renseignement serait porteuse de nouvelles pistes d'enquête. L'exemple du composant 44 (chap IV, Fig. 6B) est éloquent en ce sens que l'ajout des informations policières a permis de relier celui-ci à deux autres petits composants. Ces derniers peuvent apparaître isolés et inutiles dans l'ensemble des données du grand réseau ADN, mais leur lien avec un plus grand composant sous enquête peut être tout à fait pertinent pour l'aboutissement et l'identification d'inconnus. C'est ainsi que tout le potentiel d'intégration globale des données ADN et policières pourrait mettre au jour toutes les relations potentiellement utiles au développement des enquêtes. Il ne resterait qu'à faire le choix des dossiers à prioriser.

Outre l'aspect du renseignement tactique que nous proposons autour de l'identification des inconnus, il faut faire mention que l'utilisation des concordances ADN jointe aux données policières pourrait aussi permettre l'élaboration de renseignement stratégique, utile à un niveau plus large dans l'administration des enquêtes (Ribaux 2023). Autour des inconnus, peuvent s'ajouter des questionnements sur qui est en relation avec qui ? Où se concentrent les activités, dans les lieux et dans le temps ? Y-a-t-il une organisation particulière qui se dégage de la structure des informations en réseau ? Est-ce qu'un gang criminel serait ici une menace à la sécurité publique ?

5.7 Limitations

Une des faiblesses possibles à entrevoir dans l'approche utilisée ici réside dans les données informatiques. En effet, peu importe d'où viennent les données, plusieurs étapes des saisies manuelles d'informations généreront assurément des erreurs. Par exemple, les données ADN très bien structurées du LSJML ont nécessité diverses vérifications et corrections (voir annexe A). Les erreurs peuvent survenir à la saisie d'informations directement dans la fiche de concordance ou provenir en amont, au moment de la création du dossier informatique, et elles seraient ensuite disséminées.

La structure des données est un second point à considérer en ce qui concerne la possibilité d'utiliser des données informatiques. Ici, les informations disponibles dans la fiche de concordance du LSJML sont nombreuses et en lien, entre autres, avec la nature des délits au dossier, la date de l'événement et d'autres informations obtenues des concordances avec la BNDG. Cette étude a pu être réalisée grâce à la diversité des informations que l'on retrouve dans cette fiche. Cette dernière a été créée il y a plus de vingt ans avec l'idée d'être polyvalente, mais sans imaginer qu'elle serait fonctionnelle pour structurer les résultats de concordances en réseau. Pour une gestion minimale des concordances ADN, il n'est pas nécessaire d'avoir une fiche aussi complète, quand le but n'est que de transmettre les informations aux divers corps policiers. D'expérience, j'ai souvent vu une gestion des concordances ramenée à leur plus simple expression, dans un chiffrier, où ces dernières se résument à une liste de dossiers associés à des adresses postales, sans plus. Devant un tel désert d'informations, il devient beaucoup plus difficile de créer des réseaux, et l'assemblage des données manquantes demanderait des efforts de restructuration informatique importants. Il n'est donc pas donné à tous d'avoir accès directement à des séries de données permettant de structurer les réseaux et d'y associer des informations pertinentes pour la criminologie, à moins d'y consacrer quelques efforts supplémentaires. Ainsi, les données du LSJML sont dans une classe à part pour ce qui est de leur utilité dans des études criminologiques.

Outre les erreurs inhérentes aux saisies de données, la qualité de ces dernières peut aussi compromettre la valeur finale des analyses. À ce titre, le contrôle de la contamination des échantillons ou de la scène de crime, que ce soit par le personnel policier ou celui du laboratoire, est primordial, si on veut qu'un analyste puisse produire un travail pertinent, non aligné sur de fausses pistes, comme cela a été démontré (Lapointe *et al.* 2015). À ces contaminations par les travailleurs s'ajoute la présence incontrôlable de résultats ADN mis en banque, qui n'ont rien à voir avec le dossier criminel à l'étude. En effet, malgré l'attention portée à l'interprétation des scènes de crime, il est impossible d'avoir un contrôle absolu sur la pertinence des échantillons prélevés aux fins d'analyse, et certains se retrouvent en banque. Dans la réalité, ce groupe d'échantillons viendrait lentement augmenter la proportion des profils génétiques d'origine inconnue pour lesquels aucune enquête ou concordance au fichier des condamnés ne pourra apporter une identification. Par conséquent, sur le très long terme, on devrait s'attendre à ce que la proportion de ces inconnus « non pertinents » soit en augmentation alors qu'au fil du temps les enquêtes arriveraient à résoudre une proportion constante de délits par l'identification des profils ADN pertinents issus d'individus criminalisés. Devant ce problème intrinsèque à l'alimentation d'une banque ADN, en supposant que les contaminations par les travailleurs soient strictement maîtrisées, deux cas de figure ressortent. En premier lieu, les individus inconnus ayant à leur actif plusieurs délits et possiblement plusieurs co-délinquances pourraient être considérés comme pertinents dans de futures enquêtes. Toutefois, dans le cas du second groupe, le fait que les inconnus liés à un seul dossier demeurent dans la périphérie d'un composant, et ce, pendant de nombreuses années, pourraient laisser planer un doute sérieux sur la pertinence de leurs présences dans le composant d'activité criminelle[26]. Cette piste exploratoire pourrait facilement être déjà mise en place et permettre de clarifier certaines situations, mais évidemment il existerait toujours des exceptions, comme les individus qui, ayant appris de leurs erreurs, ne laissent plus de trace d'ADN suite à leur passage. Ces rares cas apparaîtraient eux aussi en périphérie des composants.

[26] À noter ici que l'on ne prend pas en compte les résultats uniques et isolés qui restent en banque, mais seulement les inconnus qui seraient présents dans un même dossier avec un autre profil qui, lui, a des liens de co-délinquance dans un ou plusieurs dossiers.

Outre la périphérie comme indice permettant de détecter la pertinence d'un lien, la nature du prélèvement analysé pourrait aussi donner une valeur de confiance au résultat observé. Il serait tout à fait normal d'accorder une valeur supérieure à une tache de sang, apparaissant fraîche au moment de l'intervention des enquêteurs sur la scène, comparativement à un mégot de cigarette ou une bouteille de bière qui pourrait être sur les lieux du délit depuis longtemps. Plusieurs questions doivent alors être envisagées pour clarifier ces situations. Dans l'exemple précédent, est-ce que la bouteille de bière est propre ou poussiéreuse ? Le mégot est-il seul dans un cendrier propre ou écrasé au sol et passablement abîmé ? L'analyste est donc placé devant un ensemble de d'informations variées qui exigent une recherche plus poussée, afin de mieux évaluer la pertinence des inconnus dans les composants sous enquête.

La limitation de l'échantillonnage apporte aussi son lot de biais. Les données du LSJML disponibles pour analyses sont limitées à la province du Québec, mais le crime ne connaît pas de frontière, c'est bien connu, et nombreux sont les individus qui, n'ayant qu'un seul délit criminel dans les fiches du LSJML, en ont d'autres à leur actif ailleurs au Canada. En effet, pour ce qui est de 3 % des individus non récidivistes (inscrits dans un seul dossier au Québec) on ne peut créer de liens avec leurs autres délits inscrits dans les autres provinces canadiennes puisque ces derniers ne sont inscrits que sous la forme d'une note dans la fiche du LSJML. Ainsi, un ensemble de composants qui pourraient s'avérer importants sont invisibles dans les données du LSJML. Ces individus, ayant des concordances hors Québec, sont au nombre de 199, et totalisent 261 si l'on inclut les récidivistes (sur lesquels 62 peuvent être situés dans leur réseau local). Par conséquent, le volumineux groupe des 7158 individus n'ayant qu'un délit solo demeure légèrement surévalué. Pourtant, les individus n'ayant qu'un seul délit à leur actif sont rares (2,5 %) (Farrington *et al.* 2014). Il s'ensuit donc qu'il y a un manque à gagner important d'information en lien avec le volume des activités criminelles de ces individus. Cette situation fait aussi naître l'hypothèse que les individus compétents à ne pas laisser de trace ADN durant leurs activités seraient aussi surreprésentés dans ce groupe.

L'exemple précédent apporte un éclairage plus précis sur les données globales du LSJML, qui quoique très complètes et volumineuses, peuvent présenter des éléments exceptionnels vu leur complexité. Il faut donc être à l'affût des situations particulières et porter une attention à dégager les grandes tendances. Toutefois, il ne faut pas oublier que les données ADN, même si elles sont volumineuses, demeureront toujours limitées. Elles ont l'avantage principal d'apporter des informations sur l'activité criminelle des inconnus mais l'ensemble des délits d'un individu ne demeurera toujours que partiel (De Moor *et al.* 2018). D'autant plus que le commerce des substances illicites, qui semble fréquent chez les inconnus, n'est pas évalué à sa juste hauteur à cause du manque de précisions dans les données du LSJML qui ne comptent que huit catégories, plutôt axées sur des crimes graves. C'est pourquoi il faut considérer nos analyses sur le nombre de récidives, de co-délinquance et le volume d'activité comme étant des estimations, qui permettraient de déceler des tendances, mais qui pourraient facilement être aussi en deçà de la réalité. Il demeure toutefois possible que les données ADN puissent être considérées comme un sous-groupe représentatif de la réalité, mais à la condition qu'elles apparaissent de manière totalement aléatoire dans l'ensemble des événements criminels de tous les individus impliqués (De Moor *et al.* 2018). Dans les faits, cette condition est possiblement partiellement respectée, car seuls les individus très consciencieux de la gestion de leur ADN sur une scène de crime seraient en mesure de fausser la distribution aléatoire des profils génétiques. Par contre, on ne connaît pas la proportion de ces individus dans la population concernée, et seules des études supplémentaires pourraient apporter un peu de lumière sur cet aspect de la distribution des compétences des criminels. Toutefois, compte tenu du nombre important d'individus connus laissant leur ADN sur de nombreuses scènes de crime, même en admettant qu'une certaine proportion des inconnus soient "compétents", ces derniers seraient probablement peu nombreux à pouvoir imposer leur biais de compétence.

5.8 Conclusion générale

La motivation principale derrière ce projet était de mettre en place une nouvelle approche de production de renseignement tactique, utilisant les données de concordances

ADN qui incluent des individus inconnus, dans le but de mettre au jour des pistes d'enquêtes pouvant mener à l'identification de ces inconnus.

Le modèle utilisé pour la restructuration des données de concordance en un format réseau permet d'avoir accès à une visualisation globale et synthétique de l'ensemble des relations de co-délinquances que ces individus inconnus entretiennent dans leurs activités criminelles avec les individus connus des services de police.

Le premier groupe d'analyse utilisant quelques paramètres de réseaux sociaux a permis de constater que ces inconnus étaient tout aussi bien intégrés que leurs complices connus, montrant ainsi la pertinence de ces inconnus dans les structures en réseau. Des exemples ont aussi permis de constater que le positionnement d'inconnus ayant une intermédiarité élevée pouvait avoir des conséquences importantes sur l'approche tactique du développement d'une enquête.

L'analyse de l'activité criminelle des inconnus a permis de démontrer que ces derniers avaient tendance à procéder dans leurs délits en faisant profil bas, évitant le plus possible d'attirer l'attention des policiers, qui interviennent plus promptement et en profondeur dans le cas de délits graves.

Par les exemples utilisés, il a été démontré que la structure en réseau des concordances ADN en y associant son volet dynamique, permettait de porter une attention particulière aux individus connus ayant eu un ou des inconnus comme complice et que l'ajout d'informations policières permettait de mettre au jour plusieurs délits criminels pour lesquels des informations pourraient apporter certaines pistes d'enquêtes pouvant possiblement mener à l'identification du ou des inconnus.

De manière générale, notre étude permet donc un développement de l'utilisation des concordances ADN axées sur la production du renseignement afin d'aider à l'identification des individus inconnus. Cette approche est tributaire de deux prémisses majeures. Premièrement, les données de concordances doivent être suffisamment

structurées et riches en informations pour permettre la restructuration en réseau des données et deuxièmement, la collaboration policière est essentielle pour que l'on puisse ajouter les nombreuses informations associées aux individus connus, complices des inconnus. Ce type de production de renseignement propose une approche tactique de développement de pistes d'enquêtes, et sa mise en pratique devrait permettre d'améliorer le taux d'identification des inconnus sans mettre de côté une utilisation possible pour des approches stratégiques et administrative.

Plusieurs des observations faites au cours de cette étude ont le potentiel d'être approfondies et les données utilisées ont elles aussi le potentiel de générer des sujets de recherches variés propres à la criminologie. Finalement, l'utilisation d'exemples est forcément une limitation inhérente à un projet de recherche, et on peut très bien envisager l'application à grande échelle du concept de réseau ADN pour de la production de renseignement tactique à l'ensemble des données du LSJML et des données policières du Québec pour éventuellement diffuser le concept à la grandeur du pays.

5.9 Épilogue

Au moment de finaliser la présente thèse pour son dépôt final, nous apprenons que l'analyste de la Sûreté du Québec, M. Chartrand, avec qui nous avons collaboré pour la production de renseignement décrite au Chap. IV, inspiré par les exemples de production de renseignement, a procédé à des comparaisons entre l'ensemble des données de concordances ADN du LSJML et l'ensemble des données policières à sa disposition. Selon son témoignage, cet exercice d'analyse globale des données lui a permis de mettre au jour des centaines de recoupements de dossiers, lui permettant d'associer facilement des noms d'individus à des inconnus dans d'autres dossiers.

À titre d'exemple, il explique que deux dossiers d'agressions sexuelles, espacés de seulement deux ans, n'avait pas trouvé leur aboutissement puisque l'un était classé pour insuffisance de preuve et l'autre pour retrait de plainte. Dans ce dernier dossier cinq

suspects avaient été identifiés, alors que dans celui pour lequel les preuves manquaient, on retrouvait cinq inconnus. Selon M. Chartrand, le classement de ces dossiers ne permettait pas de poursuivre les enquêtes encore moins d'y détecter un rapprochement sur la base des cinq individus mentionnés de part et d'autre. En y associant les données ADN, il constate que le LSJML possédait une fiche ADN pour un sixième individu connu dans les deux dossiers d'agression sexuelle. Les enquêtes n'ayant pas permis de porter des accusations sur cet individu, les dossiers avaient été fermés pour les raisons énumérées ci-dessus. Toutefois, on voit ici que l'assemblage des données ADN et policières a produit du renseignement en établissant un lien solide d'identification entre ces deux dossiers fermés, sur lesquels il lui aura été permis de déduire que les cinq suspects connus de l'individu fiché par ADN, sont les même que les cinq individus inconnus présent au dossier pour lequel la preuve manquait. La voie proposée à la fin de notre conclusion, d'utiliser l'ensemble des données, a trouvé son écho dans la réalité de l'opérationnel. Merci M. Chartrand pour votre curiosité, vos compétences et cet aboutissement.

RÉFÉRENCES BIBLIOGRAPHIQUES

Baechler, S., Morelato, M., Ribaux, O., Beavis, A., Tahtouh, M., Kirkbride, K. P., Esseiva, P., Margot, P. et Roux, C. (2015) Forensic intelligence framework. Part II: Study of the main generic building blocks and challenges through the examples of illicit drugs and false identity documents monitoring. *Forensic Science International* 250, 44-52.

Bastian, M., Heyman, S. et Jacomy, M. (2009) Gephi: an open source software for exploring and manipulating networks. *International AAAI conference on weblogs and social media* 8, 361-362.

Batagelj, V. and Mrvar A. (2001). A subquadratic triad census algorithm for large sparse networks with small maximum degree. *Social Networks* 23, 237-243.

Bernasco, W., Lammers, M. et Van der Beek, K. (2016) Cross-border crime patterns unveiled by exchange of DNA profiles in the European Union. *Security Journal* 29, 640-660.

Bichler, G. (2019) *Understanding Criminal Networks: A Research Guide.* University of California Press.

Bijleveld, C. et Hendriks, J. (2003) Juvenile sex offenders: Differences between group and solo offenders. *Psychology, crime and law* 9, 237-245.

Blomberg, T. G., Bales, W. D. et Piquero, A. R. (2012) Is educational achievement a turning point for incarcerated delinquents across race and sex? *Journal of youth and adolescence* 41, 202-216.

Blumstein, A., Cohen, J., Das, S. et Moitra, S. D. (1988) Specialization and seriousness during adult criminal careers. *Journal of Quantitative Criminology* 4, 303-345.

Blumstein, A., Cohen, J., Piquero, A. R. et Visher, C. A. (2010) Linking the crime and arrest processes to measure variations in individual arrest risk per crime (Q). *Journal of Quantitative Criminology* 26, 533-548.

Borgatti, S.P. Everett, M. B et Freeman, L.C. (2002) *Ucinet for Windows: Software for Social Network Analysis*. Analytical Technologies, Harvard, MA

Borgatti, S. P., Everett, M.B. et Johnson, J.C. (2018). *Analyzing social networks*, Sage.

Bright, D., Whelan, C. et Ouellet, M. (2022) Assessing variation in co-offending networks. *Global Crime* 23, 101-121.

Broséus, J., Baechler, S., Gentile, N. et Esseiva, P. (2016) Chemical profiling: A tool to decipher the structure and organisation of illicit drug markets: An 8-year study in Western Switzerland. *Forensic Science International* 266, 18-28.

Bruenisholz, E., Wilson-Wide, L., Ribaux, O. et Delémont, O. (2019) Deliberate fires: from data to intelligence. *Forensic Science International* 301, 240-253

Budowle, B., Ge, J., Chakraborty, R., Eisenberg, A. J., Green, R., Mulero, J., Lagace, R. et Hennessy, L. (2011) Population genetic analyses of the NGM STR loci. *International journal of legal medicine* 125, 101-109.

Charette, Y. et Papachristos A. V. (2017) The network dynamics of co-offending careers. *Social Networks* 51, 3-13.

Clare, J. (2011) Examination of systematic variations in burglars' domain-specific perceptual and procedural skills. *Psychology, Crime & Law* 17, 199-214.

Conrad, K. J., Riley, B. B., Conrad, K. M., Chan, Y. F. et Dennis, M. L. (2010) Validation of the Crime and Violence Scale (CVS) against the Rasch measurement model including differences by gender, race, and age. *Evaluation review* 34, 83-115.

Cressey, D. R. (1960) The theory of differential association: An introduction. *Social. Probabilistics* 8, 2.

Crispino, F., Rossy, Q., Ribaux, O. et Roux, C. (2015) Education and training in forensic intelligence: a new challenge. *Australian Journal of Forensic Sciences* 47, 49-60.

Delémont, O., Bitzer, S., Jendly, M. et Ribaux, O. (2017) The practice of crime scene examination in an intelligence-bases perspective. Dans *The Routledge International Handbook of Forensic Intelligence and Criminology*. Routledge.

De Moor, S., Vander Beken, T. et Van Daele, S. (2017) DNA Databases as Alternative Data Sources for Criminological Reseach. *European Journal on Criminal Policy and Research* 23, 175-192

De Moor, S., Vandeviver, C. et Vander Beken, T. (2018) Are DNA data a valid source to study the spatial behaviour of unknown offenders? *Science & Justice* 58, 315-322

De Moor, S., Vandeviver, C. et Vander Beken, T. (2020) Assessing the missing data problem in criminal network analysis using forensic DNA data. *Social Networks* 61, 99-106.

Egger, S. A. (1984) A working definition of serial murder and the reduction of linkage blindness. *Journal of police science and administration* 12, 348-357.

Esseiva, P., Ioset, S., Anglada, F., Gasté, L., Ribaux, O., Margot, P., Gallusset, A., Biedermann, A., Specht, Y. et Ottinger, E. (2007) Forensic drug Intelligence: An important tool in law enforcement. *Forensic Science International* 167, 247-254.

Farrington, D. P., Ttofi, M., Crago, R. W. et Coid, J. W. (2014) Prevalence, frequency, onset, desistance and criminal career duration in self-reports compared with official records. *Criminal Behaviour and Mental Health* 24, 241-253.

Freeman, L. C. (1978) Centrality in social network conceptual clarification. *Social Network* 3, 215-239.

Freeman, L. C. (2004) *The development of social network analysis*. A Study in the Sociology of Science. Empirical press, Vancouver

Gottfredson, M. R. et Hirschi, T. (1990) *A general theory of crime*. Stanford University Press.

Gendarmerie royale du Canada. *Rapport annuel de la Banque nationale de données génétique*. (2018-2019), 13. (http://www.rcmp-grc.gc.ca/pubs/nddb-bndg/index-eng.htm).

Gründ, T. et Morselli, C. (2017) Overlapping crime: Stability and specialization of co-offending relationships. *Social Networks* 51, 14-22.

Harper, W. R. et Harris D. H. (1975) The application of link analysis to police intelligence. *Human Factors* 17, 157-164.

Haynie, D. L. (2001) Delinquent peers revisited: Does network structure matter? *American Journal of sociology* 106, 1013-1057.

Haynie, D. L. (2002) Friendship networks and delinquency: The relative nature of peer delinquency. *Journal of Quantitative Criminology* 18, 99-134.

Heller, N. B. et McEwen, J. T. (1973) Applications of crime seriousness information in police departments. *Journal of criminal justice* 1, 241-253.

Hochholdinger, S., Arnoux, M., Delémont, O. et Esseiva, P. (2019A) Forensic intelligence on illicit markets: the example of watch counterfeiting. *Forensic Science International* 302, 109868.

Hochholdinger, S., Marvin, L., Arnoux, M., Esseiva, P. et Delémont, O. (2019B) Elemental analysis for profiling counterfeit watches. *Forensic Science International* 298, 177-185.

Hopkinson, D. A., Mestriner, M. A., Cortner, J. et Harris, H. (1973) Esterase D: a new human polymorphism. *Annals of Human Genetics* 37, 119-137.

Jeffreys, A. J., Wilson, V. et Thein S. L. (1985) Individual-specific ‘fingerprints’ of human DNA. *Nature* 316, 76-79.

Jeuniaux, P. P. J. M. H., Renard, B., Duboccage, L., Steuve, S., Stappers, C., Gallala, I., De Moor, S., Jonckheere, A., Mine, B., Vanhooydonck, B., Kempenaers, M., De Greef, C., Van Renterghen, P. et Vanvooren, V. (2015) Managing forensic DNA records in a divided world: the Belgian case. *Records Management Journal* 25, 269-287.

Jeuniaux, P. P. J. M. H., Duboccage, L., Renard, B., Van Renterghem, P. et Vanvooren, V. (2016) Establishing networks in a forensic DNA database to gain operational and strategic intelligence. *Security Journal* 29, 584-602.

Jeuniaux, P. P. J. M. H., De Moor, S., Robert, L., Renard, B., Stappers, C. et Vanvooren V. (2017) Reconstruction and study of offending trajectories through forensic evidence. Dans *The Routledge International Handbook of Forensic Intelligence and Criminology*. Routledge.

Johnson, S. D., Summers, L. et Pease, K. (2009) Offender as forager? A direct test of the boost account of victimization. *Journal of Quantitative Criminology* 25, 181-200.

Kelty, S. F., Julian, R. et Ross, A. (2013) Dismantling the justice silos: avoiding the pitfalls and reaping the benefits of information-sharing between forensic science, medicine and law. *Forensic Science International* 230, 8-15.

Lammers, M. (2014) Are Arrested and Non-Arrested Serial Offenders Different? A Test of Spatial Offending Patterns Using DNA Found at Crime Scenes. *Journal of Research in Crime and Delinquency* 51, 143-167.

Lammers, M., Bernasco, W. et Elffers, H. (2012) How Long Do Offenders Escape Arrest? Using DNA Traces to Analyse when Serial Offenders Are Caught. *Journal of Investigative Psychology and Offender Profiling* 9, 13-29.

Lammers, M. et Bernasco, W. (2013) Are mobile offenders less likely to be caught? The influence of the geographical dispersion of serial offenders' crime locations on their probability of arrest. *European Journal of Criminology* 10, 168-186.

Lapointe, M., Rogic, A., Bourgoin, S., Jolicoeur, C. et Séguin, D. (2015) Leading-edge forensic DNA analyses and the necessity of including crime scene investigators, police officers and technicians in a DNA elimination database. *Forensic Science International: Genetics* 19, 50-55.

Lavergne, L., Boivin, R., Baechler, S., Jeuniaux, P., Fiola, K., Séquin, D., Lefebvre, J-F. et Milot E. (2022) Determining the impact of unknown individuals in criminality using network analysis of DNA matches. Forensic Science International 331: 11142

Malm, A. et Bichler G. (2011) Networks of collaborating criminals: Assessing the structural vulnerability of drug markets. *Journal of Research in Crime and Delinquency* 48, 271-297.

Menting, B., Lammers, M., Ruiter, S. et Bernasco, W. (2016) Family matters: effect of family members' residential areas on crime location choice. *Criminology* 54, 413-433.

McGloin, J. M. et Piquero, A. R. (2010) On the Relationship between Co-Offending Network Redundancy and Offending Versatility. *Journal of Research in Crime and Delinquency* 47, 63-90.

McGloin, J. M., Sullivan, C. J., Piquero, A. R. et Bacon, S. (2008) Investigating the stability of co-offending and co-offenders among a sample of youthful offenders. *Criminology* 46, 155-188.

Milliet, Q., Delémont, O. et Margot, P. (2014) A forensic science perspective on the role of images in crime investigation and reconstruction. *Science and Justice* 54, 470-480.

Milot, E., Lecomte, M., Germain, H. et Crispino, F. (2013) The national DNA data bank of Canada: a Quebecer perspective. *Frontiers in genetics* 4, 249.

Morelato, M., Baechler, S., Ribaux, O., Beavis, A., Tahtouh, M., Kirkbride, P., Roux, C. et Margot, P. (2014) Forensic intelligence framework—Part I: Induction of a transversal model by comparing illicit drugs and false identity documents monitoring. *Forensic Science International* 236, 181-190.

Morelato, M., Beavis, A., Tahtouh, M., Ribaux, O., Kirkbride, P. et Roux, C. (2013) The use of forensic case data in intelligence-led policing: The example of drug profiling. *Forensic Science International* 226, 1-9

Moreno, J. L. (1953) *Who shall survive? Foundations of sociometry, group psychotherapy and socio-drama*. Washington DC

Morris, J. P. (1982) *Crime analysis charting: an introduction to visual investigative analysis*. Palmer Enterprises Loomis.

Morselli, C. (2009A) Hells Angels in springtime. *Trends in organized Crime* 12, 145-158.

Morselli, C. (2009B) *Inside criminal network..* Springer.

Morselli, C. (2013) *Crime and networks*. Routledge.

Morselli, C. et Boivin, R. (2016) *Les réseaux criminels.* Les Presses de l'Université de Montréal.

Morselli, C. et Ouellet, M. (2018) Network similarity and collusion. *Social Networks* 55, 21-30.

Mousseau, V., Baechler, S. et Crispino, F. (2019) Management of crime scene units by Quebec police senior managers: Insight on forensic knowledge and understanding of key stakeholders. *Science and Justice* 59, 524-532.

Ouellet, F. et Bouchard, M. (2017) Only a matter of time? The role of criminal competence in avoiding arrest. *Justice Quarterly* 34, 699-726.

Papachristos, A. V. (2017) The coming of a networked criminology? dans *Measuring crime and criminality*. Routledge, 101-140.

Peterson, M. B., Morehouse, B. et Wright, R. (2000) *Intelligence 2000: revising the basic elements*. 2nd print. Sacramento-CA: Lawrenceville-NJ: Law Enforcement Intelligence Unit-LEIU.

Pyrooz, D. C., Mc Gloin, J. M. et Decker, S. H. (2017) Parenthood as a turning point in the life course for male and female gang members: a study of within-individual changes in gang membership and criminal behaviour. *Criminology* 55, 869-899.

Quick, C. B., Fisher, R. A. et Harris, H. (1974) A kinetic study of the isozymes determined by the three human phosphoglucomutase loci PGM1, PGM2 and PGM3. *European journal of biochemistry* 42, 511-517.

Reiss, A. J. (1951) Delinquency as the failure of personal and social controls. *American Sociological Review* 16, 196-207.

Ribaux, Olivier (2023) De la police scientifique à la traçologie; Le renseignement par la trace. La fondation des Presses polytechniques et universitaires romandes (PPUR) 2iè édition, 578 p. ISBN 978-2-88915-544-6

Ribaux, O. et Margot, P. (1999) Inference structures for crime analysis and intelligence: the example of burglary using forensic science data. *Forensic Science International* 100, 193-210.

Ribaux, O., Baylon, A., Roux, C., Delémont, O., Lock, E., Zingg, C. et Margot, P. (2010A) Intelligence-led crime scene processing. Part I: Forensic intelligence. *Forensic Science International* 195, 10-16.

Ribaux, O., Baylon, A., Lock, E., Delémont, O., Roux, C., Zingg, C. et Margot, P. (2010B) Intelligence-led crime scene processing. Part II: Intelligence and crime scene examination. *Forensic Science International* 199, 63-71.

Ribaux, O. et Caneppele, S. (2017) Forensic intelligence. Dans *The Routledge International Handbook of Forensic Intelligence and Criminology*. Routledge.

Ri**b**aux, O. et Talbot Wright B. (2014) Expanding forensic science through forensic intelligence. *Science and Justice* 54, 494-501.

Ribaux, O., Walsh, S. J. et Margot, P. (2006) The contribution of forensic science to crime analysis and investigation: Forensic intelligence. *Forensic Science International* 156, 171-181.

Rossy, Q. (2016) La visualisation relationnelle au service de l'enquête criminelle. Dans *Les réseaux criminels*. Les presses de l'Université de Montréal

Rossy, Q., Ioset, S., Dessimoz, D. et Ribaux, O. (2013) Integrating forensic information in a crime intelligence database. *Forensic Science International* 230, 137-146.

Rossy, Q. et Ribaux, O. (2012) La conception de schémas relationnels en analyse criminelle: au-delà de la maîtrise des outils. *Revue Internationale de Criminologie et de Police Technique et Scientifique* 65, 345-362.

Rossy, Q. et Ribaux, O. (2014) A collaborative approach for incorporating forensic case data into crime investigation using criminal intelligence analysis and visualisation. *Science and justice* 54, 146-153.

Schroeder, J., Xu, J., Chen, H. et Chau, M. (2007) Automated criminal link analysis based on domain knowledge. *Journal of the American Society for Information Science and Technology* 58, 842-855.

Soudijn, M. R., Vermeulen, I. J. et Van der Leest, P. E. (2022) When encryption fails: a glimpse behind the curtain of synthetic drug trafficking networks. *Global Crime* 23, 1-24.

Sutherland, E.H. (1947) *Principles of Criminology*. New York. Lippincott

Sutherland, E. H., et Cressey, D.R. (1992) *Principles of criminology*, Altamira Press.

Sparrow, M. K. (1991) The application of network analysis to criminal intelligence: An assessment of the prospects. *Social Networks* 13, 251-274.

Van Mastrigt, S. B. et. Farrington D. P. (2009) Co-offending, age, gender and crime type: Implications for criminal justice policy. *The British Journal of Criminology* 49, 552-573.

Von Lampe, K. (2021) Remembering Carlo Morselli. *Trends in organized Crime* 24, 378-383.

Wigmore, J. H. (1913) *The Principles of judicial Proof as Given by Logic, Psychology, and General Experience, and Illustrated in Judicial Trials.* Boston: Little, Brown and Company

Weerman, F. M. (2003) Co-offending as Social Exchange. Explaining Characteristics of Co-offending. *British journal of criminology* 43, 398-416.

Weir, B. S. (2007) The rarity of DNA profiles. *The annals of applied statistics* 1, 358.

ANNEXE A

Gestion et vérification des données, concept de réseau et composants

1. Introduction

Les données de concordances ADN utilisées proviennent du LSJML et consistent en un fichier Excel (.csv) créé par un script R qui compile les concordances des fiches et les informations qui s'y trouvent en les regroupant par composants, c.-à-d. : les groupes d'individus qui sont liés entre eux par des délits criminels. Ceux qui ne sont pas liés par des activités de co-délinquances demeurent des individus solos (ils sont toutefois comptés dans le nombre total de composants). Toutes les données sensibles sont rendues anonymes par le LSJML. Au nombre de celles-ci on trouve le numéro de fiche, les nom et prénom de l'individu (si connu) et son numéro de FPS (identification digitale à la GRC), le numéro de dossier et celui relatif au prélèvement et l'identification du délit criminel (ID de cas). Ces données sont codées aléatoirement à chaque envoi à l'exception du numéro de fiche qui reste le même pour permettre au laboratoire de garder une clé lui permettant de retourner aux données d'origine lorsque nécessaire et aussi d'avoir dans son étude une référence stable entre divers lots de données. Dans le cadre de notre étude, pour le premier chapitre, nous avons utilisé un lot de données compilé en juillet 2018 et pour les deux autres chapitres,un second groupe compilé en juillet 2019.

Dans le fichier de 2019, on retrouve 24 641 lignes représentant tous les liens (ou concordances) individu-dossier compilés depuis la création de la banque en juillet 2000. Outre les données codées énumérées précédemment, on y retrouve aussi toute une série d'informations concernant les types de délits, la présence de combinaisons d'ADN, le lieu approximatif (poste de police ou région) associé au délit, la date d'identification de l'individu obtenue du fichier des condamnés et la date de l'événement, qui sont les plus intéressantes pour l'axe de recherche de nos travaux. Le tableau 1 donne un exemple des données extraites des fiches de concordance du LSJML, et un peu plus loin nous aborderons la gestion de ces informations et les vérifications faites sur celles-ci. Mais avant cela nous aborderons, la notion de composant dans la section suivante. Vous

constaterez que la simple lecture des informations du tableau 1 permet d'observer la structure des composants.

2. Réseau et composant

L'ensemble des données, c'est-à-dire les 24 641 délits associés aux 11 893 individus, représentent dans les faits le réseau complet des informations criminelles liées aux concordances. Toutefois, même si on peut imaginer que tous les individus criminels du Québec pouvaient être interreliés en réseau, dans la réalité ce n'est pas le cas. Il n'y a qu'une partie des délits criminels qui relie entre eux une partie des individus. Ainsi dans le réseau, on observe des composants qui sont en fait des sous-groupes d'individus (de 2 à 37 personnes) qui ont en commun, un ensemble de dossiers où chacun des individus a commis, avec au moins un autre individu, un minimum d'un délit criminel en co-délinquance.

Comme on le voit dans le tableau 1, les données sont regroupées en composants (numérotés dans la colonne X) où les individus sont identifiés par le « numéro de fiche » (colonne B) et les délits criminels par l'« ID de cas » (colonne V). Ces deux paramètres permettent de construire la structure en réseau des composants. Je vous propose une visite guidée des liens du tableau 1, que l'on peut observer dans le composant 42 en croisant les données des colonnes « B » où l'on trouvera les dix individus de ce composant et leurs dossiers associés dans la colonne « V ».

Voyez au haut de la colonne « V » le dossier « XAC2634 » (surligné en jaune) qui se retrouve à trois emplacements, ce qui se traduit par autant de complices. En effet, dans ce dossier on a retrouvé l'ADN de trois individus, les « 59091 », « 59088 » et le « 30900 » (surlignés en jaune dans la colonne « B » « numéro de fiche »). Déjà, plusieurs constatations :

Premièrement, les individus « 59091 » et « 59088 » ne sont vus nulle part ailleurs dans ce composant. Ils n'ont donc été vus qu'une seule fois dans le dossier qui leur est

attribué, et cette observation est aussi valable pour l'ensemble des résultats de tout le réseau. Ces individus présents dans ce composant ne peuvent être présents ailleurs.

Tableau 1 : Exemple des données reçues du LSJML pour les composants 42 et 43. Les couleurs sont en lien avec les descriptions dans le texte.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
Numero	Numero	Date	Prenom	Nom	Date	FPS	Annee_abi	Ville	Numero	Type	code	DateJD	Geolocalisation	date_mise	date	Match_inter	date_décès	Identite	ID Labo	Année	ID_Cas	Nombre_Individus	component	Ordre	Taille	Nombre	Date
Dossier	Fiche	Inscription			Naissance				Relatif	délit	suffixe	National		en banque	événement	provincial				complète		par prélèvement	ID	Réseau	Réseau	Individu	événement
		fiche							Prélèvement						labo											Réseau	police
QBL5556	59091	2018-08-21	YKBM	JQVD	1982-10-15	NA	17	M	WIQ7808	AD	PX	NA	SQ, Crimes économiques, Montréal,	2018-05-28	2017-11-07	NA	NA	connu	oui	2017	XAC2634	1	42	43	42	10	2017-10-12
JVO4566	59088	2018-08-21	ZHJR	LNIH	1976-10-15	NA	17	M	EDA4745	AD	PX	NA	SQ, Crimes économiques, Montréal,	2018-05-28	2017-11-07	10	NA	connu	oui	2017	XAC2634	1	42	43	42	10	2017-10-12
MKR7179	43878	2013-11-08	NA	NA	NA	NA	13	M	BMU8530	AD	PX	NA	SPVM, Incendies criminels, Montréal	2013-09-09	2013-05-04	NA	NA	inconnu	non	2013	WAY3578	1	42	43	42	10	2013-05-04
OYP8274	43878	2013-11-08	NA	NA	NA	NA	13	M	GCE4247	AD	PX	NA	SPVM, Incendies criminels, Montréal	2013-09-18	2013-05-03	NA	NA	inconnu	non	2013	KVH4097	1	42	43	42	10	2013-05-03
PJD8798	43533	2014-11-11	IAOO	UTL	NA	NA	14	M	ATN4582	AA	NA	NA	SPVM, Sections tupéfiants, région Est,	2014-11-03	2014-07-04	NA	NA	connu	non	2014	RTP3715	1	42	43	42	10	2014-07-04
UND4191	43533	2013-10-02	IAOO	UTL	NA	NA	13	M	IVE9040	AD	PX	NA	SPVM, Incendies criminels, Montréal	2013-09-09	2013-05-04	NA	NA	connu	non	2013	WAY3578	1	42	43	42	10	2013-05-04
PPB2965	43533	2013-10-02	IAOO	UTL	NA	NA	5	M	IMI8418	IN	NA	NA	SPVM, Centre opérationnel Est,	2008-07-11	2005-01-09	NA	NA	connu	non	2005	KZW8219	1	42	43	42	10	2005-01-09
HOI8747	43287	2013-09-12	ASOR	ZEME	1987-03-15	YEW5174	13	M	UOV3142	AD	NA	2013-09-10	SPVM, Incendies criminels, Montréal	2013-09-09	2013-05-04	NA	NA	connu	non	2013	WAY3578	1	42	43	42	10	2013-05-04
PYD8562	38004	2014-08-14	UKBP	EDDC	1982-03-15	EOS9154	13	M	UAS4656	AD	NA	2014-08-26	SQ, MRC de Vaudreuil-Soulanges Ouest,	2014-07-30	2013-11-12	NA	NA	connu	oui	2013	KFS4105	1	42	43	42	10	2013-11-12
CBX8332	38004	2013-09-16	UKBP	EDDC	1982-03-15	EOS9154	13	M	GKV7508	AD	PX	2013-09-24	SPVM, Incendies criminels, Montréal	2013-09-09	2013-05-04	NA	NA	connu	non	2013	WAY3578	1	42	43	42	10	2013-05-04
MPA6511	38004	2011-11-21	UKBP	EDDC	1982-03-15	EOS9154	7	M	IOM2389	HO	NA	2011-11-18	SPVM, Centre opérationnel Nord	2008-09-30	2007-04-12	NA	NA	connu	non	2007	FMN8421	1	42	43	42	10	2007-04-12
IKI9985	36942	2011-05-30	VBDE	QVLW	1976-09-15	BTZ9763	7	M	SKY4022	HO	PX	2011-05-20	SPVM, Centre opérationnel Nord	2008-09-30	2007-04-12	NA	NA	connu	non	2007	FMN8421	1	42	43	42	10	2007-04-12
HSN3243	31614	2009-02-12	KFNR	ZAUB	1989-03-15	HXN3403	7	M	VOW1553	IN	4/9	NA	SPVM, Centre opérationnel Est,	2009-01-06	NA	NA	NA	connu	non	2007	SOF5058	1	42	43	42	10	2007-07-25
DCH9825	31614	2009-02-12	KFNR	ZAUB	1989-03-15	HXN3403	4	M	BVY0625	IN	NA	2011-11-14	SQ, poste principal, MRC de Matawinie,	2005-03-15	2004-02-03	NA	NA	connu	non	2004	XGT7321	1	42	43	42	10	2004-02-03
QUY2992	30900	2018-05-31	TCNJ	IEPF	1979-07-15	NBP5805	17	M	CTB8088	AD	PX	2018-05-29	SQ, Crimes économiques, Montréal,	2018-05-28	2017-11-22	NA	NA	connu	non	2017	XAC2634	1	42	43	42	10	2017-10-12
SKG3121	30900	2008-10-16	TCNJ	IEPF	1979-07-15	NBP5805	7	M	HUF5633	HO	PX	2008-10-03	SPVM, Centre opérationnel Nord	2008-09-30	2007-04-12	NA	NA	connu	non	2007	FMN8421	1	42	43	42	10	2007-04-12
NXG7798	30900	2015-08-03	TCNJ	IEPF	1979-07-15	NBP5805	5	M	BAO0068	HO	NA	2015-08-02	Sûreté municipale de Laval, enq criminelles	2015-08-01	2005-03-10	NA	NA	connu	non	2005	HDX8715	1	42	43	42	10	2005-03-10
KAF8003	28319	2013-09-20	WZDP	NXXR	1988-08-15	HQG7020	10	M	QNW8444	HO	NA	NA	SPVM, Division des crimes majeurs,	2013-08-12	2010-09-04	NA	2010-09-04	dcd	non	2010	HDD9381	1	42	43	42	10	2010-09-04
JEL2840	28319	2009-02-04	WZDP	NXXR	1988-08-15	HQG7020	7	M	KCU6813	IN	NA	2009-10-28	SPVM, Centre opérationnel Est,	2009-01-06	NA	NA	2010-09-04	dcd	non	2007	SOF5058	1	42	43	42	10	2007-07-25
DID9142	28319	2008-08-21	WZDP	NXXR	1988-08-15	HQG7020	5	M	KXU6221	IN	2/11	NA	SPVM, Centre opérationnel Est,	2008-07-11	2005-01-09	NA	2010-09-04	dcd	non	2005	KZW8219	1	42	43	42	10	2005-01-09
YJY5873	28319	2008-08-21	WZDP	NXXR	1988-08-15	HQG7020	4	M	NUV2019	HO	NA	2009-10-28	SPVM, Anti-Gang, Montréal	2005-08-26	2004-10-15	NA	2010-09-04	dcd	oui	2004	GJE4144	1	42	43	42	10	2004-10-15
TTU0882	59457	2019-03-21	NA	NA	NA	NA	18	M	DHE3538	IN	NA	NA	SPVM, Centre opérationnel Nord,	2019-03-18	2018-11-22	NA	NA	inconnu	non	2018	EJY0882	1	43	42	42	8	2018-11-22
VTE1171	59457	2019-03-19	NA	NA	NA	NA	18	M	DHE3538	IN	NA	NA	SPVM, Centre opérationnel Est,	2019-03-13	2018-11-20	NA	NA	inconnu	non	2018	OPM4352	1	43	42	42	8	2018-11-19
MEB8785	59457	2018-08-01	NA	NA	NA	NA	18	M	UEO1377	IN	NA	NA	Service de police de Laval,	2018-07-25	2018-04-22	NA	NA	inconnu	non	2018	SST0508	1	43	42	42	8	2018-04-22
ERC0980	59457	2018-08-01	NA	NA	NA	NA	16	M	KOC0546	SR	NA	NA	SPVM, Centre opérationnel Est,	2016-05-04	2016-02-20	NA	NA	inconnu	non	2016	RKT8419	1	43	42	42	8	2016-02-20
VSH1642	58578	2018-05-03	NA	NA	NA	NA	17	M	QZN0101	AA	NA	NA	Service de police de Saint-Jérôme,	2018-04-30	2017-11-30	NA	NA	inconnu	non	2017	MUH2312	1	43	42	42	8	2017-11-30
SKX8579	58578	2018-05-03	NA	NA	NA	NA	16	M	OND4324	AA	NA	NA	SQ, MRC de La Rivière-du-Nord, Prévost	2017-05-24	2016-12-16	NA	NA	inconnu	non	2016	OTO0618	1	43	42	42	8	2016-08-26
VLR1721	58329	2018-04-06	JRYG	HDGE	1985-01-15	SYT8211	16	M	KCU6813	VQ	PX	2018-03-01	Service de police de Terrebonne,	2016-10-26	2016-07-21	NA	NA	connu	non	2016	PBX0810	1	43	42	42	8	2016-07-15
QYC7435	56389	2017-08-19	NA	NA	NA	NA	17	M	KOC0546	AD	NA	NA	SPVM, Centre opérationnel Est,	2017-08-12	2017-03-21	NA	NA	inconnu	non	2017	LLN2592	1	43	42	42	8	2017-03-21
LCX0892	56389	2017-08-19	NA	NA	NA	NA	17	M	KUQ1231	SR	8/7!	NA	SPVM, Centre opérationnel Est,	2017-03-21	2017-01-15	NA	NA	inconnu	non	2017	VXJ9049	2	43	42	42	8	2017-01-15
GDH0303	56892	2017-05-30	LIPX	NHNG	1992-02-15	PCT3248	16	M	VEW7879	AA	PX	2017-12-22	SQ, MRC de La Rivière-du-Nord, Prévost	2017-05-24	2016-12-16	NA	NA	connu	non	2016	OTO0618	1	43	42	42	8	2016-08-26
FKS0988	56892	2017-05-30	LIPX	NHNG	1992-02-15	PCT3248	16	M	VOW1553	VQ	PX	2017-12-22	Service de police de Terrebonne,	2016-10-26	2016-07-21	NA	NA	connu	non	2016	PBX0810	1	43	42	42	8	2016-07-15
KHS6500	55227	2017-03-22	CPPE	YEDR	1999-01-15	UTU1415	17	M	KOC0546	SR	NA	NA	SPVM, Centre opérationnel Est,	2017-03-21	2017-01-15	NA	NA	connu	non	2017	VXJ9049	1	43	42	42	8	2017-01-15
BJK1388	55227	2017-05-25	CPPE	YEDR	1999-01-15	UTU1415	16	M	ZDP3350	AA	PX	2017-06-21	SQ, MRC de La Rivière-du-Nord, Prévost	2017-05-24	2016-12-16	NA	NA	connu	non	2016	OTO0618	1	43	42	42	8	2016-08-26
JAV9903	53553	2017-08-19	NA	NA	NA	NA	17	M	KUQ1231	AD	NA	NA	SPVM, Centre opérationnel Est,	2017-08-12	2017-03-21	NA	NA	inconnu	non	2017	LLN2592	1	43	42	42	8	2017-03-21
LFI2596	53553	2017-02-22	NA	NA	NA	NA	16	M	KUQ1231	IN	PX	NA	SPVM, Centre opérationnel Est,	2017-02-20	2016-08-25	NA	NA	inconnu	non	2016	ZXU8990	1	43	42	42	8	2016-08-25
BMC0409	53553	2016-09-01	NA	NA	NA	NA	16	M	BSB1667	IN	NA	NA	SPVM, Centre opérationnel Est,	2016-08-29	2016-08-03	NA	NA	inconnu	non	2016	UUE1089	1	43	42	42	8	2016-08-03
QSC9985	53553	2016-09-01	NA	NA	NA	NA	16	M	KUQ1231	SR	NA	NA	SPVM, Centre opérationnel Est,	2016-05-04	2016-02-20	NA	NA	inconnu	non	2016	RKT8419	1	43	42	42	8	2016-02-20
LCX0892	53553	2017-05-12	NA	NA	NA	NA	17	M	KUQ1231	SR	PX-2/13	NA	SPVM, Centre opérationnel Est,	2017-03-21	2017-01-15	NA	NA	inconnu	non	2017	VXJ9049	2	43	42	42	8	2017-01-15
EON3882	53421	2018-05-03	NA	NA	NA	NA	17	M	JLM9831	AA	NA	NA	Service de police de Saint-Jérôme,	2018-04-30	2017-12-01	NA	NA	inconnu	non	2017	MUH2312	1	43	42	42	8	2017-11-30
ZPV5110	53421	2016-08-15	NA	NA	NA	NA	16	M	LDE4078	IN	MA(1	NA	SQ, MRC de Mékinac, Saint-Tite	2016-08-08	2015-10-11	NA	NA	inconnu	non	2016	ZUP9494	1	43	42	42	8	2015-10-10
UKP4528	53421	2016-08-15	NA	NA	NA	NA	4	M	PYL7045	IN	PEX	NA	SQ, MRC de Vaudreuil-Soulanges Est,	2005-05-06	2004-04-19	NA	NA	inconnu	non	2004	BZD9975	1	43	42	42	8	2004-04-19

Si leur présence était décelée dans un autre dossier dans un autre composant, il y aurait forcément un amalgame de tous ces dossiers dans un même composant. Deuxièmement, l'individu « 30900 » est vu dans deux autres dossiers, il s'agit donc d'un récidiviste.

Ainsi, pour cet individu, vous verrez (en vert) son dossier de récidive en co-délinquance, le « FMN8421 » qui apparaît deux fois en lien avec deux autres individus, les « 38004 » et « 36942 ». On laissera en blanc son dossier solo, le « HDX8715 » vu une seule fois. Vous remarquerez dans la co-délinquance du dossier « FMN8421 » que l'individu « 36942 » n'est vu qu'une fois; il est donc acteur solo dans le dossier. Par contre, son complice, le « 38004 » a été détecté dans deux autres dossiers : un délit effectué en solitaire, le dossier « KFS4105 », et le dossier « WAY3578 » (en bleu) partagé aussi en co-délinquance avec les individus « 43878 » et « 43533 ». Le premier individu est récidiviste dans un dossier solo (vu une fois), le « KVH4097 », et le second, lui aussi récidiviste, dans un dossier solo, le « RTP3715 », mais en co-délinquance dans un autre dossier, le « KZW8219 » (en violet). L'enchaînement se poursuit avec ce dernier dossier dans lequel on a retrouvé l'ADN de l'individu « 26319 », un récidiviste à quatre délits dont deux en solo, les dossiers « GJE4144 » et « HDD9361 » et le dernier lien observé dans le dossier « SOF5058 » (en orange), en co-délinquance avec l'individu « 31614 » dont on a aussi trouvé l'ADN dans un dossier solo, le « XCT7321 ».

Il s'agit là d'un exercice intéressant qui montre comment des événements, échelonnés dans le temps, peuvent être en fait reliés entre eux par des individus qui pratiquent la co-délinquance[27]. Comme vous pouvez le constater, tous ces délits se sont déroulés entre le 3 février 2004 et le 12 octobre 2017 (colonne AB), soit sur un peu plus de 13 ans, à Montréal, dans sa banlieue Ouest et Nord ainsi que dans les Laurentides (voir la géolocalisation dans la colonne N). Peut-on imaginer que l'ensemble de ces informations soit présent à l'esprit de certains enquêteurs chevronnés afin qu'ils puissent focaliser leur attention sur l'inconnu « 43878 » présent dans le réseau 42? Poser la question c'est, je crois, y répondre.

---

[27] À vos crayons, je vous invite à retrouver les liens présents dans le composant 43 avec les données de la deuxième moitié du Tableau 1.

Si vous vous faites l'exercice de dessiner les interrelations observées dans le composant 42, vous devriez obtenir un sociogramme comme celui de la Figure 1 (à gauche). Dans le deuxième chapitre, vous trouverez ce même composant dans sa forme unimodale (c.-à-d. : montrant uniquement les individus) dans la Figure 2B, à la position 44. Vous pouvez comparer cette structure avec celle présentée à droite dans la Figure 1.

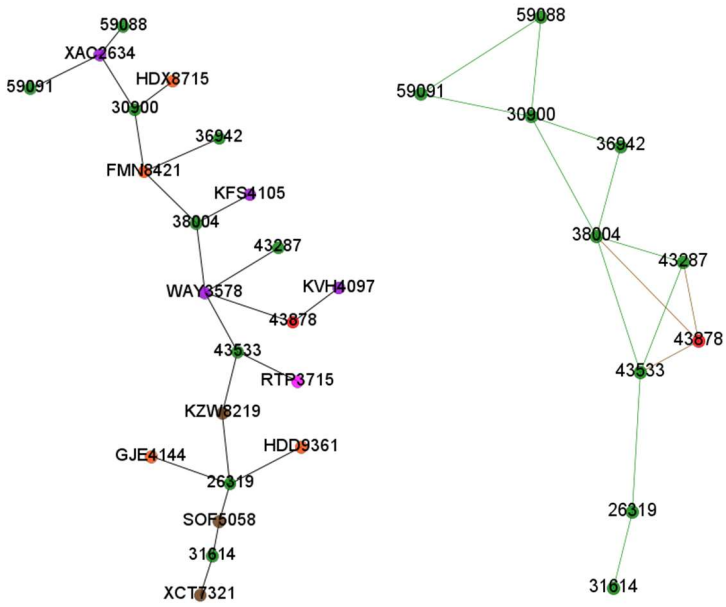


Figure 1 : Représentation du composant 42 en sociogramme bimodal à gauche et unimodal à droite, avec les libellés de nœuds. Les connus en vert, les inconnus en rouge, les autres couleurs correspondant à divers types de délits. Les sociogrammes ont été constitués à partir de données extraites du tableau 1 (colonne B et V) et traités par le logiciel Gephi (Bastian *et al*. 2009).

3. Gestion, vérification et correction des données

Dans la présente section, nous passerons en revue chacun des 28 types de données auxquels, outre leur description, j'ajouterai le type de corrections qui ont parfois été nécessaires et comment ces données ont été utilisées dans le processus d'analyse en réseau ou pour certains aspects criminologiques.

Lors de la création du fichier codifié, les transformations effectuées à partir des données d'origine conservées au LSJML génèrent des valeurs qui sont précédées d'un « = » et de «"» qu'il faudra retirer avant de poursuivre le travail de vérification qui sera

fait dans sur une version sauvegardée en format .xls. Il faut mentionner que plusieurs de ces données sont saisies manuellement et que, invariablement, quelques coquilles peuvent s'y glisser. Il faut donc effectuer une vérification complète des données pour s'assurer de corriger ou d'enlever certaines données aberrantes qui peuvent fausser certains calculs ou bloquer des tris ou filtrations lors des analyses. On retrouve facilement, dans chaque catégorie, les divers formats (attendus et aberrants) de données en procédant à des filtrations. On y remarquera les inscriptions « NA » qui dénotent l'absence d'informations. Une mention est inscrite dans les paragraphes appropriés lorsqu'on observe des informations comportant des données aberrantes et lorsque les corrections ont été apportées, s'il a été possible de la faire.

La colonne Numéro_Dossier (codé) (colonne A)

Ce numéro correspond à l'identifiant unique créé pour nommer l'échantillon mis en banque. Lors de la vérification, on constate que les 24 641 lignes sont constituées de 24 254 numéros uniques, 192 doublons et 1 trio. Les échantillons qui ne sont pas uniques correspondent à ceux qui sont constitués d'une combinaison complète de deux ou trois personnes (un seul cas de cette espèce) et qui ont été mis en banque comme tels. Ce faisant, le « même » échantillon (lire le même numéro), s'il y a concordance, se retrouvera possiblement dans deux fiches. À titre d'exemple, voyez le «numéro de_dossier» « LCX0892 » du composant 43 qui est présent deux fois, permettant l'identification des individus « 55869 » et « 53553 » (colonne B) dans le même prélèvement « QUQ1231 » (colonne J) en provenance du même dossier « VXJ9049 » (colonne V). Ce sont uniquement ces cas de combinaisons parfaites de deux individus qui peuvent être détectés dans un même prélèvement, qui peuvent être à la source d'un doublons de «numéro de dossier». Ces cas particuliers sont notés avec la valeur « 2 » dans la colonne W qui indique le nombre d'individus pour les combinaisons complètes. Cette valeur n'est d'aucune utilité pour construire les composants du réseau, mais, comme vous le verrez ci-dessous, cette dernière grâce à sa logique d'unicité, et à la présence des combinaisons décrites dans la colonne W, elle a permis de détecter des coquilles présentes dans les numéros de la colonne « ID de Cas ».

Dans les données globales, les valeurs de « 2 » de la colonne W qui indiquent les combinaisons complètes sont au nombre de 188. Il y a donc un écart de quatre doublons (192 – 188 = 4) entre les valeurs de combinaisons et les «numéro de dossier». Une recherche parmi les numéros de la colonne « ID de Cas » présentant des valeurs différentes pour le même «numéro de dossier» a permis d'identifier ces numéros qui sont « NDE5418 », « IZE0094 », « NDE5203 » et « HLL7400 ». Les huit « ID de Cas »associés à ces quatre «numéro_dossier» sont « DAT5739 », « AIN8266 », « PML6672 », « PFK6408 », « SCK6023 », « AQM0433 », « YAY1372 » et « MNP3731 ». Il s'agit donc de quatre paires de dossiers (cas) différents qui ont chacune le même «numéro de dossier», pour des prélèvements différents. Une erreur de nomenclature s'est probablement glissée dans ces cas d'exceptions pour que le même «numéro de dossier» soit attribué aux échantillons différents mis en banque pour des dossiers différents. Pour rectifier la situation et donner un «numéro de dossier» différent aux « ID de Cas »différents, il a été décidé, pour une série de quatre « ID de Cas », de remplacer par un « A » la première lettre de leur «numéro de dossier». Vérification est faite que ces nouveaux «numéro de dossier» sont uniques dans la colonne « A ». Ainsi, comme on le voit dans le tableau 2, une série de quatre « ID de Cas »conserve le «numéro de dossier» d'origine (lignes du haut) et l'autre groupe de quatre « ID de Cas » hérite du même numéro mais que l'on a modifié en y ajoutant un « A » initial (lignes du bas).

Tableau 2 : Résumé des corrections de quatre « numéro de dossier » en fonction de « ID de Cas » différents.

observé		corrigé	
«Numéro_dossier»	« ID_Cas »	«Numéro_dossier»	« ID_Cas »
NDE5418	DAT5739	NDE5418	DAT5739
IZE0094	AIN8266	IZE0094	AIN8266
NDE5203	PML6672	NDE5203	PML6672
HLL7400	PFK6408	HLL7400	PFK6408
NDE5418	SCK6023	ADE5418	SCK6023
IZE0094	AQM0433	AZE0094	AQM0433
NDE5203	YAY1372	ADE5203	YAY1372
HLL7400	MNP3731	ALL7400	MNP3731

Le «numéro de_dossier» réel contient un amalgame de plusieurs informations qui seront séparées par le script de recherche des composants (voir plus loin).

La colonne Numéro_Fiche (codé) (colonne B)

Le numéro de fiche est l'identifiant d'un profil génétique unique[28] et par extension d'un individu. On comprendra sans peine qu'il s'agit d'un élément central pour la construction des composants. C'est pour cette raison que le numéro de fiche est la seule donnée codée qui reste constante dans les divers envois de données en provenance du LSJML permettant du coup de retracer les individus dans des composants qui, eux, peuvent subir des modifications au fil du temps par l'ajout de dossiers ou d'autres individus.

[28] L'utilisation du terme unique se veut ici une simplification pour apporter l'idée que le profil génétique observé provient d'une personne. La notion de rareté est fonction de la valeur probante d'un profil génétique et une fiche est créée lorsque cette valeur probante est suffisante pour permettre la comparaison à l'aveugle de profils mis en banque, en provenance d'autres délits ou du fichier de condamnés.

La répétition d'un numéro de fiche représente le nombre de récidive d'un individu. Elle fait directement état du compte des délits d'un individu. Les contrevenants identifiés par la BNDG dans un seul dossier apparaissent une seule fois dans la liste de la colonne « B ».

Dans cette colonne, les jumeaux ont un descriptif qui commence par un « J ». Ils sont au nombre de 16 (vus sur 36 « ID de Cas »). Il a été parfois observé, en fonction du «numéro de dossier» associé aux échantillons provenant de jumeaux, que le script de détection des composants détecte les jumeaux comme des mélanges de deux individus, leur attribuant une valeur de « 2 » dans le compte des combinaisons de la colonne « W ». Une correction manuelle est faite pour ramener ces valeurs à « 1 » pour les jumeaux.

La colonne Date_inscription_fiche (réelle) (colonne C)

Cette date correspond à la date de création de la fiche. Elle n'est d'aucune utilité pour la construction des composants, mais a été utilisée pour remplacer (par rapprochement) dix-huit dates d'identifications manquantes de la colonne « M » alors que « oui » était inscrit la colonne « ID de labo » (voir Tableau 3). Cette valeur peut aussi, en cas de problème, aider à vérifier la cohérence d'autres dates, puisqu'elle doit être postérieure à la date de mise en banque du profil en provenance du plus vieil événement présent dans la fiche.

Tableau 3 : Ajout manuel de « Date.ID.Nationnal » par l'utilisation de la « Date_inscription_fiche » pour 18 fiches.

No. Fiche	Date de « C » recopiée vers « M »	No. Fiche	Date de « C » recopiée vers « M ».
31497	2009-04-24	42057	2013-04-11
43899	2013-11-06	40704	2012-10-25
55206	2017-03-22	34752	2010-10-21
21090	2010-08-21	39681	2012-07-09
35802	2010-12-16	23160	2005-02-02
26730	2013-03-19	40287	2012-09-11
43038	2013-08-07	23688	2005-04-14
25158	2006-04-19	60654	2018-12-13
57672	2018-02-13	59772	2018-09-13

La colonne Prénom et Nom (codés) (colonnes D et E)

Cette information n'est d'aucune utilité pour la construction des composants. Leurs présences doivent toutefois être associées à des individus connus ayant une date d'identification au National (colonne M) ou une « IDLabo » marquée comme « oui » (colonne T).

La colonne Date_Naissance (réelle partielle) (colonne F)

Il s'agit d'une donnée qui n'a été utilisée que de manière très exploratoire, puisqu'elle est absente pour les sujets centraux de notre étude, c'est-à-dire les inconnus. Mis à part quelques exceptions qui étaient complètes, le LSJML a fourni une date «aaaa-mm » sans préciser la journée pour éviter des possibilités d'identification. Il a été décidé d'arrondir manuellement au 15 du mois ces dates de naissance, comme on peut le voir dans l'exemple du tableau 1.

Certaines aberrations sont observées dans le format qui se présente parfois en aa-mm-jj au lieu de aaaa-mm-jj, ou parfois dans la valeur de l'année qui était incohérente.

Des corrections ont été effectuées : par exemple une valeur 62-06-19 a été réécrite sous le format 1962-06-19, et des valeurs d'années comme 2018 et 2042 qui devaient en fait être 1918 et 1942 ont été corrigées.

La colonne FPS (codé) (colonne G)

Cette information n'est d'aucune utilité pour la construction des composants. Il s'agit du numéro unique d'identification des empreintes digitales d'un individu enregistré dans les dossiers de la GRC et au fichier des condamnés. On voit que cette donnée est associée aux individus connus, et qu'elle est absente pour les inconnus.

Les colonnes

Année_abr (réelle) (colonne H),
Ville (réelle) (colonne I),
Numéro_Relatif_Prélèvement (codé) (colonne J)

Il s'agit ici de trois valeurs inscrites par le script de détection des composants qui sont obtenues de la décomposition du «Numéro de Dossier » afin d'isoler le « ID de cas » (voir plus bas). Cette information n'est d'aucune utilité pour la construction des composants.

Toutefois, la valeur indiquée dans la colonne « M » qui représente les données fournies par le LSJML, pourrait un jour s'avérer inutile, car elle permettrait de faire les distinctions entre les villes canadienne si les données de concordances étaient un jour intégrées à l'échelle nationale.

La colonne Type_de_délit (réelle) (colonne K)

Il s'agit ici d'une liste de huit abréviations décrivant les types de délits effectués par les individus. On en trouvera une description dans la section « data » dans les articles soumis.

À la suite de saisies erronées, des incohérences peuvent se glisser dans cette liste d'informations. En effet, nous avons rencontré des situations où dans un même dossier (« ID de Cas »identique) lié à deux individus on trouve deux types de délits différents. Dans la réalité, il n'est pas impossible que deux individus ayant participé à une activité criminelle soient accusés différemment. Ce genre de situation rare est toutefois incompatible avec une approche de montage en réseau, car les informations liées au nœud du délit sont différentes pour deux individus, et elle est facile à détecter puisqu'un message d'erreur apparaît dans le logiciel Gephi au moment de l'importation des données. Une telle observation est plus facile à corriger qu'une vérification systématique de toutes les informations sur les co-délinquances. Le tableau 4 donne la liste des trois situations où une erreur dans le type de délit a été corrigée. Évidemment, pour tenter de déceler la source de l'erreur et obtenir le véritable délit, on s'en remettra au laboratoire ou au service de police, si on travaille en appui à une enquête, ou bien on pourrait arbitrairement choisir le type de délit le moins important (statistiquement plus fréquent) si on travaille à des évaluations globales. Ce type d'erreur est dans les faits si rare, comme on l'a vu, qu'il n'y aurait pas de conséquences importantes dans une perspective de statistique générale.

Tableau 4; Trois groupes de deux individus en co-délinquance pour lesquels la description du délit est incohérente. (ND : non désigné, HO : homicide, VQ : vol qualifié)

Individu (fiche)	Dossier (ID de Cas	Délit observé	Délit corrigé
40368	BMV9816	ND	ND
40389	BMV9816	HO	ND
24123	ZBO4580	AA	VQ
25377	ZBO4580	VQ	VQ
43662	CXG5006	AA	VQ
41610	CXG5006	VQ	VQ

La colonne code_suffixe (réelle) (colonne L)

Ici aussi, il s'agit d'une valeur inscrite par le script de détection des composants qui sont obtenus de la décomposition du «Numéro de Dossier » afin d'isoler le « ID de cas » (voir plus bas).

Ces données sont généralement associées aux profils génétiques complexes et de combinaisons de profils où parfois il reste des éléments supplémentaires qui n'ont pu être soustraits du profil principal. Par exemple, les dénominations PEX ou PX indiquent que le profil est déposé en banque comme source unique, mais extrait d'une combinaison. S'il est suivi d'une série de chiffres comme « -5/9, -2/8, -0/12, -4/8, etc.», il s'agira d'un profil partiellement extrait où, dans le premier exemple, 5 loci apparaissent en banque avec un ou deux allèles (fragment ADN) et 9 loci avec trois allèles et plus, et ainsi de suite pour les autres exemples. Si un « (x » (ou x = 1 à 11) suit cette valeur, il s'agit alors d'un profil mis en banque avec x allèles obligatoires qui devront être présents si on veut que le logiciel de comparaison accepte la concordance. Si les suffixes sont MI ou MA, il s'agit encore de profils extraits de combinaisons, mais cette fois-ci dans une forme « Mineur ou Majeur ». Les points d'exclamation ! » quant à eux représentent des profils extraits de manière tentative aux fins de comparaison avec le fichier national, mais indiquent qu'une version plus complexe de la combinaison est conservée au niveau local pour comparaison avec l'ensemble de ses allèles.

Ces données n'ont pas été utilisées dans le cadre de la présente thèse, mais pourraient servir à qui voudrait mettre l'accent sur les résultats que l'on peut obtenir à partir de profils complexes. Au total, il y a 2 160 concordances qui possèdent un profil génétique avec un suffixe.

Toutefois, pour qui voudrait s'y attarder, il faut prendre note que quelques données aberrantes y ont été observées. Le tableau 5 en donne la liste et les corrections effectuées. De plus, les données d'origine obtenues du laboratoire sont marquées d'une apostrophe « ‘ » devant de nombreuses valeurs. Ce caractère typographique doit être enlevé et toute la colonne sera alors mise au format texte.

Tableau 5 : Correction pour les valeurs suffixe observées comme aberrantes, auxquelles s'ajoutent 89 valeurs « PEXSTA » changées en « PEX ».

Fiche	Valeur observée	Valeur corrigée	Fiche	Valeur observée	Valeur corrigée
20919	BIPEX	PEX	61359	FU	NA
21096	PEXMAJ	MA	21723	PEXMAJ	MA
23538	PEXMJ	MA	35154	C	NA
27816	3/10V	3/10	35937	8*5	8/5
42747	ESTA	NA	44007	AMI-11/4	MI-11/4
42219	BPX	PX	20781	8/5PEX	PEX-8/5
32703	11/2V	11/2	27594	11/4V	11/4
35682	PX*1	PX(1	45423	PEX1	PEX(1
39015	PX-11*4*4	PX-11/4(4	31116	RM8STA	NA
36021	APX	PX	24939	/11	2/13
26625	PX9/4V	PX9/4			

La colonne Date.ID.Nationnal (réelle) (colonne M)

Il s'agit de la date à laquelle la BNDG a mis au jour une concordance entre le profil génétique du dossier criminel et un individu inscrit au fichier des condamnés. Ces dates ne sont présentes que pour les individus connus (identité, colonne S), et elles ont le format aaaa-mm-jj.

Chez les individus connus, cette date peut servir d'un moment charnière qui permet de distinguer leurs activités criminelles inscrites avant cette date alors qu'ils étaient inconnus, des activités inscrites après, pour ces individus maintenant connus.

La colonne Géolocalisation (réelle) (colonne N)

Il s'agit de la dénomination du service de police qui a enregistré le délit. La localisation des délits est un élément important pour des études de criminologie ou pour

l'avancement d'enquêtes complexes, mais dans tous les cas, il est préférable d'avoir accès à la localisation précise GPS du lieu du délit. Ici la localisation par poste de police est beaucoup plus grossière et moins utile. De plus, dans les présentes données les descriptions d'un même poste de police sont souvent variables suite à diverses saisies de données. Pour toutes ces raisons, la géolocalisation n'a pas été intégrée aux analyses de la présente thèse.

La colonne Date_mise_en_banque (réelle) (colonne O)

Cette information au format aaaa-mm-jj n'est d'aucune utilité pour la construction des composants. Cette date peut à tout le moins aider à découvrir une certaine irrégularité ou incohérence sur d'autres dates, car on sait que la date de mise en banque se situe pour la plupart des cas entre quelques mois et un an après l'événement. La vérification démontre qu'elles sont toutes ≥ 2000.

La colonne Date_événement_labo (réelle) (colonne P)

Il s'agit de la date enregistrée dans les informations centrales du LSJML, relative à un délit. Cette date a servi à compléter les dates absentes au dossier de police (colonne AB). Une valeur aberrante de 782 a été remplacée par un NA.

La colonne Match_inter_provincial (réelle) (colonne Q)

Il s'agit d'une information qui précise si l'individu identifié au LSJML a aussi été détecté dans un autre laboratoire du Canada. On y décrit, par une lettre, le labo d'enregistrement de la concordance. (H : pour Halifax, O : pour Ottawa, T : pour Toronto, E : pour Edmonton et V : pour Vancouver)

Ces informations donnent une idée intéressante des individus qui sont actifs sur un territoire plus élargi, hors Québec. Compte tenu des dates et des types de délit, il y a là des informations intéressantes qui sont prêtes à être éventuellement complétées par les autres délits des autres provinces. Il est à noter que ces informations sont parfois associées à des inconnus qui doivent être présents dans un minimum deux dossiers pour

puissent être inclus dans les fiches de concordances. Les individus inconnus qui n'ont qu'un seul dossier au Québec, mais qui sont présents dans les autres provinces, ont une fiche de suivi puisque dans les faits une ou des concordances existent. Cette particularité est expliquée dans la section « data » du premier article.

La colonne Date Décès (réelle) (colonne R)

Il s'agit d'une autre information intéressante qui pourrait être développée mais qui ne touchait pas notre champ de recherche concernant les inconnus, puisqu'à partir du moment où un individu décède on arrive généralement à l'identifier. D'ailleurs pour contourner cette situation et conserver les données de concordances et d'activités de ces individus, leur statut de décédé (dcd) a été changé à « connu » lors de nos analyses. Ces dates ont un format aaaa-mm-jj et aucun individu autre que les « dcd » n'ont une date de décès. (Heureusement).

La colonne identité (réelle) (colonne S)

Ici, on a trois descriptions désignant si un individu est connu, inconnu ou décédé (dcd dans le fichier). Premièrement, un compte d'inconnus ≥1 dans les composants (en utilisant le « Component_ID » plus loin) permet de filtrer les composants avec inconnu(s) des composants qui n'en ont pas (valeur de 0). Deuxièmement, les inconnus qui figurent dans cette colonne et sont présents dans un seul dossier au LSJML, présentent un ou des dossiers dans d'autres provinces (voir sous l'item « Match_inter_provincial » colonne Q). Et finalement, certains inconnus présentaient une information « ID de labo » marquée « oui ». Cette situation de saisie erronée doit être vérifiée auprès du laboratoire pour une mise à jour.

Dans le cas de certains inconnus on a une « Date.ID.National », mais aucune information sur leur identité (nom, prénom, FPS). Une vérification doit être faite auprès du laboratoire pour s'assurer qu'il s'agit bien de jeunes contrevenants pour lesquels la GRC ne peut libérer ces informations. De tous ceux qui ont été observés, il en est resté

14 sans précisions en ce qui a trait à leur statut réel. Ces 14 « Numéro_Fiche » sont les suivants : 25926, 41154, 55329, 60870, 40479, 43761, 56688, 60903, 40122, 44490, 57501, 40368, 45240 et 59271.

La colonne IDLabo (réelle) (colonne T)

Dans cette colonne, on indique si l'identité de l'individu a été obtenue d'une analyse d'un mandat ADN délivré dans le cadre d'une enquête, ou si elle a été obtenue de la BNDG dont il a été fait mention ci-dessus. On y retrouve tout simplement des « oui » et « non ».

On a observé quatre individus ayant une « ID-Labo » inscrite comme « oui » mais sans date de naissance. Il s'agit des fiches 23160, 29313, 32922 et 53193. Aucune action n'a été entreprise aux fins de correction.

La colonne Année_complète (réelle) (colonne U)

Il s'agit du pendant de l'année abrégée inscrite dans la colonne H. Ces informations peuvent aider à corriger un résultat aberrant observé dans la colonne AB intitulée « Date_Evénement_Police ».

La colonne ID_Cas (codé) (colonne V)

Ce numéro anonyme représente le numéro de dossier qui prend ici le nom de « cas ». Ces numéros sont uniques, et leur répétition démontre une Co-délinquance, qui s'observe quand un minimum de deux profils distincts ou une combinaison de deux personnes sont associés au même « ID_Cas ».

La co-délinquance peut être vue de manière globale en comptant le nombre de fois qu'une « ID_Cas » est observée. On obtient ainsi le nombre d'individus liés ou actifs dans ce délit. En faisant le même calcul, mais en y soustrayant une valeur « 1 » on obtient le nombre de complices actifs en compagnie d'un individu particulier. Un des aspects intéressants de ce calcul de complice est qu'une valeur de « 0 » est attribuée aux délits effectués en solo. Ainsi, en soustrayant ces délits solos du nombre de dossiers en

récidive attribuables à un individu (répétition du « Numéro_fiche »), on obtient le nombre de dossiers effectués en co-délinquance.

Une correction a été effectuée sur les composants 331 et 612 où une « ID_Cas » était dédoublée sans mention de combinaison (point suivant). L'observation a été faite grâce à un calcul de co-délinquance, dont le résultat démontre que ces individus avaient tous des délits solos et un délit sur une même « ID_Cas » tout en étant dans dans des composants différents. Les deux lignes d'informations en double ont été éliminées.

La colonne Nombre_Individus_par_Prélèvement (réelle) (colonne W)

Ici, une valeur de « 2 » indique la présence d'une combinaison complète de deux individus. Il y a 188 mélanges d'une telle nature dans les données et un trio. Des corrections ont été effectuées et elles sont décrites dans les sections « Numéro de Dossier » et «Numéro de Fiche » à propos des jumeaux.

La colonne Component_ID (réelle) (colonne X)

Il s'agit d'un nombre de 1 à 11 910 désignant les divers composants observés. C'est une valeur intéressante, inscrite par le script R de recherche des composants, qui permet de suivre les divers composants analysés en y joignant le nombre d'individus observés (voir plus bas). Cette valeur est évaluée en fonction de l'ordre et de la taille du réseau (deux points suivants).

La colonne Ordre_Réseau (réelle) (colonne Y)

Il s'agit d'une valeur de description technique inscrite par le script R de recherche des composants qui va de 3 à 155.

La colonne Taille_Réseau (réelle) (colonne Z)

Il s'agit d'une troisième valeur obtenue du script de recherche des composants, qui correspond à deux fois le nombre de prélèvements observés dans un composant. Les deux composants en exemple dans le tableau 1 ont tous deux 21 prélèvements (colonne J) et par conséquent une « Taille_Réseau » de 42. Ces valeurs vont de 2 à 156. S'il est

nécessaire dans certaines études de mettre en corrélation des éléments en fonction du nombre d'échantillons traités dans les analyses liées à un composant, la « Taille_Réseau » serait un paramètre de choix.

Le tableau 6 présente les résultats de corrections effectuées sur 9 composants pour lesquels la « Taille_réseau » n'était pas le double du nombre de prélèvements. Aucune investigation du script n'a été tentée pour expliquer cette différence par rapport à l'ensemble des autres résultats.

Tableau 6 : Correction de la « Taille Réseau » de 9 composants

Composant	Taille_Réseau lue	Nombre de prélèvements	Taille_Réseau corrigée
38	52	18	36
45	42	19	38
612	14	4	8
866	10	3	6
1159	8	3	6
1727	6	2	4
1907	6	2	4
2011	6	2	4
3435	4	1	2

Cette correction permet d'avoir des données cohérentes sur la « Taille_Réseau » pour qui voudrait les utiliser.

La colonne Nombre_Individu_Réseau (réelle) (colonne AA)

Il s'agit ici d'une autre valeur obtenue du script de recherche des composants qui donne le nombre total d'individus dans le composant, valeur importante qui décrit bien, sous un autre paramètre, la taille d'un composant. C'est une valeur que nous avons utilisée régulièrement dans nos travaux, soit pour classer les composants en ordre soit pour corréler avec d'autres paramètres. Les valeurs observées vont de 1 à 37.

Vérification est faite que les jumeaux ne sont pas comptés deux fois dans les composants où ils sont présents.

La colonne Date_Evénement_Police (réelle) (colonne AB)

Il s'agit de la date de l'événement telle qu'elle a été consignée dans l'un des répertoires du dossier d'analyse. Elle est indiquée dans un format aaaa-mm-jj. Certaines données sont totalement absentes et, dans 22 cas, seule l'année était inscrite. Des corrections ont été effectuées en ajoutant un -06-06 aux dates. Pour les dates totalement absentes, nous avons utilisé l'année inscrite dans la colonne U, en s'assurant qu'elle soit cohérente avec la date de mise en banque (colonne O) et possiblement avec la « Date.ID.National » (colonne M).