Version pre-print de l'article paru dans *Réseaux*, n°214-215.

Comment se forment les publics d'une carte de crimes ?

Une analyse computationnelle de traces textuelles[1]

Sylvain Parasie (Université Paris Est Marne-la-Vallée, LISIS)
sylvain.parasie@univ-paris-est.fr

&

Jean-Philippe Cointet (médialab, Sciences Po)
jeanphilippe.cointet@sciences-po.fr

[1] Nous remercions Benjamin Ooghe, qui a participé à l'élaboration de la base de données que nous exploitons dans ce texte. Nous remercions également Madeleine Akrich, Valérie Beaudouin, Bilel Benbouzid, Jean-Samuel Beuscart, Dominique Cardon, Alexandre Mallard et Kevin Mellet pour leurs conseils sur des versions antérieures de cet article.

Résumé

Dans cet article, nous nous demandons dans quelle mesure un public au sens fort du terme peut se former autour des occurrences diffusées par les plateformes journalistiques en ligne – c'est-à-dire pas seulement un ensemble d'individus qui consomment des informations, mais un être collectif qui partage des interprétations communes. Pour répondre à cette question, nous avons analysé un corpus de 28 828 commentaires postés sur « The Homicide Report », une plateforme apparue en 2010 qui fournit une information standardisée sur tous les homicides commis à Los Angeles. Nous avons eu recours à une méthode d'analyse textuelle très peu utilisée en sciences sociales, qui repose sur des algorithmes d'apprentissage supervisé. Cette enquête conduit à deux résultats. D'une part, nous montrons que les internautes parviennent à élaborer des interprétations communes à partir des occurrences qui leur sont adressées, en combinant trois façons de « faire public » qui empruntent aux médias traditionnels. D'autre part, nous montrons que l'exploitation sociologique des inscriptions textuelles permet de réduire le fossé entre les enquêtes quantitatives sur les audiences – qui se situent à grande échelle mais échouent à saisir des interprétations – et les études plus qualitatives – qui saisissent des interprétations à une échelle très locale.

Mots-clés

Publics de l'information ; journalisme ; faits divers ; analyse textuelle ; apprentissage supervisé.

Avec l'essor des dispositifs d'information en ligne, nous accédons de plus en plus à des « occurrences », c'est-à-dire à des informations qui concernent des faits ou des incidents souvent uniques, et qui sont attachés à des espaces, des temps et des acteurs étroitement circonscrits (cf. Molotch et Lester, 1996). Un crime est commis sur une personne tel jour dans tel quartier ; un niveau de pollution de l'air est mesuré à tel endroit sur une période de temps donné ; telle école obtient tel taux de réussite à un examen, etc. Toutes ces occurrences sont diffusées sous la forme d'articles, de cartes qui représentent leur répartition sur un territoire donné, ou encore sous la forme de classements ou de graphiques. Bon nombre d'applications mobiles ou de sites web permettent ainsi aux individus de prendre connaissance de ces informations factuelles, circonscrites dans le temps et dans l'espace, et dont ils prennent connaissance sur la base de leurs intérêts personnels.

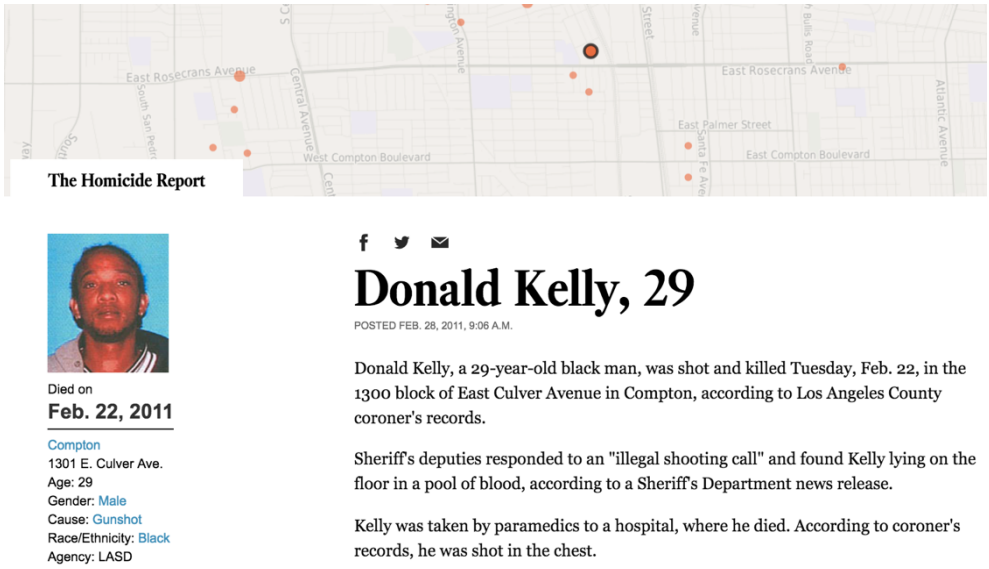
Le public de ces dispositifs d'information nourrit de vives interrogations. Essayistes et chercheurs se sont inquiétés du passage d'une configuration dans laquelle les individus n'accédaient qu'à un petit nombre d'occurrences soigneusement sélectionnées par les organisations médiatiques, à une nouvelle configuration dans laquelle ils ont de plus en plus la possibilité de sélectionner eux-mêmes parmi un grand nombre d'occurrences sur la base de leurs intérêts personnels. Dans cette nouvelle configuration, les individus se contenteraient de prendre connaissance des occurrences qui les touchent personnellement – en tant que riverains, parents ou usagers des services collectifs – et se détourneraient des informations ayant trait aux affaires publiques (Sunstein, 2002, 2018 ; Prior, 2007). Si bien qu'ils ne seraient plus en mesure de former un « public » au sens fort du terme – c'est-à-dire un collectif d'individus qui partagent des interprétations tout en étant physiquement à distance les uns des autres (Tarde, 1901).

On sait toutefois peu de choses sur la façon dont les d'individus s'agrègent autour de ces informations dès lors que celles-ci sont diffusées sous la forme d'occurrences. Un tel projet est loin d'être évident, surtout quand on sait à quel point l'étude empirique de la formation des publics a toujours posé des difficultés aux sociologues (Quéré, 2003). Récemment, de nombreux chercheurs ont tenté de mesurer la fragmentation du public en ligne, en enquêtant quantitativement sur la manière dont les individus se rassemblent autour de certains contenus d'information selon que ceux-ci s'alignent ou non sur leurs préférences idéologiques (Bakshy et al., 2015; Flaxman et al., 2016; Fletcher & Nielsen, 2017 notamment). Mais ces nombreux travaux présentent deux limites majeures pour qui souhaiter enquêter empiriquement sur la façon dont les individus s'agrègent autour d'un grand nombre d'occurrences accessibles à travers des dispositifs d'information en ligne. En premier lieu, ces travaux analysent des phénomènes de polarisation

idéologique, considérant uniquement les contenus d'information du point de vue de leur orientation politique. De façon assez paradoxale, peu de chercheurs en sciences sociales se sont intéressés au public des informations qui correspondent le plus aux normes de l'objectivité journalistique – des informations factuelles, standardisées et dont la production engage moins la subjectivité journalistique. En second lieu, la grande majorité de ces recherches exploitent des données d'audience, ce qui ne leur permet pas d'accéder à des interprétations. On sait en effet depuis les études de réception, que la consultation d'un même contenu peut conduire les récepteurs à produire des interprétations totalement opposées (Hall, 1994). Dans cet article, nous nous demandons donc dans quelle mesure un public au sens fort du terme peut-il se former autour d'occurrences – c'est-à-dire pas seulement un ensemble d'individus qui consomment des informations, mais un être collectif qui partage des sujets de conversation et des interprétations communes.

Notre analyse porte sur « The Homicide Report », une plateforme apparue en 2010 sur le site du *Los Angeles Times*, et qui fournit une information standardisée sur tous les homicides commis dans la métropole californienne. Plutôt que de couvrir un petit nombre d'homicides qu'ils jugeaient intéressants d'un point de vue éditorial, les journalistes ont décidé de ne plus faire aucune sélection, et de couvrir tous les homicides d'une façon factuelle et standardisée. Pour l'internaute, cette plateforme se présente sous la forme d'une carte, sur laquelle apparaissent une myriade de points signalant l'endroit précis où un homicide a eu lieu. Tous ces homicides sont présentés à travers un ensemble d'informations standardisées liées à la victime (nom, âge, genre, ethnicité) et au crime (date, adresse, lieu, causes, circonstances). À chaque homicide correspond une page web, sur laquelle on trouve une photo de la victime ainsi qu'un court article rédigé de façon automatique à partir de ces informations structurées. La capture d'écran ci-dessous représente le meurtre de Donald Kelly par un point rouge cerclé de noir sur une carte et un ensemble d'informations standardisées (figure 1). Ce jeune homme noir de 29 ans a été assassiné par arme à feu, dans le quartier de Compton, le 28 février 2011. Chaque année, ce sont en moyenne 750 occurrences de ce type qui sont ainsi diffusées via *The Homicide Report*, et qui rencontrent une audience importante, comme en témoigne les dizaines de milliers de commentaires postés sur la plateforme depuis sa création.

Fig. 1 – Une occurrence parmi tant d'autres : le meurtre de Donald Kelly
(Source : http://homicide.latimes.com/)



The Homicide Report présente à nos yeux une valeur expérimentale pour étudier la formation des publics autour de ces nouveaux dispositifs d'information. D'abord parce qu'il fournit un grand nombre d'occurrences, qui n'ont pas été sélectionnées pour leur valeur journalistique, et que les internautes peuvent consulter séparément les unes des autres ou mettre en équivalence au moyen de cartes ou de listes générées à partir de critères factuels. Ensuite parce que chaque occurrence peut être commentée ou discutée sur la plateforme elle-même, le sujet de la violence urbaine étant particulièrement vif aux États-Unis. Nous avons donc procédé à une analyse quantitative des commentaires publiés sur la plateforme sur une période de sept ans, en mobilisant une méthode d'analyse textuelle qui est aujourd'hui très rarement utilisée en sciences sociales. Celle-ci s'appuie sur une technique de classification du texte dont la mise en œuvre repose sur des algorithmes d'apprentissage supervisé. À partir de la plateforme du *Los Angeles Times*, nous avons réuni un corpus de 28 828 commentaires d'internautes associés à 4 506 homicides commis entre février 2010 et décembre 2016.

Dans cet article, nous montrons que les internautes parviennent à élaborer des interprétations communes à partir des occurrences qui leur sont adressées. Suivant une approche de sociologie pragmatique, nous envisageons ici le public non comme une réalité stable et figée, mais plutôt

comme le processus par lequel des internautes en viennent à partager des interprétations (Céfaï et Pasquier, 2003). Nous montrons que ce processus implique plusieurs façons de « faire public », qui ne peuvent être saisies par le sociologue qu'à travers la mise au point de nouvelles méthodes.

L'article s'organise de la façon suivante. Dans une première partie, nous isolons, à partir de la littérature, plusieurs façons de constituer un public autour d'occurrences. Dans une deuxième partie, nous présentons les choix méthodologiques que nous avons effectués pour analyser la formation d'un public par les traces numériques offertes par la plateforme du *Los Angeles Times*. Puis nous consacrons les trois dernières parties de l'article aux trois façons de « faire public » qui émergent de l'analyse des traces numériques.

1. Comment « faire public » autour d'occurrences ?

En parcourant la littérature, il est possible de distinguer trois modèles qui sont autant de façons de construire des collectifs autour d'occurrences. Ces modèles, nous les appelons le « Colisée invisible », la « multitude de riverains » et le « collectif d'enquête ». S'ils ont rarement fait l'objet d'investigations empiriques, ces modèles vont nous aider à orienter notre enquête sur la plateforme du *Los Angeles Times*.

1.1 Le « Colisée invisible »

Le premier modèle est décrit par Gabriel Tarde au tout début du 20e siècle, dans *L'opinion et la foule* (Tarde, 1901). Évoquant la chronique judiciaire, le sociologue s'étonne que le récit d'un seul drame criminel fasse « converger pendant des semaines entières tous les regards d'innombrables spectateurs épars, immense et invisible Colisée ». À travers cette expression, Tarde souligne le pouvoir des journaux à faire en sorte que des individus en viennent à discuter d'une occurrence unique, et ce alors même qu'ils sont physiquement séparés les uns des autres et qu'ils n'ont aucun lien avec les personnes impliquées.

Le modèle que nous appelons « Colisée invisible » est lié à l'apparition des médias de masse. Il repose sur une sélection drastique des occurrences qui apparaîtront dans les journaux parmi le très grand nombre d'occurrences qui se produisent dans le monde. Le cas des journaux métropolitains aux États-Unis s'identifie bien à ce modèle. Dans les dernières décennies du 19e siècle, ces journaux cessent de couvrir le plus grand nombre possible

d'informations locales, pour ne traiter qu'un petit nombre d'occurrences en cherchant à intéresser l'ensemble des habitants de la métropole (Nord, 2001). Ce faisant, ils font émerger un public métropolitain, qui s'intéresse et converse autour de quelques occurrences souvent liées aux institutions et aux espaces centraux de la ville.

La couverture des occurrences s'accompagne alors d'un certain effacement de leurs propriétés contextuelles. Lorsqu'un crime fait la une des journaux, il n'est pas si important de savoir dans quel lieu précis il a été commis, ni quelles sont les identités précises de toutes les personnes impliquées. Plusieurs chercheurs ont ainsi montré qu'à partir des années 1960, les faits divers deviennent pour les journalistes l'occasion de parler de sujets de société, si bien que les propriétés de l'occurrence elle-même deviennent secondaires dans le traitement journalistique (Barnhurst et Mutz, 1997).

Ce modèle a souvent été vanté pour ses vertus socialisatrices, c'est-à-dire pour sa capacité à faire que des individus éloignés les uns des autres partagent des préoccupations communes. En montrant empiriquement que les citoyens partagent un petit nombre de thèmes d'intérêt, les études d'*agenda-setting* ont d'ailleurs montré que les médias parviennent à imposer les sujets de conversation des citoyens (McCombs et Shaw, 1972). Mais ce modèle a aussi été abondamment critiqué. De nombreux chercheurs de sciences sociales ont ainsi pointé le décalage entre la valeur que les journalistes attribuent à un crime et la réalité sociale du crime à une époque et dans un contexte donné. L'implication d'une célébrité ou la présence d'une circonstance étrange renforce ainsi la valeur journalistique du crime (Roshier, 1973 ; Berthaut et al., 2009). Pour les journalistes de données, le processus de sélection des occurrences s'apparente d'ailleurs à un biais qu'ils veulent de corriger en mobilisant les technologies informatiques (Parasie et Dagiral, 2013a, 2013b).

1.2 Le collectif d'enquête

On trouve chez John Dewey une autre manière de former des collectifs autour des occurrences. Dans *Le public et ses problèmes* (1927), Dewey soulignait à quel point la plupart des informations que l'on trouve dans le journal sont difficiles à interpréter et à intégrer dans le cours des événements :

> Par « nouvelles », on entend un fait qui vient juste d'arriver et qui n'est nouveau que parce qu'il dévie de ce qui est ancien et régulier. Mais la *signification* de ce fait dépend de sa relation à ce qu'il apporte et à la nature de ses conséquences sociales. Sa portée ne peut

> être déterminée que si le nouveau est placé en relation à l'ancien, à ce qui s'est passé et à ce qui a été intégré dans le cours des événements. Sans coordination, ni consécution, les événements ne sont pas des événements mais de simples occurrences, des intrusions ; un événement implique ce dont il provient. (Dewey, 1927 : 178)

Dewey jugeait alors que la majorité des informations publiées par les journaux constituent des « brèches de continuité ». Pour qu'un véritable public apparaisse, il fallait selon lui qu'un processus d'enquête se mette en place, au terme duquel sont élaborés un ensemble de jugements partagés. L'enjeu principal selon Dewey, c'est que les diverses personnes qui composent ce public mettent en œuvre une enquête collective, de façon à élaborer des jugements publics sur le problème qu'ils rencontrent (Zask, 2008). Or cette enquête porte tout particulièrement sur l'ensemble des informations parcellaires qui lui parviennent, notamment par le biais des journaux. L'enjeu, pour ce public, est alors d'enquêter sur ces occurrences, afin de les interpréter et de les mettre en série. Le travail interprétatif est ici très important, et il nécessite des ressources importantes pour que le public soit en mesure de formuler des jugements.

Ce modèle du « collectif d'enquête » diffère beaucoup du précédent. Il implique que les individus ne partagent pas seulement des sujets d'attention et de préoccupation, mais qu'ils élaborent collectivement de nouvelles interprétations à la lumière des multiples occurrences dont ils prennent connaissance. Un ensemble d'individus très différents les uns des autres en viennent à enquêter sur un grand nombre d'occurrences éclatées – qu'il s'agisse de crimes, d'accidents ou de pollutions. C'est en enquêtant sur le lien entre ces multiples occurrences, en identifiant des schèmes d'explication, que ces individus se constituent en public et élaborent des jugements partagés sur leur problème. C'est là un processus dont le terme n'est jamais définitif, et qui implique des ressources cognitives importantes.

Plusieurs traditions sociologiques se sont intéressées à cette deuxième façon de faire public. On peut penser aux études qui ont porté sur les « affaires » à travers lesquelles se constituent des collectifs qui s'indignent en interprétant des occurrences uniques ou multiples (Claverie, 1994) ; ou aux études qui analysent la façon dont des collectifs identifient des signaux pour alerter l'opinion de risques sociotechniques (Chateauraynaud et Torny, 2005). De façon plus générale, la sociologie de la mobilisation a souligné la capacité des mouvements sociaux à élaborer des cadres cognitifs pour interpréter les nombreuses occurrences provenant notamment des médias (Benford et Snow, 2000, Scheufele, 1999). Des études récentes suggèrent que les technologies numériques, et en particulier les médias sociaux, offrent de

nouvelles possibilités pour interpréter ces occurrences et permettent aux individus d'identifier des problèmes sociaux et de s'organiser (Bennett et Segerberg, 2013 ; Lim, 2012).

1.3 La multitude des riverains

Une dernière façon de former des collectifs autour d'occurrences s'appuie sur la proximité des individus vis-à-vis de ces occurrences. Cette proximité est souvent géographique, à la manière des premiers journaux métropolitains apparus aux États-Unis au début du 19e siècle. Ceux-ci contenaient un grand nombre d'informations hétérogènes et toujours associées à un lieu précis (Nord, 2001). L'occurrence sollicite l'intérêt de l'individu en tant que riverain, parent, consommateur ou usager des services collectifs. Quel que soit le ressort de la proximité, il en résulte la coexistence d'un grand nombre de tout petits collectifs, qui s'agrègent autour de chaque occurrence.

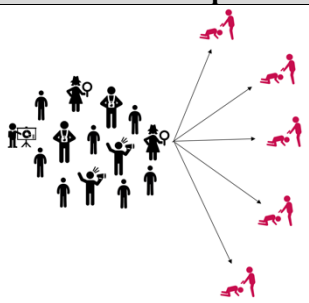
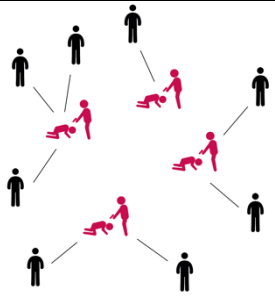
Dans ce modèle, les occurrences n'ont de valeur que si elles sont précisément attachées à un lieu. Elles n'ont pas à être sélectionnées ou même éditorialisées. Elles conduisent rarement les individus à former des jugements publics, ni même vraiment à converser les uns avec les autres. Le citoyen n'est pas vraiment sollicité par des occurrences qui servent surtout à faire des choix de consommation – la bonne école, le bon quartier, etc.

Cette « multitude des riverains » apparaît parfois comme un modèle repoussoir, associé à l'émiettement ou à la dissolution du public. Elle est alors utilisée pour décrire les formes médiatiques antérieures aux médias de masse, ou pour pointer le risque de dissolution de l'espace public associés qui accompagnerait l'essor des médias en ligne. On y fait ainsi souvent référence dans les débats qui accompagnent l'essor d'internet et la transformation des médias (Missika, 2006). C'est ainsi que le juriste américain Cass Sunstein a rencontré un certain succès en alertant sur les risques associés aux possibilités de personnalisation de l'information offertes par internet (Sunstein, 2002). Plus récemment, les débats autour des « bulles de filtre » supposément produites par les sites de réseaux sociaux ont à nouveau mis ce risque sur le devant de la scène (Pariser, 2011). Dès lors que les gens ne sont plus contraints de s'intéresser à des informations qui sont éloignés d'eux, ils finiraient par ne plus partager de préoccupations communes.

Ces trois modèles sont autant de façons de former des collectifs autour d'occurrences multiples et éclatées (figure 2). Ces trois modèles ont rarement fait l'objet d'enquêtes à proprement parler, et l'on sait peu de

choses sur la façon dont les publics se constituent empiriquement – et encore moins dans les contextes numériques. Mais surtout, les dispositifs en ligne qui sont liés au journalisme de données présentent plusieurs particularités : ils donnent accès à un très grand nombre d'occurrences, qui sont traitées de façon standardisée et peuvent être mises en équivalence de plusieurs façons par le biais d'algorithmes. Dans la suite de l'article, nous allons voir que ces dispositifs en ligne combinent les trois façons de faire public que nous avons identifiées.

Fig. 2 – Trois façons de faire public autour d'occurrences

	Colisée invisible	**Collectif d'enquête**	**Multitude de riverains**
			
Nombre d'occurrences	Quelques-unes	Un grand nombre	Un très grand nombre
Forme des collectifs	Grands ensembles d'individus éloignés les uns des autres	Regroupement d'acteurs hétérogènes (citoyens, militants, chercheurs, journalistes, etc.)	Nombreux petits regroupements d'individus proches
Ressort des collectifs	Le partage de sujets d'indignation ou de conversation	L'élaboration de jugements publics visant la résolution d'un problème	Un lien de proximité lié à un quartier, une famille, des amis, etc.
Intermédiaires	Journalistes et organisations de presse	Médias, militants, chercheurs	Entreprises privées et institutions publiques
Critiques	Les journalistes sélectionnent des occurrences arbitraires ou sensationnelles	Un public difficile à faire émerger	Dissolution de l'espace public

2. Suivre la formation d'un public via ses traces numériques

L'exploitation sociologique des traces numériques nourrit aujourd'hui de fortes attentes. On attend notamment de cette « sociologie numérique » qu'elle concilie la profondeur de l'analyse qualitative avec l'étendue de l'analyse quantitative (Lazer et al., 2009). La plupart des chercheurs constatent toutefois qu'un ensemble de difficultés nouvelles apparaissent avec ce type de méthodes, si bien qu'elles sont aujourd'hui loin d'être stabilisées (Marres, 2017 ; Venturini, Cardon et Cointet, 2014 ; Cointet et Parasie, 2018). C'est ainsi que nous avons entrepris d'étudier la formation d'un public via les traces textuelles offertes par la plateforme du *Los Angeles Times*. En quittant le chemin balisé des méthodes sociologiques conventionnelles, nous empruntons un sentier où le bricolage est la norme.

2.1 Une plateforme originale

Lorsque les journalistes du *Los Angeles Times* lancent la plateforme « *Homicide Report* » en 2010, ils veulent rompre avec le modèle du « Colisée invisible » que nous avons présenté. Ils critiquent la tendance des journalistes, y compris ceux du *Los Angeles Times*, à couvrir massivement un petit nombre de meurtres, au détriment de la grande majorité des homicides considérés comme étant dépourvus de valeur éditoriale (Young et Hermida, 2015). Megan Garvey, la rédactrice en chef du *Homicide Report*, explique ainsi que la couverture des homicides est à la fois marquée par la recherche du sensationnel et par un fort biais racial :

> L'adolescente blanche qui a été tuée – ce qui constitue l'exception à la règle – attire énormément d'attention (...) C'est la même chose avec une fusillade de masse. Mais on est arrivé à une situation où les personnes qui se font tuer jour après jour – c'est-à-dire les hommes noirs qui ont entre 17 et 22 ans et qui vivent dans un quartier pauvre – ne sont pas considérés comme ayant la moindre valeur journalistique, du fait des contraintes de l'imprimé et de tout le reste. (Source : Reid, 2014)

Quelques journalistes thématisent comme un problème le fait que leur journal couvre 10 % seulement des meurtres commis chaque année dans la métropole californienne. Ils entreprennent donc de couvrir tous les homicides, d'abord par le biais d'un blog puis au moyen d'une base de données alimentée par le bureau des médecins légistes et par la police de Los Angeles. La recherche de l'exhaustivité va de pair avec la volonté de traiter tous les homicides de la même façon, à travers un ensemble d'informations standardisées obtenues auprès des autorités, et qui concernent la victime, le lieu du crime, les causes et les circonstances du meurtre.

En offrant une couverture exhaustive et standardisée des meurtres commis à Los Angeles, les journalistes veulent donc rompre avec le modèle du « Colisée invisible ». Dans leurs discours publics, ils font référence à deux formes de publics différents. La première forme est proche de celle que nous avons appelé la « multitude des riverains » : il s'agit d'intéresser des personnes qui vivent dans les quartiers où les meurtres ont lieu. Le journaliste de données en charge du projet déclare ainsi que « c'est quelque chose qui peut intéresser les gens qui se soucient de ce qui se passe à proximité de l'endroit où ils vivent » (cité par Young et Hermida, 2015). Ils mentionnent tout particulièrement les familles des victimes, qui « souffrent de ne jamais voir la mort de leur fils apparaître dans le journal » (Jill Leovy, fondatrice de « The Homicide Report », citée par Roderick, 2013). Mais ils font aussi référence à une autre forme de public qui correspond davantage à un « collectif d'enquête ». Ils espèrent en effet que « les lecteurs auront une vision plus réaliste des personnes qui meurent » (*ibid.*), et qu'ils pourront mieux comprendre les causes de la violence urbaine. Lorsqu'ils justifient la possibilité laissée aux internautes de commenter chaque homicide sur la plateforme, les journalistes expliquent que cela doit permettre aux proches de la victime de lui rendre hommage, mais aussi à tout ceux qui le souhaitent de discuter des causes de la violence et de l'intervention de la police.

Nous avons voulu tirer profit de l'important volume de commentaires publiés sur la plateforme – 28 364 commentaires postés entre janvier 2010 et décembre 2016 – pour étudier la façon dont un public se forme autour d'occurrences. Au-delà de l'opportunité offerte par ce corpus, nous avons rapidement été confronté à plusieurs difficultés.

2.2 Une opportunité et des obstacles

Les traces numériques de cette plateforme offrent l'opportunité de saisir la façon dont se forme un public en dépit du caractère multiple et éclaté des occurrences. Il y a plusieurs raisons à cela. D'abord, ces traces permettent d'accéder à une matière textuelle d'une grande richesse, et que l'on peut précisément relier à chaque occurrence. La dimension argumentative est centrale dans la formation d'un public, et on peut ici saisir la façon dont les gens prennent la parole sur un meurtre pour exprimer leur douleur, proposer des explications, partager ou contester l'opinion des autres internautes, etc. Ensuite, cette matière argumentative présente l'intérêt de n'avoir pas été sollicitée par le sociologue, à la différence de ce que l'on obtiendrait par le biais d'un questionnaire. Une parole se manifeste sans que le chercheur ne puisse imposer ses propres catégories de façon *a priori*. Enfin, on peut

attendre de ces traces qu'elles permettent de concilier la profondeur d'une enquête qualitative sur les publics à l'étendue d'une enquête quantitative par questionnaire. La profondeur proviendrait de la riche matière textuelle, tandis que l'étendue procéderait du fait que tous les homicides qui ont eu lieu pendant une période de trois ans apparaissent sur la plateforme et sont susceptibles d'être commentés.

Nous avons donc constitué une nouvelle base de données à partir de trois jeux de données collectés depuis la plateforme par le biais d'algorithmes *ad hoc*[2]. Ces trois jeux de données concernent (1) les informations que la plateforme délivre au sujet de chaque victime (nom, âge, genre et ethnicité) ; (2) les informations diffusées par la plateforme concernant chaque homicide (date, lieu, causes, circonstances, scène de crime) ; (3) l'ensemble des commentaires postés sur la plateforme en lien avec des homicides particuliers (nom de l'auteur du commentaire, date de publication, texte, nombre de signes du commentaire).

Ces données présentent cependant trois difficultés, qui tiennent à leurs conditions de production, au peu d'informations concernant les locuteurs, ainsi qu'à la nature du matériau textuel. Examinons rapidement ces difficultés.

En premier lieu, ces données nous font dépendre entièrement des catégories forgées par les journalistes du *Los Angeles Times* pour décrire les homicides et leurs victimes. Celles-ci résultent de décisions éditoriales faites à partir des informations fournies par les autorités. Comme notre propos n'est pas d'analyser le crime dans la métropole californienne, il importe peu que les données de la plateforme renvoient à la réalité des homicides commis[3]. Notre projet étant de saisir la façon dont les internautes s'assemblent autour de ces occurrences, il suffit d'accéder à la représentation du crime à laquelle les internautes accèdent eux-mêmes. En revanche, nous dépendons totalement de la façon dont la plateforme sollicite et encadre la participation des internautes – notamment à travers la modération qu'exerce la rédaction sur les commentaires de façon *a priori*.

Deuxièmement, les données extraites de la plateforme nous renseignent peu sur les locuteurs. On dispose du nom par lequel l'internaute se signale sur la plateforme, mais on ignore aussi bien son état civil, sa profession, son statut socio-économique que sa situation familiale ou ses préférences politiques.

[2] Il s'agit d'algorithmes d'extraction web élaborés par Jean-Philippe Cointet dans le cadre de la plateforme CorText.

[3] Le fait que les journalistes privilégient les informations provenant des médecins légistes, qu'ils croisent ensuite avec celles de la police, garantit cependant une bonne qualité de l'information.

Bref, il s'agit là d'une contrainte majeure pour qui veut analyser la formation d'un public.

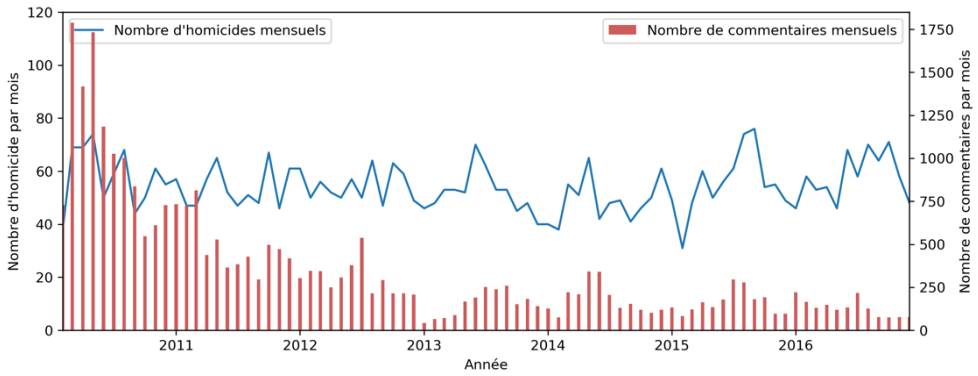
Enfin, nous avons collecté un matériau textuel volumineux, mais qui présente plusieurs difficultés pour l'analyse. Il est très majoritairement en langue anglaise, mais comporte souvent des expressions argotiques et des références culturelles dont la compréhension nous échappe en partie. Surtout, il s'agit de mettre au point une méthode pour saisir, à travers ce matériau relativement inerte, des actions d'interprétation mises en œuvre par les internautes.

Passons maintenant aux solutions que nous avons trouvées pour atténuer ces difficultés liées aux caractéristiques des locuteurs et au traitement du matériau textuel.

2.3 Qui sont les locuteurs ?

Si nous avons au départ peu d'informations sur les auteurs des commentaires, l'exploration des données collectées nous permet d'en apprendre davantage. Comme pour la plupart des plateformes en ligne (cf. Beuscart, Dagiral et Parasie, 2016 : 99-102), la participation est très inégale. Comme on le voit sur la figure 3, elle est d'abord inégalement répartie dans le temps, alors que le nombre d'homicides est lui relativement stable pendant les sept ans. Très utilisée entre 2010 et 2013, la plateforme a ensuite recueilli moins de commentaires à partir de cette date, sans doute au profit des réseaux sociaux.

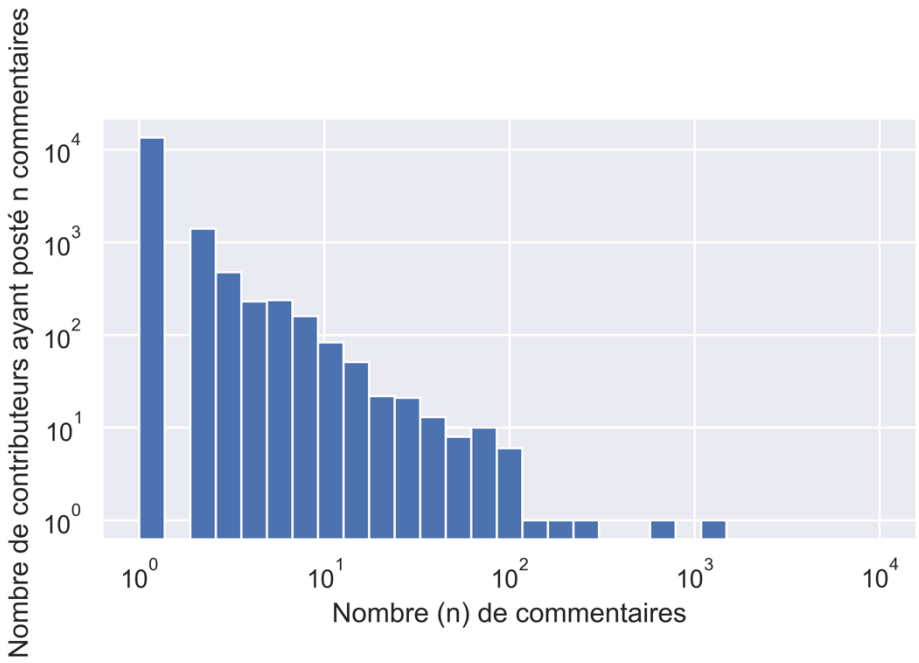
Fig. 3 – Distribution des homicides et des commentaires sur « The Homicide Report » (2010-2016)



On constate ensuite que la participation est très différente selon les contributeurs (figure 4). Sur un total de 16 147 contributeurs[4], 83 % ont posté un seul commentaire, tandis que 1,3 % des contributeurs ont posté plus de 10 commentaires.

Fig. 4 – Distribution des contributeurs selon le nombre de commentaires postés

[4] Il n'est pas du tout certain qu'à chaque pseudonyme corresponde un locuteur unique. Deux éléments permettent pourtant de supposer que la différence entre le nombre de locuteurs et le nombre de pseudonymes n'est pas très importante. D'une part, une partie des locuteurs s'identifient auprès des autres par le biais de leur pseudonyme. D'autre part, les locuteurs qui interviennent peu affichent souvent leur proximité avec la personne disparue.



C'est ainsi que nous avons été conduit à distinguer deux populations parmi les auteurs, selon le nombre d'homicides qu'ils commentent. Nous avons choisi de séparer, d'une part, des « contributeurs actifs », qui commentent au moins dix homicides différents, et, d'autre part, des « contributeurs ponctuels », dont la prise de parole porte sur un plus petit nombre d'occurrences. La population des « contributeurs actifs » se compose de 76 auteurs, qui sont, à eux seuls, à l'origine de 16 % de l'ensemble des commentaires[5]. En parcourant leurs publications, on s'aperçoit que ces « contributeurs actifs » poursuivent en effet un agenda politique spécifique, certains défendant systématiquement l'action de la police tandis que d'autres dénoncent les violences policières. *Syscom3*, qui est de loin l'auteur qui intervient le plus, avec 1 142 commentaires publiés au sujet de 449 homicides. De façon systématique, il prend la parole pour dénoncer la responsabilité des gangs dans la violence urbaine, ce qui le conduit souvent à « blâmer la victime » lorsqu'il suspecte que celle-ci était liée à un gang. À l'opposé, *Jag* intervient régulièrement pour dénoncer la discrimination raciale dont se rend coupable la police de Los Angeles. Sur la période, il est l'auteur de 350 commentaires qui concernent 186 homicides.

2.4 Comment saisir des énonciations ?

[5] Dans les espaces de discussion en ligne, le rôle des contributeurs les plus actifs est souvent structurant (Graham et Wright, 2014).

Pour qu'elle nous permette de saisir le processus par lequel des internautes en viennent à faire public autour des occurrences, l'analyse du matériau textuel doit se plier à plusieurs contraintes. Il s'agit d'abord de ne pas considérer ce matériau comme un ensemble d'*énoncés* inertes, mais plutôt comme un ensemble d'*énonciations*. Autrement dit, il s'agit de saisir le mouvement par lequel les locuteurs en viennent à faire un travail d'interprétation de l'occurrence, en lien avec les autres locuteurs. Nous nous appuierons ici sur le cadre élaboré par Luc Boltanski dans ses travaux pionniers sur l'analyse des prises de parole en public (Boltanski, Schiltz et Darré, 1984). S'inspirant de la sémiotique, il envisage toute prise de parole comme la construction de relations entre différents « actants ». Dans le cas qui nous occupe, les actants sont le locuteur qui prend la parole ; la victime du meurtre ; les responsables du drame ; et le public dont on sollicite la compassion ou que l'on prend à témoin d'une injustice. Toute prise de parole sur *The Homicide Report* s'apparente ainsi à une certaine façon de relier ces actants.

Mais il s'agit également de ne pas nous en remettre aveuglément aux algorithmes. Le sociologue doit en effet pouvoir assumer la responsabilité de l'interprétation qu'il élabore. C'est la raison pour laquelle nous avons passé un temps considérable à nous familiariser avec ses particularités afin de définir des modalités pertinentes de codage des commentaires. Ces six modalités sont des variables binaires qui s'inspirent du cadre pragmatique exposé dans le paragraphe précédent :

1. **[*Lien locuteur-victime*]** Cette variable concerne le lien entre le locuteur et la victime. Dans son commentaire, l'auteur exprime ou non un lien personnel avec la personne assassinée. L'expression du lien peut prendre différentes formes selon que le locuteur s'adresse directement à la victime (« *you were special* »), rend manifeste un lien familial ou amical avec elle (« *he was a dear friend* »), exprime des souvenirs partagés avec la victime (« *we had so many good laughs* »), ou manifeste un lien de connaissance d'une autre manière (« *I lost someone very special* ») ;
2. **[*Affection vis-à-vis de la victime ou de ses proches*]** Cette variable correspond à des situations où le locuteur exprime son affection à la personne disparue (« *rest in peace* », « *you will not be forgotten* », « *You're Forever in our Hearts* »), ou à ses proches en leur adressant ses condoléances (« *my prayers go to the family* ») ;
3. **[*Évaluation morale des individus*]** Cette variable identifie la présence d'un jugement sur la valeur morale de la victime, de ses proches ou des suspects. Elle peut se manifester sous la forme de jugements positifs (« *he was a loving and devoting husband and father* », « *my nephew was so loving* » ; « *He stood out as an*

employee, always smiling, dressed nicely and eager to help people in need ») ou de jugements négatifs (« *this guy was evil* » ; « *He was a known notorious gangster* » ; « *a 15 year old punk* »). Dans ce second cas, l'auteur du commentaire se contente d'établir une responsabilité individuelle, blâmant la victime pour son style de vie, ou d'autres personnes (parent, policiers, etc.) pour leurs comportements. Ici, l'homicide est exclusivement considéré comme la conséquence d'actes individuels.

4. [***Recherche d'une responsabilité collective***] Cette variable correspond à des situations dans lesquelles l'auteur du commentaire mentionne des responsabilités plus générales, invoquant des êtres collectifs qui vont au-delà des individus impliqués dans l'homicide. Nous distinguons trois sous-variables associées à trois formes distinctes de généralisation :
    a. [***Problèmes publics***] Cette sous-variable identifie les commentaires qui associent l'homicide à un ou plusieurs problèmes plus généraux (violence urbaine, culture des gangs, violence policière, crise du système éducatif, etc.). Elle n'exige pas que le locuteur identifie des responsabilités ou des solutions, mais seulement que l'occurrence soit mise en série et considérée comme renvoyant à un problème plus global (« *with all the violence in our society* », « *these killing fields neighborhoods* », « *plenty of murders like this* », « *We as society have to come to accept that certain young teens are damaged goods* ») ;
    b. [***Institutions***] Cette sous-variable identifie les commentaires qui évoquent la responsabilité d'une institution (l'Etat, les tribunaux, la loi, la police, l'école, les églises, etc.). Ces institutions peuvent être aussi bien pointées comme la source de la violence (« *the failure of the religious leaders* ») que convoquée comme une solution (« *start DEMANDING action by your elected representatives* ») ;
    c. [***Groupes sociaux/ethniques***] Cette dernière sous-variable identifie les commentaires dans lesquels des groupes sociaux ou ethniques sont mentionnés (« *black on black, latino on latino crime* », « *low income residents* », « *You white people are funny as hell* »).

2.5 Une catégorisation par apprentissage supervisé

Du fait de la variété des objets qu'elles mobilisent, et de la multiplicité des styles d'écriture individuels, chacune de ces variables renvoie à des formes lexicales et syntaxiques hétérogènes. Dans ces conditions, une approche strictement lexicométrique est très difficile à mettre en œuvre, puisqu'il est

pratiquement impossible de faire la liste de toutes les expressions correspondant à chaque variable. Plus satisfaisant en principe, un codage strictement humain engagerait toutefois des ressources trop importantes. C'est pourquoi nous avons opté pour des algorithmes d'apprentissage supervisé pour catégoriser l'intégralité des messages de la plateforme en fonction de ces différentes modalités.

Les méthodes de classification d'un corpus textuel par apprentissage supervisé ont été jusqu'à aujourd'hui très peu utilisées en sciences sociales (Hillard et al., 2008 ; Burscher et al., 2015). Il n'existe aucune procédure stabilisée permettant d'évaluer, du point de vue des sciences sociales, la qualité de la catégorisation d'un corpus textuel par apprentissage supervisé. L'évaluation est d'autant plus délicate que les algorithmes en *machine learning* souffrent parfois d'une certaine opacité au sens où il n'est souvent pas possible d'expliciter de façon simple les critères qui sont retenus par le classifieur pour opérer. L'algorithme d'apprentissage que nous utilisons ne fait pas exception puisqu'il s'agit d'un réseau de neurones. Pratiquement nous utilisons le logiciel Prodigy[6] qui présente l'avantage d'offrir à la fois une interface d'annotation et un moteur d'inférence permettant de construire un classifieur de textes. Déjà utilisé par des chercheurs en science politique et en communication (Liang et al., 2018), *Prodigy* combine un module d'analyse linguistique (incluant analyse morphosyntaxique, vecteurs sémantiques, parseur sémantique), une interface utilisateur et un module d'apprentissage actif[7] couplé à cette dernière pour permettre l'apprentissage d'une catégorie sur le corpus.

Pratiquement, nous avons entraîné un classifieur différent pour chacune de nos 6 variables en suivant les trois étapes suivantes :

(1) Nous avons d'abord établi une liste d'une dizaine d'expressions dont nous pensons qu'elles sont susceptibles d'être des bons marqueurs pour chaque catégorie (par exemple, les termes « *nephew* » ou « *mother* » pour ce qui concerne le lien entre locuteur et victime). Prodigy se sert par la suite de ces listes pour l'aider dans la sélection des commentaires proposés à l'annotateur. Cet échantillon n'est pas aléatoire puisqu'il vise à permettre au logiciel d'apprendre plus rapidement et efficacement à identifier les commentaires rentrant dans l'une ou l'autre des modalités ;

[6] https://support.prodi.gy/

[7] L'apprentissage actif renvoie à la façon dont le corpus d'apprentissage est volontairement biaisé de façon à proposer au réseau de neurones des exemples qui sont les plus susceptibles de guider le classifieur dans sa tâche. L'apprentissage actif permet de rééquilibrer la distribution des exemples fournis à la machine lorsqu'une catégorie de documents est rare dans un corpus.

(2) Les deux auteurs de l'article ont ainsi codé manuellement environ 800 commentaires uniques pour chaque variable (ce qui a permis de produire des mesures de confiance inter-codeurs, cf. annexe 1). À partir de ce codage humain, le logiciel infère un classifieur en utilisant 75 % des commentaires étiquetés pour l'apprentissage du réseau et 25 % pour son évaluation. Cette évaluation permet de mesurer la qualité du classifieur en terme de précision et de rappel. Ces mesures étant satisfaisantes (cf. annexe 1), nous avons ensuite appliqué chaque modèle à l'ensemble du corpus pour construire un codage de l'intégralité du corpus de commentaires.

(3) Afin de tirer parti des possibilités offertes par le logiciel, nous avons élaboré un protocole destiné à évaluer, au-delà des mesures de qualité des classifieurs calculés sur un échantillon d'évaluation par essence biaisé, la pertinence sociologique de la façon dont le logiciel catégorise les commentaires du corpus. Nous avons ainsi procédé à une évaluation *ex post* d'un échantillon de 300 commentaires tirés aléatoirement dans le corpus, en comparant les catégorisations produites par le logiciel et celles effectuées manuellement (cf. annexe 2).

2.6 Cinq ensembles de commentaires

Examinons les principales formes d'énonciation qui traversent le corpus. Il apparaît d'abord que l'ensemble du corpus se divise en deux grandes parties : les commentaires dans lesquels le locuteur affiche un lien personnel avec la personne disparue (58 % du corpus) ; les commentaires dans lesquels aucun lien n'est exposé entre le locuteur et la personne disparue (42 % du corpus).

Dans la première partie du corpus, on peut distinguer un ensemble de commentaires dans lesquels le locuteur exprime son affection vis-à-vis de la personne disparue (et/ou de ses proches), mais sans se prononcer sur sa valeur morale. Il s'agit d'« hommages centrés sur l'amour » (32,6 % du corpus), dans lesquels le locuteur exprime un amour inconditionnel envers la victime. L'évaluation morale de la victime est ici sans objet, comme dans le commentaire ci-dessous :

> Olga, je n'arrive pas à croire que tu sois partie si vite... Je chérirai toujours nos merveilleux souvenirs de l'école... Je t'aime et tu me manqueras toujours... Tu nous manqueras à jamais... Que Dieu bénisse tes enfants et ta famille dans ces moments difficiles... Puisses-tu reposer dans la paix céleste...

> Pour toujours, Letty... [Letty M., 14 mai 2010 ; homicide d'Olga Martinez]

Parmi les commentaires où s'exprime un lien personnel avec la personne disparue, on identifie ensuite les « hommages centrés sur la morale » (22 % du corpus). Il s'agit de commentaires qui se focalisent sur les qualités morales de la personne disparue. Le locuteur ne manifeste pas seulement ici son amour à l'égard de la personne disparue, mais il met aussi en avant ses qualités – il était un « fils merveilleux », un « père aimant », une « femme forte », etc. :

> Un des meilleurs jeunes hommes qu'il m'ait été donné de rencontrer. Un bon père, un frère formidable, un fils aimant. Sean n'a jamais rencontré une personne qu'il n'ait pas aimée. Son cœur était aussi grand que son amour pour qui le rencontrait. Nos cœurs sont brisés, notre perte est grande. Aidez-nous à trouver la justice, pour que nous puissions faire notre deuil et trouver la paix. Une vie qui nous a été prise si tôt. [Sherree, 9 août 2015 ; homicide de Sean Sylvester]

Dans la seconde partie du corpus, où le locuteur n'exprime aucun lien avec la victime, nous avons identifié trois ensembles distincts. Il y a d'abord les « hommages à distance » (15,2 % du corpus), à travers lesquels les contributeurs expriment leur affection vis-à-vis de la personne disparue et de ses proches, sympathisent avec leur douleur sans les connaître personnellement :

> Christopher semblait être un être humain formidable, prometteur, doué et généreux. C'est si triste pour sa famille. J'espère que ceux qui sont responsables de son meurtre seront jugés dans toute la mesure de la loi. J'adresse mes prières quotidiennes à la famille de Christopher. J'aimerais savoir s'il y a un moyen d'aider financièrement sa famille. [Linda Willson, 17 juin 2015 ; homicide de Christopher Jermaine Handy]

Un autre ensemble regroupe les « évaluations morales à distance » (14,3 % du corpus). Les auteurs évoquent ici la valeur morale de la victime, de son entourage familial ou amical, ou des personnes impliquées dans l'homicide. Puisque les locuteurs ne connaissaient pas personnellement la victime, leur évaluation morale se focalise sur ses activités – était-elle membre d'un gang ? Était-elle impliquée dans des activités illégales ? Ou s'agissait-il au contraire d'une victime innocente qui se trouvait au mauvais endroit au mauvais moment ? La particularité de ces commentaires, c'est qu'ils ne relient pas le meurtre précis à un ensemble d'éléments systémiques (tels que la pauvreté des quartiers, les politiques sociales ou les discriminations

subies par les minorités), mais se concentrent uniquement sur des explications morales.

> ... « Robert était une belle personne en dehors de son activité de gangster, il était une personne très respectée et une bonne personne ».
> Dans quel type de culture ? Dans le monde normal, où les gens ne commettent pas de crimes violents, son comportement est considéré comme sociopathe. Une personne qui a une famille reste en dehors de ce style de vie de façon à subvenir à leurs besoins.
> ... « il n'est pas juste que les gens disent du mal de la personne décédée et de sa famille ».
> Que peut-on dire d'un individu qui se rend en voiture à une fusillade dans un quartier résidentiel ?
> ... « Non, Robert ne possédait pas de pistolet »
> Et comment expliquez-vous qu'il portait une arme et que celle-ci était chargée ?
> ... « Personne n'est parfait »
> Mais peu d'individus adoptent un style de vie violent et agissent comme si c'était normal. Ce propos est une phrase que les défenseurs des gangs utilisent pour justifier l'injustifiable.
> ... « Ne vous mettez pas à la place de Dieu »
> Ce n'est pas nous qui NOUS PRENONS POUR DIEU en décidant de tirer sur une autre personne. [Syscom3, 29 octobre 2010 ; homicide de Robert Earl Gipson]

Le commentaire ci-dessus incarne cette « morale à distance ». L'auteur s'en prend aux proches de la victime qui lui attribuent un ensemble de qualités personnelles, pour pointer le caractère profondément immoral de la vie vécue par la victime. L'interprétation qui est ici élaborée est strictement individuelle : la personne disparue a choisi un « style de vie violent », ce qui la rend responsable de sa propre mort et de l'ensemble des dégâts occasionnés.

Enfin, un dernier ensemble de commentaires regroupe les « quêtes de responsabilités collectives » (14,4 % des commentaires). On entend par là des commentaires qui ne manifestent aucun lien entre le locuteur et la victime, et qui mentionnent des êtres collectifs – des groupes sociaux, des institutions – ou envisagent l'occurrence comme un problème public et pas uniquement comme un problème de morale individuelle. Ces commentaires sont souvent beaucoup longs et moins directement reliés aux détails de l'homicide particulier auxquels ils sont associés :

> Je ne veux plus que le moindre dollar de mes impôts vienne financer ces politiques sociales qui ont échoué. J'adorerais que l'Etat-providence soit transformé en système de travail obligatoire : vous vous pointez et vous travaillez une journée pour faire ce dont la ville a besoin, balayer les rues, nettoyer les graffitis, ramasser les poubelles, etc. L'autre chose que j'adorerais, c'est que nos prisonniers soient mis au travail. D'abord en cultivant pour se nourrir, ensuite en faisant le travail pour lequel on embauche les immigrés illégaux. On ferait d'une pierre deux coups : les prisonniers paieraient pour leur séjour en prison, et ils apprendraient une éthique de travail. Mais je crains que ce commentaire ne soit pas publié parce qu'il n'est pas politiquement correct. [*Eye Opener*, 31 décembre 2010 ; homicide de Cesar Guerrero]

Ces 5 ensembles regroupent 87,2 % des commentaires publiés sur *The Homicide Report* entre 2010 et 2016. Ils dessinent des façons spécifiques – et conflictuelles – de donner du sens aux multiples occurrences diffusées par la plateforme. En nous appuyant sur l'identification de ces formes d'énonciation, nous allons maintenant mettre au jour la coexistence de trois formes de public, qui se forment chacune selon des processus distincts. Examinons-les successivement.

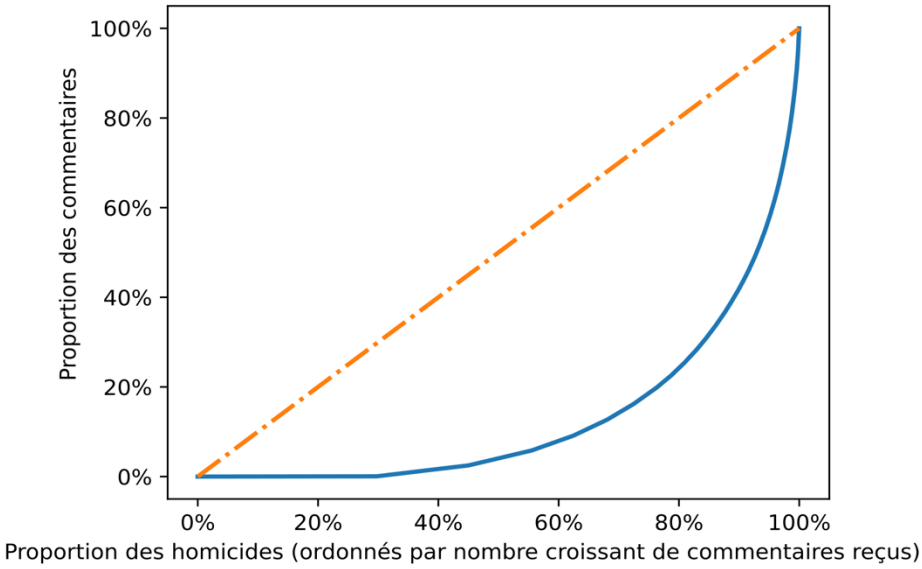
3. Se rassembler autour d'une minorité d'occurrences choisies

Une première façon de « faire public », caractéristique des médias de masse, consiste à rassembler un grand nombre d'individus autour d'un petit nombre d'occurrences sélectionnées pour leur valeur éditoriale. C'est ce que nous avons appelé, à la suite de Gabriel Tarde, le « Colisée invisible ». Or, nous allons voir que les utilisateurs de la plateforme du *Los Angeles Times* s'agrègent d'une manière qui emprunte à cette façon de faire public. Alors même qu'ils peuvent accéder à n'importe quelle occurrence, les contributeurs concentrent leur attention sur un petit nombre d'occurrences, qui sont sélectionnées selon un ensemble de critères partagés.

Le graphique ci-dessous donne à voir la distribution des commentaires entre l'ensemble des homicides commis sur la période (figure 5). Si toutes les occurrences suscitaient un même nombre de commentaires, la distribution prendrait alors la forme de la droite en pointillée. Au lieu de cela, elle correspond à la courbe noire, ce qui marque une forte concentration sur un petit nombre d'occurrences – 80 % des commentaires portent sur 30 % des

homicides. Certes, la sélection opérée par les internautes qui prennent la parole est moins forte que celle qu'effectuent les journalistes du *Los Angeles Times* version imprimée. Nous avons en effet calculé que 100 % des articles du journal consacrés aux homicides portent sur seulement 10 % d'entre eux[8]. Mais s'il est légèrement moins prononcé, on observe bel et bien un processus de sélection des occurrences par les internautes.

Fig. 5 – Distribution du nombre de commentaires reçus par homicides



Or cette sélection n'est pas seulement le résultat de choix individuels qui seraient indépendants les uns des autres. Il repose au contraire sur un ensemble de critères relativement partagés. C'est là un autre point commun avec la façon dont les publics médiatiques classiques s'assemblent autour des occurrences. Journalistes et internautes ont beau mettre en œuvre des critères qui diffèrent profondément, leur sélection repose sur des critères cohérents.

Examinons d'abord les critères mobilisés par les journalistes du *Los Angeles Times* lorsqu'ils couvrent un homicide dans le journal imprimé. La régression statistique effectuée sur une période de trois ans (cf. annexe n°3, colonne A) indique qu'ils couvrent de façon préférentielle les meurtres qui se produisent dans les quartiers réputés plus sûrs, et qui touchent des

[8] À partir du site du LATimes.com, nous avons vérifié manuellement si chacun des homicides commis sur la période avait fait ou non l'objet d'un article dans le journal.

victimes plutôt âgées. Ils écartent systématiquement les homicides qui se passent dans la rue, et couvrent davantage ceux qui se produisent à la suite d'un cambriolage ou qui impliquent les forces de police. Ces résultats confirment les recherches conduites aux États-Unis depuis plusieurs décennies (Roshier, 1973 ; Katz, 1987) : les journalistes privilégient les occurrences exceptionnelles, qui ne sont pas liées à la violence par arme à feu qui touche surtout les jeunes hommes noirs et hispaniques des quartiers pauvres[9].

Les homicides qui font parler les internautes ne sont pas du tout ceux qui retiennent l'attention des journalistes. La régression statistique (cf. annexe n°3, colonne B) indique en effet que les internautes commentent en priorité les homicides qui ont lieu dans les quartiers pauvres, et qui touchent les jeunes noirs tués par arme à feu. Ils se désintéressent ainsi des meurtres dont sont victimes les personnes plus âgées dans les quartiers plus riches. L'intérêt des internautes se concentre donc en priorité sur les occurrences qui sont systématiquement négligées par les médias traditionnels, et qui constituent la masse des homicides commis dans la métropole californienne – ceux qui frappent les jeunes noirs, souvent en lien avec les gangs.

La sélection opérée par les journalistes et les internautes n'est toutefois pas totalement différente. D'abord parce qu'il existe un critère commun – le fait que la police soit impliquée dans l'homicide[10] –, auquel les internautes accordent cependant un poids beaucoup plus important. Ensuite parce que le fait qu'un homicide soit couvert par le journal augmente ses chances d'être commenté par les internautes.

L'essor de ce type de plateformes ne signale donc pas la disparition des façons de faire public associées aux médias de masse. Plusieurs recherches actuelles montrent même que ces façons de s'assembler autour de contenus culturels ne sont pas systématiquement bouleversée avec les technologies en ligne (Beuscart, Beauvisage et Maillard, 2012). Dès lors que les journalistes ne filtrent plus les occurrences, les internautes eux-mêmes mettent en œuvre un processus de sélection des occurrences. Le « Colisée invisible » demeure d'actualité, même si les internautes ont désormais plus de poids dans l'élaboration des critères qui président à la sélection des occurrences.

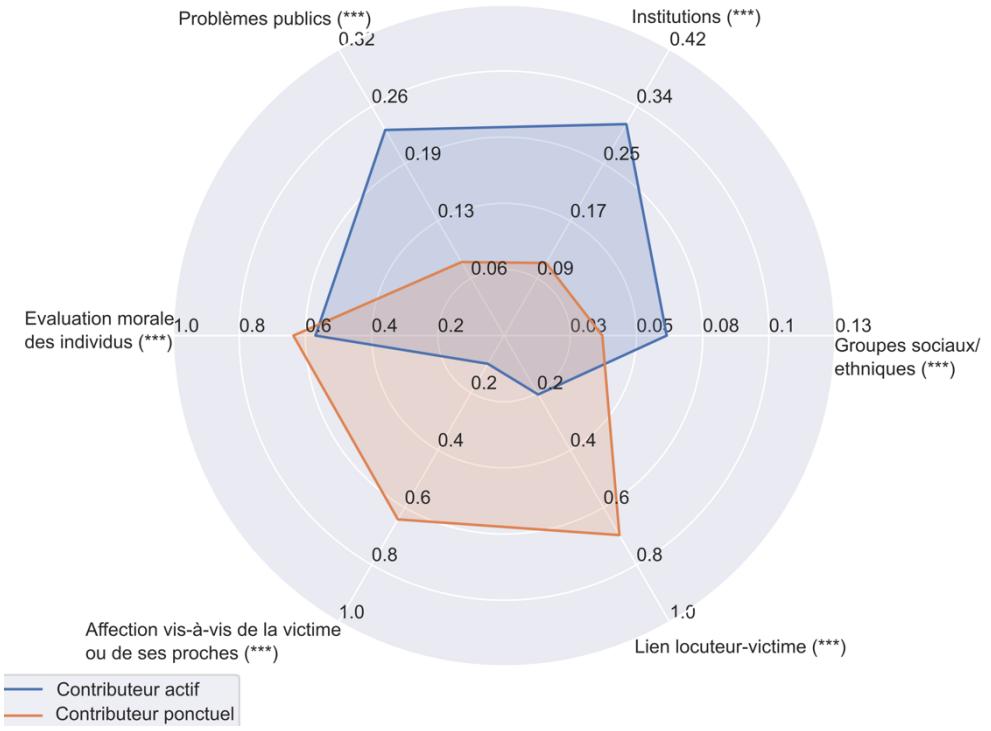
4. Comment en vient-on à enquêter collectivement ?

[9] Sur la période, les homicides qui frappent de jeunes hommes noirs et hispaniques tués par arme à feu représentent 65 % de l'ensemble des homicides.
[10] Sur la période, 7,2 % des homicides impliquent la police.

Une deuxième façon de faire public rassemble des personnes qui enquêtent sur les occurrences dans le but de trouver des solutions à un problème qui les affecte collectivement. Ici, le problème est celui de la violence urbaine, qui touche la plupart des grandes métropoles américaines, et tout particulièrement Los Angeles avec 700 personnes qui perdent la vie chaque année. Aux États-Unis, ce problème donne lieu à des débats structurés – autour de la relégation sociale, économique et raciale qui frappe certains quartiers, du rapport à la violence qui règne dans ces quartiers, du rôle joué par la police à l'encontre des minorités, etc. – et s'inscrit dans des formes de politiques urbaines spécifiques (Donzelot et al., 2003 : 323-359). Or, la plateforme *Homicide Report* est aussi investie par des individus qui s'intéressent aux occurrences, non parce que celles-ci les touchent personnellement, mais plutôt parce qu'elles sont l'occasion d'échanger autour du problème de la violence urbaine. Nous allons voir que ces locuteurs mettent en équivalence un très grand nombre d'homicides, en cherchant collectivement, à partir d'un répertoire cognitif et politique, à identifier des schèmes explicatifs plus généraux du problème de la violence.

Intéressons-nous à ceux que nous avons appelés les « contributeurs actifs », et qui commentent de nombreux homicides différents. En retenant le seuil de 10 homicides commentés, nous avons une population de 76 auteurs qui, à eux seuls, sont à l'origine de 16 % de l'ensemble des commentaires. Ces locuteurs s'intéressent à de nombreux homicides, à l'image de *Syscom3* qui publie plus de 1 000 commentaires au sujet de 368 homicides. On peut penser que ces contributeurs actifs exploitent les possibilités de mise en équivalence offertes par la plateforme, en cherchant à identifier rapidement les meurtres intéressants. Au-delà du nombre considérable d'homicides qu'ils commentent, ces locuteurs se distinguent aussi des autres usagers de la plateforme par la forme de leur argumentation. Non seulement ils ne manifestent aucun lien personnel avec la victime, mais ils mobilisent presque exclusivement des êtres éloignés de la personne décédée – ils parlent surtout des institutions, des problèmes publics, des groupes sociaux et ethniques, de la société américaine. Sur le diagramme ci-dessous, on voit ainsi que les contributeurs actifs mobilisent des formes d'argumentation très différentes des contributeurs ponctuels (figure 6).

Fig. 6 –Diagramme en radar de la participation selon le type de contributeurs



Lecture : *le polygone bleu dessine le profil argumentatif des commentaires des contributeurs actifs. Il relie six points correspondant aux proportions des commentaires de contributeurs actifs qui ont été codés dans chaque variable. Ainsi, 30 % des commentaires de contributeurs actifs mentionnent les institutions, alors que seuls 10 % des commentaires de contributeurs ponctuels comportent de telles mentions.*

Test exact de Fisher : *p < .05; **p < .01; ***p < .001

Il apparaît alors que les interventions des contributeurs actifs s'inscrivent nettement plus souvent dans les ensembles « quête de responsabilités collectives » et « morale à distance » – alors que ceux des locuteurs locaux s'inscrivent dans les ensembles « hommages centrés sur l'amour » et « hommages centrés sur la morale ».

L'analyse d'un échantillon de commentaires montre que ces contributeurs actifs poursuivent souvent un agenda politique particulier. C'est le cas de *Syscom3*, dont la rhétorique consiste à accuser la victime d'être responsable de son sort en ayant fait le choix d'une existence irresponsable, immorale et violente :

> Comment peut-on aimer ce type ? Il est la principale raison de l'effondrement de ces quartiers. Il est comme le signe avant-coureur

> de la malédiction et du crime. On peut attendre de sa famille qu'elle dise les choses habituelles sur son compte. Mais pourquoi en serait-il de même pour nous autres ? Il a commencé à chercher les problèmes, et il savait exactement à quoi s'attendre. Il savait que ses délits entraîneraient une réaction violente. Combien de dizaines de milliers de dollars ont été dépensés pour réparer les dommages qu'il avait causé à la propriété des gens ! Ciel, soutenez-vous les activités des mécréants et des barbares comme ce type ? Pourquoi ? Êtes-vous tellement immergés dans le monde violent et sociopathe auquel il appartenait que vous y voyez un comportement normal et acceptable ? [Syscom3, 15 avril 2010 ; homicide de José Castillo]

À l'opposé du spectre politique, *Jag* intervient lui pour pointer les discriminations dont sont victimes les minorités de la part de la police de Los Angeles :

> Les personnes défavorisées, qui sont en majorité des minorités, sont toujours les laissés pour compte. La plupart sont obligées de plaider coupable pour éviter de rester plus de temps en prison pour des crimes qu'elles n'ont pas commis. J'évite de jouer la carte de la race, mais quand un clochard blanc se fait tuer par des policiers à Fullerton, on demande des comptes à ces policiers. Mais quand un policier tue un hispanique ou un noir, la justice ne bouge pas. Syscom3, tu aimes les statistiques et je me demandais si tu avais les chiffres des gens qui sont accusés de crime et qui sont effectivement condamnés. Je suis presque certain qu'on est moins dur avec les personnes blanches qu'avec les minorités. C'est triste, mais vrai. [Jag, 17 novembre 2012 ; homicide d'Amondo Casillas]

Comme on le voit dans ce commentaire, certains contributeurs actifs s'interpellent les uns les autres. Ils fréquentent la plateforme depuis longtemps, si bien qu'un nouvel homicide est pour eux l'occasion de poursuivre une discussion en cours depuis plusieurs mois, voire plusieurs années. D'une occurrence à l'autre, leur argumentation présente une grande continuité. Ils défendent en effet les mêmes positions tranchées, qui opposent les défenseurs de la police (*cop supporters*) aux critiques de la police (*police bashers*). Manifestement, ces positions sont étroitement associées à des formes d'activisme qui trouvent leur origine en dehors de la plateforme. Ce « collectif d'enquête » s'inscrit bien dans cette tradition sociologique qui souligne la capacité des mouvements sociaux à élaborer des « cadres cognitifs » offrant des façons cohérentes d'interpréter des occurrences disparates (Benford et Snow, 2000).

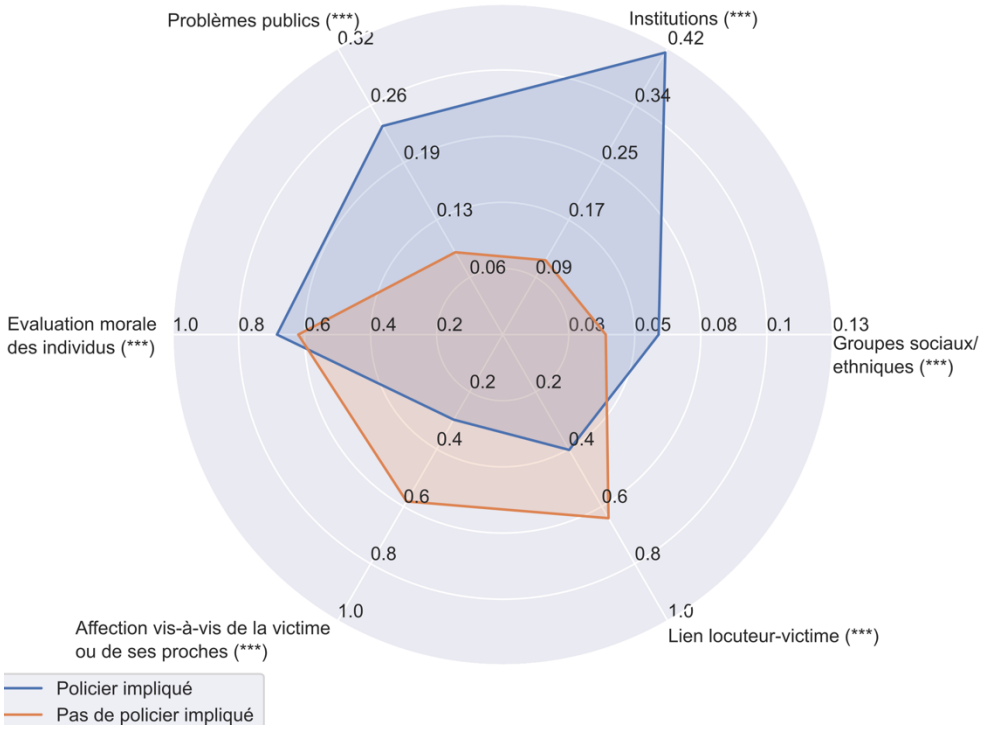
Quand une organisation de presse cesse de sélectionner les occurrences qu'elle donne à voir à son public, une deuxième façon de faire public apparaît ainsi qui, tout en étant très différente de la précédente, existe elle aussi depuis longtemps.

5. À quelles conditions des riverains forment-ils un public ?

Une dernière façon de faire public, que nous avons appelé la « multitude de riverains », est habituellement convoquée comme une figure repoussoir, associée à la dissolution de l'espace public. Dès lors que les individus s'intéressent uniquement aux occurrences qui les touchent personnellement, ils en viendraient à ne plus partager les mêmes préoccupations et à perdre l'horizon du bien commun. Jamais interrogé empiriquement, cet argument mérite d'être considérablement nuancé. Le cas de la plateforme *Homicide Report* montre en effet qu'à certaines conditions, des internautes peuvent en venir à partager des interprétations, orientées vers le bien public, alors même qu'ils s'agrègent autour d'un grand nombre d'occurrences qui les affectent personnellement. Ces conditions concernent les particularités de l'homicide ainsi que les interactions entre contributeurs actifs et ponctuels.

Considérons uniquement les contributeurs ponctuels, c'est-à-dire ceux qui commentent moins de dix homicides. Rappelons que 67 % d'entre eux expriment un lien personnel avec la victime (figure 6). Or, il apparaît d'abord que lorsque l'occurrence présente certaines caractéristiques, ces contributeurs ponctuels interviennent différemment, en mobilisant davantage des registres d'interprétation faisant intervenir les institutions, les groupes sociaux ou ethniques, ainsi le registre des problèmes publics. C'est le cas lorsque l'homicide a directement impliqué la police – autrement dit, lorsqu'un policier est à l'origine du décès de la victime. Sur le diagramme ci-dessous (figure 7), on voit ainsi que dès lors que la victime est tombée sous les coups ou les balles de la police, les contributeurs ponctuels interviennent davantage selon dans le registre de la « quête de responsabilités collectives ». Ils mentionnent beaucoup plus les institutions, interprètent l'homicide particulier comme un problème public, en évoquant davantage la couleur de peau et les problèmes de discrimination. Comme le suggèrent les recherches ethnographiques menées dans les quartiers centraux de Los Angeles, on peut supposer que ces interprétations prennent appui sur un ensemble de normes partagées par la population noire de ces quartiers (Costa Vargas, 2006).

Fig. 7 – Diagramme en radar de la participation des contributeurs ponctuels selon que la police ait été ou non impliquée dans l'homicide



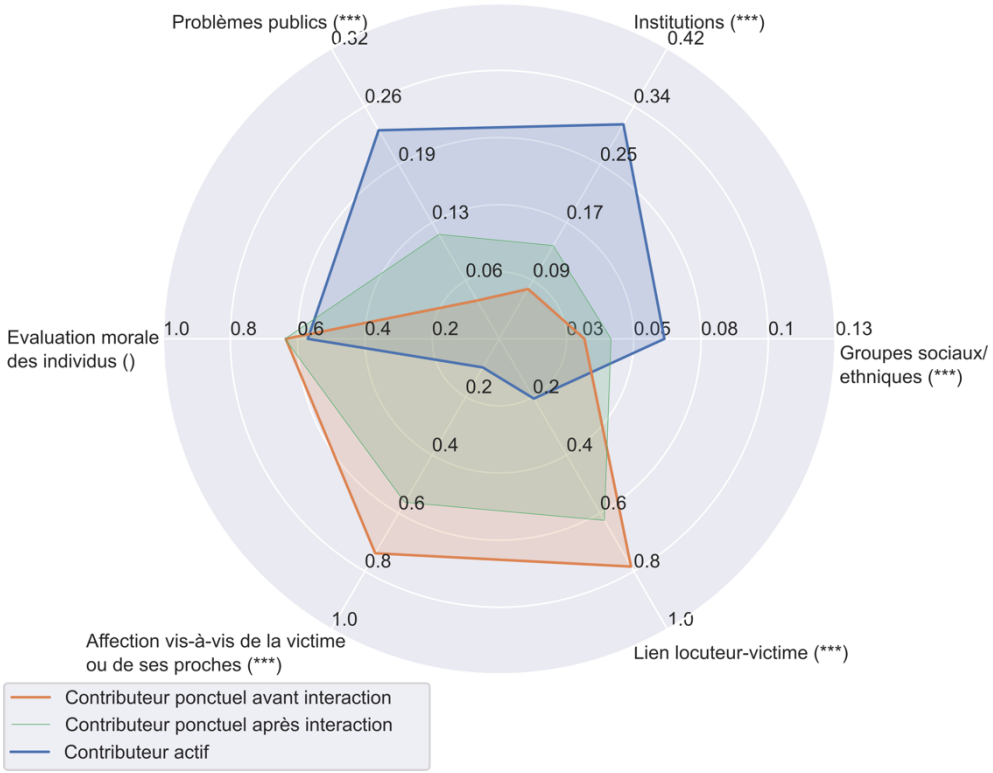
Lecture : *le polygone bleu dessine le profil argumentatif des commentaires des contributeurs ponctuels, dans le cas où un policier est à l'origine du décès. Il relie les points correspondant à la proportion de ces commentaires qui ont été codés dans chaque variable. Ainsi, 42 % des commentaires de contributeurs ponctuels mentionnent les institutions lorsqu'un policier est à l'origine du décès, contre 10 % lorsque la police n'est pas à l'origine du décès.*

Test exact de Fisher : *p < .05; **p < .01; ***p < .001

Une seconde condition favorise la mobilisation des registres publics par les contributeurs ponctuels. Il s'agit des interactions qu'ils ont avec les contributeurs actifs. Ces derniers, nous l'avons vu, mobilisent très majoritairement les registres de la « morale à distance » et de la « quête de responsabilités collectives ». Leurs interventions sont souvent violentes pour les locuteurs locaux, à la fois parce qu'elles proposent des interprétations générales et sont porteuses de jugements tranchés sur la personne disparue. Or, l'analyse quantitative montre que dès lors que des

contributeurs actifs se mettent à commenter un homicide, cela a des effets sensibles sur la participation des locuteurs locaux (figure 9).

Fig. 8 –Diagramme en radar de la participation des contributeurs ponctuels selon qu'il y ait ou non interaction avec des contributeurs actifs



Lecture : *le polygone rouge dessine le profil argumentatif des commentaires des contributeurs ponctuels dans le cas où aucun contributeurs actifs n'a encore commenté l'homicide. Le polygone vert dessine le profil argumentatif des commentaires des contributeurs ponctuels après qu'au moins un contributeur actif soit intervenu.*

Test exact de Fisher : *p < .05; **p < .01; ***p < .001

Mais une fois qu'au moins un contributeur actif a posté un commentaire sur l'homicide, alors les contributeurs ponctuels modifient leurs registres d'intervention. Ils mentionnent moins souvent leur lien à la victime, expriment moins souvent leur amour envers la victime, pour évoquer

davantage les problèmes publics, les institutions et les groupes sociaux. Autrement dit, la présence des contributeurs actifs les oblige à considérer l'occurrence d'un point de vue politique et moral. Comme l'avait analysé Luc Boltanski lorsqu'il étudiait les tensions entre les régimes de l'amour et de la justice (Boltanski, 1990), ce changement de registre argumentatif est vécu de façon particulièrement violente par les contributeurs ponctuels qui ont perdu un proche.

Dès lors que certaines conditions sont réunies, qui sont liées à certaines caractéristiques de l'occurrence et à l'interaction avec les contributeurs actifs, les contributeurs ponctuels argumentent davantage en termes publics. Contrairement à ce qu'on postule souvent rapidement, les « multitudes de riverains » sont donc susceptibles de se constituer en public, c'est-à-dire de partager des jugements qui ne limitent pas à l'espace du proche.

Conclusion

Cette enquête sur la formation des publics des dispositifs en ligne diffusant l'information sous la forme d'occurrences – ici des crimes, mais il pourrait tout aussi bien s'agir de mesures de pollution ou d'évaluations d'écoles – apporte deux enseignements principaux. Premièrement, elle montre que dès lors que les organisations de presse mobilisent les technologies numériques pour renoncer à sélectionner les occurrences dignes d'être diffusées, le public ne disparaît pas subitement. D'abord parce qu'un grand nombre d'internautes sélectionnent les occurrences qui les affectent personnellement, et ne les considèrent que sur le registre de la proximité. Mais surtout parce que d'autres façons de faire public, qui sont davantage liées aux médias traditionnels, peuvent encore être observées. Plus précisément, une proportion non négligeable des internautes en viennent à partager des interprétations plus générales et orientées vers la morale ou les problèmes publics – soit parce qu'ils sont eux-mêmes militants, soit parce qu'ils interagissent avec des militants, soit encore parce qu'ils sont touchés par des problèmes qui suscitent leur indignation. D'une certaine manière, ce résultat s'inscrit dans la continuité des recherches contemporaines, qui montrent que les publics en ligne se transforment plus qu'ils changent de nature.

Deuxièmement, cette enquête se présente comme une alternative aux recherches qui envisagent les publics de l'information en ligne uniquement

à partir des mesures d'audience. L'exploitation des inscriptions textuelles laissées par les internautes constitue certes une limitation – il n'est plus alors possible de dire quelque chose du public beaucoup plus large qui consulte l'information sans poster de commentaires – mais elle permet d'accéder à des interprétations et de les traiter quantitativement. La méthode computationnelle proposée dans cet article présente l'originalité de combiner une théorisation *a priori* du matériau textuel (à partir d'un cadre pragmatique ou actantiel) avec la mise en œuvre d'un codage de ce matériau par apprentissage supervisé. Articulée à une procédure de validation sociologique, cette méthode vient réduire le fossé entre les enquêtes quantitatives sur les audiences – qui se situent à une grande échelle d'individus mais échouent à saisir des interprétations – et les études plus qualitatives – qui saisissent des interprétations mais à une échelle très locale.

Comme nous l'avons vu tout au long de cet article, l'exploitation sociologique des traces numériques oblige les chercheurs à ruser devant le manque d'informations entourant les locuteurs et l'opacité des traitements algorithmiques. Mais pour peu que soit défini et mis en œuvre un protocole d'enquête – ce à quoi nous espérons avoir contribué à travers cet article –, ce type d'enquête offre l'opportunité d'étudier la façon dont se forment des publics, en mettant à distance le risque de réification.

Références

Bakshy, E., Messing, S., & Adamic, L. (2015), « Exposure to ideologically diverse news and opinion on Facebook », *Science*, vol.348, n°6239, p.1130-1132.

Barnhurst, K. et D. Mutz (1997), « American Journalism and the Decline of Event-Centered Reporting », *Journal of Communication*, vol.47, n°4, p.27-53.

Benford, Robert D. et David A. Snow (2000), « Framing processes and social movements : An overview and assessment », *Annual review of sociology*, vol.26, n°1, p.611–639.

Bennett, W. L. et Segerberg, A. (2013), *The logic of connective action : Digital media and the personalization of contentious politics*. Cambridge University Press.

Berthaut, J., É. Darras et S. Laurens (2009), « Pourquoi les faits-divers stigmatisent-ils ? » L'hypothèse de la discrimination indirecte, *Réseaux*, vol.5, n°157-158, p.89-124.

Beuscart, J.-S., É. Dagiral et S. Parasie (2016), *Sociologie d'internet*, Paris, Armand Colin.

Beuscart, J.-S., T. Beauvisage et S. Maillard (2012), « La fin de la télévision ? Recomposition et synchronisation des audiences de la télévision de rattrapage », *Réseaux* vol. 5, n°175, p.43-82.

Boczkowski, P. J. (2010), *News at Work. Imitation in an age of information abundance*, Chicago, University of Chicago Press ; ainsi que le livre dirigé par J. Jouët et R. Rieffel (dir.) (2014), *S'informer à l'ère numérique*, Rennes, Presses universitaires de Rennes, coll. « Res publica ».

Boltanski, L., M.-A. Schiltz et Y. Darré (1984), « La dénonciation », *Actes de la recherche en sciences sociales*, vol.51, n°1, p.3-40 ; L. Boltanski, M.-N. Godet (1995), « Messages d'amour sur le téléphone du dimanche », *Politix*, vol.8, n°31, p.30-76.

Boltanski, L. (1990), *L'Amour et la Justice comme compétences*, Paris, Métailié.

Burscher, B., R. Vliegenthart, et C. H. De Vreese (2015), « Using supervised machine learning to code policy issues : Can classifiers generalize across contexts ? », *The ANNALS of the American Academy of Political and Social Science*, vol.659, n°1, p.122-131.

Céfaï, D. et D. Pasquier (dir.) (2003), *Les sens du public*. *Publics politiques, publics médiatiques*, Paris, PUF, CURAPP/CEMS.

Chateauraynaud, F. (2003), *Prospéro. Une technologie littéraire pour les sciences humaines*, Paris, CNRS Éditions, p.47-64.

Chateauraynaud, F., D. Torny (2005), *Les sombres précurseurs. Une sociologie de l'alerte et du risque*, Paris, Ehess.

Claverie, É. (1994), « Procès, Affaire, Cause. Voltaire et l'innovation critique », *Politix*, vol.7, n°26 p.76-85.

Cointet, J.-P., S. Parasie (2018), « Ce que le big data fait à l'analyse sociologique des textes. Un panorama critique des recherches contemporaines », *Revue française de sociologie*, vol.59, n°3, p.533-557.

Costa Vargas, J. H. (2006), *Catching Hell in the city of Angels. Life and meanings of blackness in South Central Los Angeles*, Minneapolis, University of Minnesota Press.

Dewey, J. (1927), *Le public et ses problèmes*, Publications de l'Université de Pau, Farrago/Léo Scheer, traduit par Joëlle Zask

Donzelot, J., C. Mével et A. Wyvekens (2003), *Faire société. La politique de la ville aux États-Unis et en France*, op.cit.,

Flaxman, S., S. Goel et J. M. Rao (2016), « Filter bubbles, echo chambers, and online news consumption », *Public Opinion Quarterly*, vol.80, p.298-320.

Fletcher, R. et R. Kleis Nielsen (2017), « Are news audiences increasingly fragmented? a cross-national comparative analysis of cross-platform news audience fragmentation and duplication », *Journal of Communication*, vol.67, n°4, p.476-498.

Graham, T. et S. Wright (2014), « Discursive equality and everyday talk online : the impact of "superparticipants" », *Journal of Computer-Mediated Communication*, vol.19, n°3, p.625-642.

Hall, Stuart (1994), « Codage/décodage », *Réseaux*, n°68 (1977), p.27-39.

Hillard, D., Purpura, S., & Wilkerson, J. (2008), « Computer-assisted topic classification for mixed-methods social science research », *Journal of Information Technology & Politics*, vol.4, n°4, p.31-46.

Hirschman, Albert O (1995), *Défection et prise de parole. Théorie et applications*, Paris, Fayard, coll. « L'espace du politique », trad. C. Besseyrias (1970).

Katz, Jack (1987), « What makes crime "news" ? », *Media, Culture and Society*, vol.9, p.47-75.

Lazer D., Pentland A. S., Adamic L., et al. (2009), « Computational social science », *Science*, 323, 5915, p.721-722.

Lemieux, C. (2000), *Mauvaise presse. Une sociologie compréhensive du travail journalistique et de ses critiques*, Paris, Métailié.

Liang, Y., Jabr, K., Grant, C., Irvine, J., & Halterman, A. (2018), « New Techniques for Coding Political Events across Languages », *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, p.88-93.

Lim, M. (2012), « Clicks, cabs, and coffee houses : Social media and oppositional movements in Egypt, 2004–2011 », *Journal of communication*, vol.62, n°2, p.231-248.

Marres N. (2017), *Digital sociology : the reinvention of social research*, Cambridge, Polity Press.

Missika, J.-L. (2006), *La fin de la télévision ?*, Paris, Seuil, coll. « La république des idées ».

McCombs, Maxwell E. et Donald L. Shaw (1972), « The agenda-setting function of mass media », *The Public Opinion Quaterly*, vol. 36, n°2, p.176-187.

Molotch, Harvey et Marilyn Lester (1996), « Informer : une conduite délibérée de l'usage stratégique des évènements », *Réseaux*, n°75, p.23-41 (1974).

Morrison, P. A. et I. S. Lowry (1994), « A riot of color : The demographic setting », in M. Baldasare (dir.), *The Los Angeles riots : Lessons for the urban future*, Boulder, Westview, pp.19-46.

Nord, D. P. (2001), *Communities of journalism : A history of American Newspapers*, Chicago, University of Illinois Press.

Parasie, S., Cointet, J.-P. (2012), « La presse en ligne au service de la démocratie locale » *Une analyse morphologique de forums politiques, Revue française de science politique*, vol. 62, n°1, pp. 45-70.

Parasie, Sylvain, Eric Dagiral (2013a), « Data-driven journalism and the public good : "Computer-assisted-reporters" and "programmer-journalists" in Chicago », *New media & society*, vol.15, n°6, p.853-871.

Parasie, Sylvain, Eric Dagiral (2013b), « Des journalistes enfin libérés de leurs sources ? Promesse et réalité du 'journalisme de données' », *Sur le journalisme*, vol.2, n°1, pp. 52-63.

Pariser, E. (2011), *The filter bubble : What the Internet is hiding from you*, London, Penguin.

Prior, Markus (2007), *Post-broadcast democracy : How media choice increases inequality in political involvement and polarizes elections*, Cambridge University Press.

Quéré, L. (2003), « Le public comme forme et comme modalité d'expérience », in D. Céfaï et D. Pasquier (dir.), *Les sens du public*. *Publics politiques, publics médiatiques*, Paris, PUF, CURAPP/CEMS, pp.113-134.

Reid, A. (2014), « How homicide report tells the 'true story' of LA's violent crime" », Journalism.co.uk. Disponible à l'adresse : http://www.journalism.co.uk/news/how-the-homicide-report-tells-the-true-story-of-la-s-violent-crime/s2/a555713/

Roderick, K. (2013), « Homicide Report Gets New Life at the LA Times », *LA Observed*. Disponible à l'adresse : http://www.laobserved.com/archive/2013/03/homicide_report_gets_new.php.

Roshier, R. (1973), « The selection of crime news by the press », in S. Cohen and J. Young (dir.), *The Manufacture of news*, Beverly hills, Sage, pp.28-39.

Scheufele, D. A. (1999), « Framing as a theory of media effects », *Journal of communication*, vol.49, n°1, p.103-122.

Sunstein, C. R. (2002), *Republic.com*, Princeton, Princeton University Press.

Sunstein, C. R. (2018), *# Republic : Divided democracy in the age of social media*, Princeton, Princeton University Press.

Tarde, G. (1989), *L'opinion et la foule*, Paris, PUF (1901).

Venturini, T., D. Cardon et J.-P. Cointet (2014), « Méthodes digitales. Présentation », *Réseaux*, vol.6, n° 188, pp.9-21.

Young, M. L. et A. Hermida (2015), « From Mr. And Mrs. Outlier to central tendancies. Computational journalism and crime reporting at the *Los Angeles Times* », *Digital journalism*, vol.3, n°3, pp.381-397

Zask, J. (2008), « Le public chez Dewey : une union sociale plurielle », *Tracés*, vol.2, n°15, pp.169-189.

Annexe n°1 – évaluation *ex ante* des classifieurs

Pour chaque variable, le logiciel utilise un échantillon des commentaires codés par nous-mêmes, appelé échantillon de test, pour évaluer la qualité du modèle statistique inféré qui servira par la suite à coder l'ensemble du corpus. Ces commentaires sont volontairement exclu de l'échantillon d'apprentissage qui sert effectivement à entraîner le classifieur. Les étiquettes de l'échantillon de test sont ainsi comparées avec celles prédites par le modèle généré au terme de l'apprentissage. Les « tables de confusion » ci-dessous synthétisent les résultats de ce double codage (machine et humain) pour chaque variable. Un ensemble de métriques habituelles sont ensuite été produites, qui évaluent la qualité du codage *ex ante* opéré par la machine.

1. Lien locuteur-victime		machine		Total
		non	oui	
humain	non	34	11	45
	oui	6	74	80
Total		40	85	125

Précision : 0,87
Rappel : 0,92
F score : 0,90
Baseline : 0,64
Précision de score : 0,86

2. Affection vis-à-vis de la victime ou de ses proches		machine		Total
		non	oui	
humain	non	74	8	82
	oui	9	130	139
Total		83	138	221

Précision : 0,94
Rappel : 0,94
F score : 0,94
Baseline : 0,63
Précision de score : 0,92

3. Évaluation morale des individus		machine		Total
		non	oui	
humain	non	46	19	65
	oui	6	42	48

Total	52	61	113

Précision : 0,69
Rappel : 0,87
F score : 0,77
Baseline : 0,58
Précision de score : 0,78

4a Problèmes publics		machine		Total
		non	oui	
humain	non	150	3	153
	oui	15	28	43
Total		165	31	196

Précision : 0,90
Rappel : 0,65
F score : 0,76
Baseline : 0,78
Précision de score : 0,91

4b. Institutions		machine		Total
		non	oui	
humain	non	147	6	153
	oui	20	47	67
Total		167	53	220

Précision : 0,89
Rappel : 0,70
F score : 0,78
Baseline : 0,70
Précision de score : 0,88

4c. Groupes sociaux/ethniques		machine		Total
		non	oui	
humain	non	138	1	139
	oui	10	17	27
Total		148	18	166

Précision : 0,94
Rappel : 0,63
F score : 0,76
Baseline : 0,84

Précision de score : 0,93

Annexe n°2 – évaluation *ex post* des classifieurs

Pour chaque variable, le modèle généré au terme de l'apprentissage a été testé sur un échantillon aléatoire de 300 commentaires. Un nouveau codage humain a été réalisé, qui a ensuite été comparé au codage opéré par la machine. Les « tables de confusion » ci-dessous synthétisent les résultats de ce double codage pour chaque variable. Un ensemble de métriques habituelles sont ensuite été produites, qui évaluent la qualité du codage opéré par la machine. Ce contrôle *ex post* nous a semblé nécessaire de par le caractère volontairement biaisé du corpus d'annotation qui est une conséquence de la nature active de notre procédure d'annotation. De fait, notre évaluation ex post sur un corpus de commentaires entièrement aléatoire (après exclusion du corpus d'apprentissage utilisé) montrent que la performance des classifieurs est très légèrement inférieure que l'évaluation *ex ante*.

1. Lien locuteur-victime		machine		Total
		non	oui	
humain	non	98	42	140
	oui	13	154	167
Total		111	196	307

Précision : 0,78
Rappel : 0,92
F score : 0,85
Baseline : 0,54
Précision de score : 0,82

2. Affection vis-à-vis de la victime ou de ses proches		machine		Total
		non	oui	
humain	Non	101	18	119
	Oui	28	160	188
Total		129	178	307

Précision : 0,9
Rappel : 0,85
F score : 0,87
Baseline : 0,62
Précision de score : 0,85

3. Évaluation morale des individus	machine		Total
	non	oui	

humain	non	88	36	124
	oui	59	124	183
Total		147	160	307

Précision : 0,77
Rappel : 0,68
F score : 0,72
Baseline : 0,6
Précision de score : 0,69

4a Problèmes publics		machine		Total
		non	oui	
humain	non	270	6	276
	oui	13	17	30
Total		283	23	306

Précision : 0,74
Rappel : 0,57
F score : 0,64
Baseline : 0,9
Précision de score : 0,94

4b. Institutions		machine		Total
		non	oui	
humain	non	245	7	252
	oui	18	36	54
Total		263	43	306

Précision : 0,84
Rappel : 0,67
F score : 0,74
Baseline : 0,82
Précision de score : 0,92

4c. Groupes sociaux/ethniques		machine		Total
		non	oui	
humain	non	292	3	295
	oui	4	7	11
Total		296	10	306

Précision : 0,7

Rappel : 0,64
F score : 0,66
Baseline : 0,96
Précision de score : 0,98

Annexe n°3 – Régression linéaire du nombre de commentaires et de la couverture dans l'édition imprimée du *Los Angeles times* selon les caractéristiques de la victime, de l'homicide et du quartier (méthode des moindres carrés ordinaires)

	A	B	C	D
	Couverture dans l'édition imprimée du *Los Angeles Times*	Nombre de commentaires par mois (tous contributeurs)	Nombre de commentaires par mois (contributeurs actifs seulement)	Nombre de commentaires par mois (contributeurs ponctuels seulement)
Âge de la victime (ans)				
0-19	-0.0037	0.4081***	0.1176***	-0.4416*
	(0.0207)	(0.1088)	(0.0308)	(0.2284)
20-24	0.0030	0.0702	0.0403	0.1376
	(0.0206)	(0.1081)	(0.0306)	(0.2268)
25-31	-0.0243	0.0972	0.0294	0.2565
	(0.0202)	(0.1063)	(0.0300)	(0.2229)
32-43	-0.0275	-0.0811	-0.0336	0.5024**
	(0.0203)	(0.1069)	(0.0302)	(0.2244)
44-97	0.0001	-0.3054***	-0.0679**	0.3831*
	(0.0205)	(0.1076)	(0.0304)	(0.2258)
Sexe de la victime				
Femme	0.1129	-0.1109	0.0263	-1.4866
	(0.1056)	(0.5557)	(0.1571)	(1.1657)
Homme	0.0176	-0.1789	0.0053	-0.9258
	(0.1063)	(0.5592)	(0.1581)	(1.1732)
Ethnicité de la victime				
Noire	0.0417	0.3713**	0.0609	-0.2240
	(0.0306)	(0.1613)	(0.0456)	(0.3384)
Hispanique	0.0259	0.2220	0.0796*	-0.0864
	(0.0295)	(0.1550)	(0.0438)	(0.3251)
Asiatique	0.0582	-0.2214	-0.0362	0.2762
	(0.0456)	(0.2399)	(0.0678)	(0.5034)
Blanche	0.0528	0.1262	0.0335	0.1791
	(0.0326)	(0.1719)	(0.0486)	(0.3606)
Autre	-0.0650	-0.3539	-0.0841	0.2216
	(0.1242)	(0.6534)	(0.1847)	(1.3708)
Circonstances de				

l'homicide				
Violence conjugale	-0.0781***	-0.3502**	-0.0794*	0.1260
	(0.0302)	(0.1589)	(0.0449)	(0.3334)
Fusillade au volant	-0.0734***	0.1245	-0.0065	0.3493
	(0.0259)	(0.1368)	(0.0387)	(0.2869)
Bagarre	-0.0624***	-0.1293	-0.0292	-0.3473
	(0.0224)	(0.1181)	(0.0334)	(0.2477)
Implication de la police	0.0918***	0.7587***	0.3902***	-0.3681
	(0.0274)	(0.1446)	(0.0409)	(0.3033)
Fête	0.0702	-0.0640	-0.1376	-0.5746
	(0.0619)	(0.3255)	(0.0920)	(0.6828)
Cambriolage	0.1576***	-0.0687	0.0455	0.3996
	(0.0394)	(0.2082)	(0.0589)	(0.4368)
Fusillade à pied	-0.0535**	0.0078	-0.0814***	0.1509
	(0.0209)	(0.1102)	(0.0312)	(0.2313)
Cause de l'homicide				
Traumatisme	-0.0010	-0.0053	0.0253	-0.6884**
	(0.0316)	(0.1660)	(0.0469)	(0.3482)
Arme à feu	0.0157	0.0020	0.0135	-0.3001
	(0.0191)	(0.1004)	(0.0284)	(0.2106)
Autre	0.0949**	-0.0563	0.0025	-0.2447
	(0.0380)	(0.2001)	(0.0566)	(0.4197)
Poignardé	-0.0098	-0.0571	-0.0419	0.2191
	(0.0245)	(0.1290)	(0.0365)	(0.2706)
Étranglement	-0.0452	0.2328	-0.0220	-0.3680
	(0.0445)	(0.2341)	(0.0662)	(0.4911)
Mention du mot "gang" dans l'article	0.0170**	0.1265***	0.0370***	-0.2104**
	(0.0086)	(0.0454)	(0.0128)	(0.0953)
Pas de couverture dans l'édition imprimée du *Los Angeles Times*	–	-0.2152***	-0.0753***	0.2471*
		(0.0644)	(0.0182)	(0.1351)
Ethnicité majoritaire des habitants du quartier				
Asiatique	-0.0611	0.1699	0.0788	-0.9299**
	(0.0422)	(0.2223)	(0.0629)	(0.4664)
Noire	0.0032	0.1030	-0.0567	0.3914
	(0.0241)	(0.1270)	(0.0359)	(0.2664)
Hispanique	-0.0230	0.0058	-0.0115	0.1751

	(0.0177)	(0.0933)	(0.0264)	(0.1956)
Revenu moyen du quartier (log)	-0.0160 (0.0261)	0.0488 (0.1372)	-0.0142 (0.0388)	0.2563 (0.2877)
Taux d'homicides du quartier (log)	-0.0778*** (0.0242)	0.0521 (0.1278)	0.0268 (0.0361)	-0.2769 (0.2680)
Intercept	0.6506* (0.3368)	0.4707 (1.7716)	0.2837 (0.5009)	2.2773 (3.7166)
R2	0.07	0.09	0.12	0.04
Nombre d'observations : 1,699				