Can a public emerge from a crime map?

Jean-Philippe Cointet, Sylvain Parasie

▶ **To cite this version:**

Jean-Philippe Cointet, Sylvain Parasie. Can a public emerge from a crime map?: A computational analysis of online comments. Réseaux : communication, technologie, société, 2019, n°214-215, 10.3917/res.214.0209 . hal-03406731

HAL Id: hal-03406731
https://hal.science/hal-03406731

Submitted on 28 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CAN A PUBLIC EMERGE FROM A CRIME MAP?

A computational analysis of online comments[1]

Jean-Philippe COINTET
Sylvain PARASIE

La Découverte | "Réseaux"

2019/2 No. 214-215 | pages 209 to 250

[1] Both authors contributed equally to the article. We would like to thank Benjamin Ooghe-Tabanou, who helped develop the database we use here. We would also like to thank Madeleine Akrich, Valérie Beaudouin, Peter Bearman, Bilel Benbouzid, Jean-Samuel Beuscart, Dominique Cardon, Alexandre Mallard, Kevin Mellet, and Alix Rule for their feedback on the various previous draft versions of this article.

DOI: 10.3917/res.214.0209

With the rapid growth of online services provided by news organizations, we increasingly have access to "occurrences", that is, information about often isolated facts or incidents associated with clearly defined spaces, times and actors (see Molotch and Lester, 1996). A crime is committed against a person on a particular day in a particular neighbourhood; a level of air pollution is measured in a particular place over a defined period of time; a particular school achieves a particular success rate on an exam; and so on. All of this information on occurrences is communicated in the form of articles, maps of their distribution across a given area, rankings or graphs. Many mobile applications and websites thus enable individuals to find this factual information, delimited in time and space, which they can browse through freely.

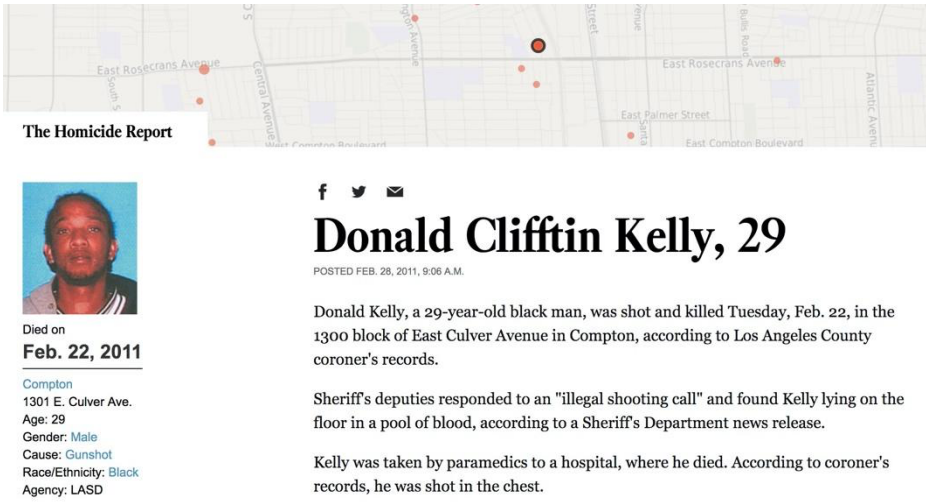
The very nature of the publics that form around these news platforms generates intense debate. Writers and researchers have raised concerns about the shift from individuals having access to only a small number of occurrences carefully selected by news organizations, to a new configuration where they themselves are increasingly able to pick from a wide range of occurrences based on their personal interests. In this new configuration, individuals are said to read about only those occurrences that affect them personally – as residents, relatives or users of collective services –, and to steer away from information on public affairs (Sunstein, 2002, 2018; Prior, 2007). As a result, it is argued, they are no longer be able to form a "public" in the strong sense of the word, that is, a collective of individuals who share interpretations despite being physically distant from one another (Tarde, 1901).

Yet little is known about how individuals aggregate around these news platforms when they provide news in the form of occurrences. This question is particularly challenging, especially considering just how difficult it has always been for sociologists to empirically research the emergence of publics (Quéré, 2003). Recently, many researchers have attempted to measure the public's fragmentation online by quantitatively investigating how individuals converge around specific news content, based on whether or not it aligns with their ideological preferences (see in particular Bakshy *et al.*, 2015; Flaxman *et al.,* 2016; and Fletcher and Nielsen, 2017). However, for anyone wishing to empirically study how individuals converge around a large number of occurrences accessible through online news platforms, these studies present two major limitations. First, they analyse ideological polarization, examining news content only through the prism of its political orientation. Somewhat paradoxically, few social scientists have studied the public of the news that best meets journalistic objectivity standards – factual and standardized information, the production of which involves less

journalistic subjectivity. Second, this stream of research overwhelmingly relies on audience metrics, which does not allow for an understanding of how individuals make sense of the news. As reception studies have shown, individuals who browse the same content can interpret it in very different and even contradictory ways (Hall, 1994). In this article, we therefore consider the extent to which a "public" can emerge around occurrences, that is, not just as a set of news consumers, but as a collective entity that shares common topics of conversation and common interpretations.

Our analysis focuses on "The Homicide Report", a news platform that was launched in 2010 on the *Los Angeles Times* website, providing standardized information on all homicides committed in the Californian metropolis. Rather than covering a small number of homicides that they deemed to be of editorial interest, the journalists decided to stop making any selection and to cover all homicides in a factual and standardized way. For Internet users, this platform consists of a map on which a myriad of dots pinpoint the precise location at which each homicide took place. All these homicides are presented through a set of standardized information about the victim (name, age, gender, ethnicity) and the crime (date, address, location, causes, circumstances). Each homicide has a dedicated webpage, featuring a photograph of the victim and a short article automatically generated with this structured information. The screenshot below presents the murder of Donald Kelly with a red dot circled in black on a map and a set of standardized information (Figure 1). This 29-year-old black man was shot dead in the Compton neighbourhood on 28 February 2011. Every year, an average 750 such occurrences are shared through "The Homicide Report", reaching a large audience, as evidenced by the tens of thousands of comments posted on the platform since its creation.

Figure 1. One of many occurrences: the murder of Donald Kelly – http://homicide.latimes.com/, accessed on 10 April 2019



Source: *Los Angeles Times*.

From our point of view, "The Homicide Report" offers valuable experimental ground to study the emergence of publics around these emerging news platforms. First, it provides a large number of occurrences that have not been selected for their journalistic value and which users can browse individually or compare, using maps or lists generated from factual criteria. Second, each occurrence can be commented on or discussed on the platform itself, in a context where urban violence is a particularly intense topic in the United States. We therefore carried out a quantitative analysis of the comments published on the platform over a seven-year period, applying a textual analysis method that is very rarely used in the social sciences at present. This method is based on a text classification technique, performed using supervised learning algorithms. Through the *Los Angeles Times* platform, we compiled a corpus of 28,828 comments by Internet users, concerning 4,506 homicides committed between February 2010 and December 2016.

In this article, we show that users develop shared interpretations based on the occurrences presented to them. Following a pragmatic sociology approach, we consider the public here not as a stable and fixed reality, but as the process whereby individuals come to share interpretations (Céfaï and Pasquier, 2003). We show that this process involves several ways of forming

a public, which sociology can grasp only by developing new methods.

This article is organized as follows. In the first section, we review the literature to distinguish several ways of forming a public around occurrences. In the second section, we present the methodological choices we made to analyse the emergence of a public through the digital traces offered by the *Los Angeles Times* platform. We then devote the last three sections of the article to the three ways of forming a public identified through the analysis of the digital traces.

HOW DO PUBLICS FORM AROUND OCCURRENCES?

Three models can be distinguished in the literature, each corresponding to a way of building collectives around occurrences. We call these models the "invisible Coliseum", the "multitude of residents" and the "collective of inquiry". Although empirical research on these models is rare, they help guide our study on the *Los Angeles Times* platform.

The "invisible Coliseum"

The first model was described by Gabriel Tarde at the very beginning of the twentieth century, in *L'opinion et la foule* (Tarde, 1901). Discussing the court chronicle, the sociologist was amazed that the account of a single criminal drama could cause "the gaze of countless scattered spectators, an immense and invisible Coliseum", to "converge for weeks on end". Tarde thus emphasized the power of newspapers to get individuals to discuss a single occurrence, even when they are physically separate from one another and have no connection to the people involved.

The model we call the "invisible Coliseum" is linked to the emergence of mass media. It relies on a strict selection of the occurrences that will feature in newspapers, from the extensive number of occurrences that take place in the world. The case of metropolitan newspapers in the United States fits this model well. In the last decades of the nineteenth century, these newspapers stopped covering as much local information as possible and featured only a small number of occurrences, in an attempt to capture the interest of all inhabitants in the metropolis (Nord, 2001). In so doing, they generated a metropolitan public that was interested in and converged around a few occurrences often linked to the institutions and central areas of the city.

To some extent, the coverage of occurrences therefore erased their contextual features. When a crime made the headlines, it was not so important to know the precise location where it had been committed or the precise identities of all the people involved. Several scholars have shown that from the 1960s onwards, news stories thus became an opportunity for journalists to report on social issues, such that the properties of the occurrence itself became secondary in journalistic coverage (Barnhurst and Mutz, 1997).

This model has often been praised for its socializing virtues, that is, its ability to bring together individuals distant from one another to share common concerns. By empirically showing that citizens share a small number of topics of interest, agenda-setting studies have also shown that the media succeed in imposing topics of conversation on citizens (McCombs and Shaw, 1972). But this model has also been widely criticized. Many social scientists have pointed out the disconnect between the value that journalists attribute to a crime and the social reality of that crime at a given time and in a given context. The involvement of a celebrity, or exceptional circumstances, for example, increase the journalistic value of a crime (Roshier, 1973; Berthaut *et al.*, 2009). Data journalists see the occurrence selection process as a bias that they endeavour to correct through computational technology (Parasie and Dagiral, 2013a, 2013b).

The collective of inquiry

John Dewey's work sheds light on another way of forming collectives around occurrences. In *The Public and its Problems* (1927), Dewey stresses the extent to which much of the information found in newspapers is difficult to interpret and to integrate into the course of events:

> "News" signifies something which has just happened, and which is new just because it deviates from the old and regular. But its *meaning* depends upon its relation to what it imports, to what its social consequences are. This import cannot be determined unless the new is placed in relation to the old, to what has happened and been integrated into the course of events. Without coordination and consecutiveness, events are not events, but mere occurrences, intrusions; an event implies that out of which a happening proceeds. (Dewey, 1927, p.139)

Dewey deems that most of the news published by newspapers constitutes "breaches of continuity". According to him, for a true public to emerge, an investigative process has to be set in motion that can lead to the production

of a set of shared judgments. The main challenge, Dewey argues, is for the various individuals who make up this public to set a collective inquiry in motion, so as to formulate public judgments about the problem they face (Zask, 2008). Yet this inquiry focuses primarily on all the fragmented information that reaches it, especially through newspapers. The challenge, for this public, is thus to investigate these occurrences in order to interpret them and put them into series. Interpretive work plays a major role here, and requires significant resources for the public to be able to formulate judgements.

This "collective of inquiry" model differs considerably from the first model. It implies that individuals not only share topics of attention and concern, but also collectively develop new interpretations in light of the multiple occurrences they discover. A set of individuals very different from one another come to investigate a large number of scattered occurrences – be they crimes, accidents or pollution. It is by investigating the link between these multiple occurrences, by identifying explanatory patterns, that these individuals come to form a public and develop shared judgements about their problem. The terms of this process are never set in stone, and significant cognitive resources are mobilized.

Several sociological traditions have investigated this second way of forming a public. This includes for instance studies on the "affairs" through which collectives are formed, and the indignation of these collectives, based on their interpretation of isolated or multiple occurrences (Claverie, 1994); or studies analysing the way in which collectives identify signals to alert public opinion to sociotechnical risks (Chateauraynaud and Torny, 2005). More broadly, the sociology of collective action has highlighted social movements' ability to develop cognitive frameworks to interpret the numerous occurrences shared by the media in particular (Benford and Snow, 2000; Scheufele, 1999). Recent studies have suggested that digital technology, and social media in particular, offer new possibilities for interpreting these occurrences and allow individuals to identify social problems and organize (Bennett and Segerberg, 2013; Lim, 2012).

The multitude of residents

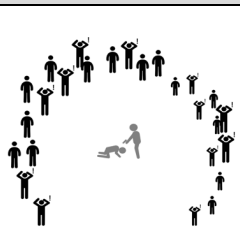
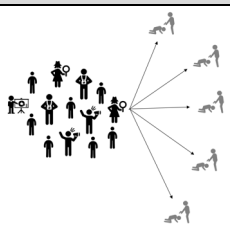
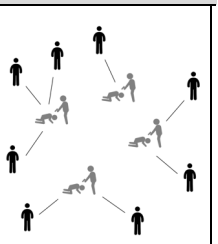
The third way of forming collectives around occurrences hinges on individuals' proximity to these occurrences. This proximity is often geographical, as in the case of the first metropolitan newspapers that appeared in the United States in the early nineteenth century. These newspapers featured a large amount of heterogeneous information, always

associated with a specific location (Nord, 2001). Here, an occurrence appeals to the individual's interest as a resident, a relative, a consumer or a user of collective services. Whatever the nature of the proximity at play, it results in the coexistence of a large number of very small collectives, which aggregate around each occurrence.

In this model, occurrences have value only if they are associated with a precise location. They do not have to be selected or even editorialized. They rarely lead individuals to form public judgments, or even really to converse with one another. The occurrences are not primarily intended to create a collective interpretation, but rather to help individual consumers to make an informed choice and find the right school or the right neighbourhood for example.

This "multitude of residents" model is sometimes presented as a foil, associated with the crumbling or dissolution of the public. It is thus used to describe media forms that predate mass media, or to point out the risk of dissolution of the public space allegedly associated with the rise of online media. It is often referred to in debates surrounding the rise of Internet and the transformation of the media (Missika, 2006). US jurist Cass Sunstein, for example, had some success in sounding the alarm about the risks associated with the possibilities of online news personalization (Sunstein, 2002). More recently, the debates surrounding the "filter bubbles" supposedly produced by social networking sites have once again brought this risk to the fore (Pariser, 2011), with the case being made that once people are no longer compelled to take an interest in information from which they are removed, they no longer share common concerns.

Figure 2: Three ways of forming publics around occurrences

	Invisible Coliseum	Collective of inquiry	Multitude of residents
			
Number of occurrences	A few	A large number	A very large number
Form of the collectives	Large groups of individuals physically distant from one another	Grouping of heterogeneous actors (citizens, activists, researchers, journalists, etc.)	Numerous small groupings of close individuals
Driver of the collectives	Shared topics of indignation or conversation	The development of public judgements with a view to resolving a problem	Proximity link to a neighbourhood, a family, friends, etc.
Intermediaries	Journalists and press organizations	Media, activists, researchers	Private companies and public institutions
Criticisms	Journalists select arbitrary or sensational occurrences	Difficulty of making a public emerge	Dissolution of public space

These three models each represent a way of forming collectives around multiple and fragmented occurrences (Figure 2). They have rarely been properly studied, and little is known about how publics are empirically formed – particularly in digital contexts. More importantly, the online platforms associated with data journalism have several particularities: they give access to a very large number of occurrences; these are processed in a standardized way; and they can be compared in several ways using algorithms. In the next sections of this article, we will see that these online platforms combine the three ways of forming publics that we have identified.

Following the formation of a public through its digital traces

There are high expectations surrounding the sociological use of digital traces. In particular, this "digital sociology" is expected to reconcile the depth of qualitative analysis with the breadth of quantitative analysis (Lazer *et al.*, 2009). However, most researchers have pointed out that a set of new

difficulties has emerged with such methods, which are therefore currently far from stabilized (Marres, 2017; Venturini, Cardon and Cointet, 2014; Cointet and Parasie, 2018). We therefore set out to study the formation of a public based on the textual traces offered by the *Los Angeles Times* platform. Leaving the well-trodden path of conventional sociological methods, we embarked on a journey where experimentation is the norm.

An original platform

When the *Los Angeles Times* journalists launched the "Homicide Report" platform in 2010, their intention was to break with the "invisible Coliseum" model presented above. They criticized the tendency of journalists, including at the *Los Angeles Times*, to massively cover a small number of murders, at the expense of the vast majority of homicides considered to lack editorial value (Young and Hermida, 2015). Megan Garvey, the editor-in-chief of the "Homicide Report", explained that the coverage of homicides was both sensationalist and racially biased:

> The white teenage girl who was killed – which is the outlier, the exception to the rule – gets a lot of attention (...) Or a mass shooting. But the people who are getting killed day-in, day-out, the 17- to 22-year-old black male living in a poor neighborhood, those homicides had gotten to the point, with constraints in print and everything else, where they were not newsworthy. (Reid, 2014)

A few journalists took issue with the fact that their newspaper covered only 10% of the murders committed annually in the Californian metropolis. They thus set out to cover all homicides, first through a blog and then through a database populated by the Los Angeles Medical Examiner's Office and the Los Angeles Police Department. The search for comprehensiveness went hand in hand with the desire to treat all homicides in the same way, using a set of standardized information provided by public authorities regarding the victim, the place of the crime, and the causes and circumstances of the murder.

By providing exhaustive and standardized coverage of the murders committed in Los Angeles, these journalists thus sought to break away from the "invisible Coliseum" model. In their public statements, they referred to two different types of public. The first is similar to what we have called the "multitude of residents", focused on capturing the interest of those who live in the neighbourhoods where the murders take place. As the data journalist managing the project stated, "that is something that could be of interest to people who care about what happens near they live" (quoted by Young and

Hermida, 2015). They referred specifically to the families of the victims, who suffered from "their sons' deaths [never being] covered by the press" (Jill Leovy, founder of "The Homicide Report", quoted by Roderick, 2013). The second type of public to which they referred bears greater resemblance to an "collective of inquiry". They hoped that it would "give readers a much more real view of who is dying" (*ibid.*), and allow them to better understand the causes of urban violence. Regarding the possibility given to Internet users to comment on each homicide on the platform, the journalists justified this on the grounds that it would enable the victims' friends and families to pay tribute to them, as well as allowing anyone who wished to do so to discuss the causes of the violence and police intervention.

We wished to take advantage of the large volume of comments published on the platform – 28,364 comments posted between January 2010 and December 2016 – to study the way a public emerges around occurrences. Aside from the opportunity afforded by this corpus, we soon encountered several challenges.

An opportunity and obstacles

The digital traces on this platform afford an opportunity to capture the way in which a public emerges in spite of the multiple and fragmented nature of occurrences, for several reasons. First, these traces afford access to very rich textual material, which can be linked precisely to each occurrence. The discursive dimension is central to the formation of a public, and we can capture here the way that people speak out about a murder to express their pain, offer explanations, share or challenge the opinion of other Internet users, etc. Second, this textual material has the benefit of not having been elicited by a sociologist, contrary to what would be obtained from a questionnaire. Individuals express themselves without the researcher being able to impose his or her own categories first. Finally, these traces can be expected to afford both the depth of a qualitative study of the public and the breadth of a quantitative survey. Depth is provided by the wealth of textual material, while breadth stems from the fact that all the homicides that have occurred over the last seven years appear on the platform, and are likely to be commented on.

We therefore built a new database, based on three sets of data collected on the platform using dedicated scripts. These three sets of data relate to: (1) the information provided by the platform about each victim (name, age, gender and ethnicity); (2) the information shared by the platform regarding each homicide (date, place, causes, circumstances, crime scene); (3) all the

comments posted on the platform in relation to particular homicides (name of each comment's author, date of publication, text).

However, these data presented us with three difficulties, linked to the conditions of their production, the limited information available about the authors, and the nature of the textual material. Consider briefly these difficulties.

First, with these data, we were entirely reliant on the categories created by the *Los Angeles Times* journalists to describe homicides and their victims. These are the result of editorial decisions made on the basis of information provided by public authorities. Since our goal here was not to analyse crime in the Californian metropolis, the accuracy with which the data on the platform reflected the reality of the homicides committed was of little importance[2]. As our aim was to capture the way in which users converged around these occurrences, all we needed to access was the representation of the crime to which they had access. We were nevertheless entirely reliant on the way in which the platform solicited and managed Internet users' participation – particularly the way in which the editorial staff moderated comments upstream.

Second, the data extracted from the platform tell us little about the authors of the comments. They include each contributor's username on the platform, but we know nothing of their civil status, profession, socio-economic status, family situation or political preferences. In short, this is a major constraint for the analysis of the formation of a public.

Finally, while we collected a large volume of textual material, it presented several analytical challenges. Most of it was in English, but often included slang and cultural references that partially escaped us. Above all, we needed to develop a method to capture interpretive actions implemented by Internet users from this relatively inert material.

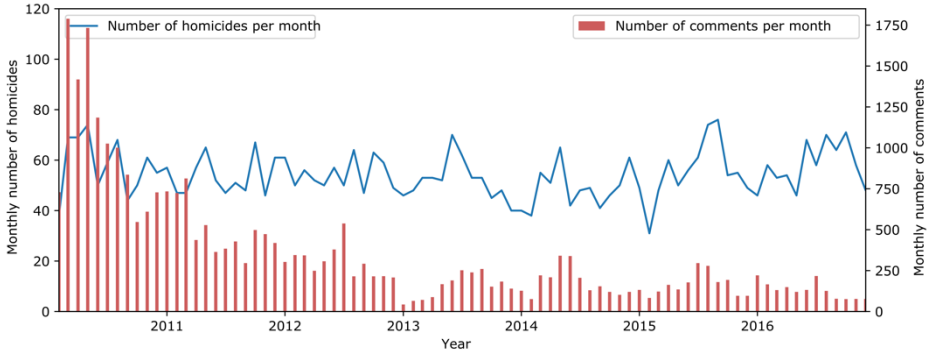
Let us now turn to the solutions we implemented to alleviate these difficulties concerning the characteristics of contributors and the processing of the textual material.

[2] However, the fact that the journalists prioritized information from forensic doctors, which they then cross-checked with that from the police, guarantees good quality information.

Who are the comment authors?

While we initially had little information about the comment authors, exploring the data collected afforded greater insight. As with most online platforms (see Beuscart, Dagiral and Parasie, 2016: 99-102), participation on the "Homicide Report" platform varies widely. First, as Figure 3 shows, it is unevenly distributed in time, whereas the number of homicides had been relatively stable over the past seven years. While it was widely used between 2010 and 2013, the platform received fewer comments after that date, probably due to the rise of social networks.

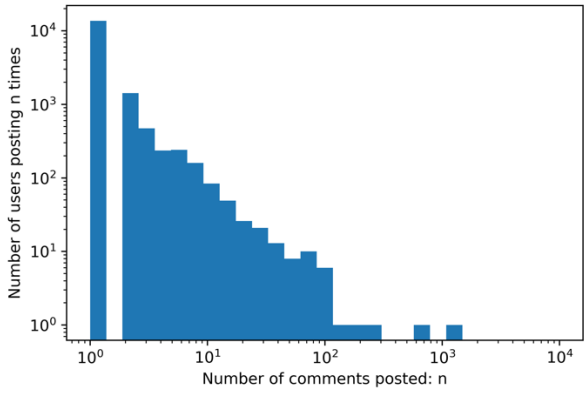
Figure 3. Distribution of homicides and comments on "The Homicide Report" (2010-2016)



Moreover, participation differed considerably from one contributor to another (Figure 4). Out of a total of 16,147 contributors[3], 83% posted just one comment, while 1.3% of contributors posted more than ten comments.

[3] It is by no means certain that each pseudonym corresponds to a single individual. However, two points allow us to assume that the difference between the number of comment authors and the number of pseudonyms is not very significant. First, some comment authors identified themselves to others using their pseudonym. Second, posters who did not comment very often signalled their closeness to the deceased.

Figure 4. Distribution of contributors according to their number of comments



We thus distinguished between two populations of comment authors, based on the number of homicides on which they commented: "superposters" commented on at least ten different homicides, whereas "occasional posters" commented on a smaller number of occurrences. The "superposter" population is comprised of 76 authors, who alone account for 16% of all comments[4]. Studying their publications, it became apparent that these "superposters" were indeed pursuing a specific political agenda, with some systematically defending the actions of the police while others denounced police violence. *Syscom3* was by far the most active superposter, with 1,142 comments published about 449 homicides. He systematically denounced gangs' responsibility for urban violence, often "blaming the victim" when he suspected that they were affiliated with a gang. *Jag*, on the contrary, regularly spoke out against racial discrimination by the Los Angeles Police Department. During this period, he wrote 350 comments regarding 186 homicides.

How to interpret statements

[4] In online discussion spaces, the most active superposters often play a structuring role (Graham and Wright, 2014).

In order to grasp how users converge around occurrences to form a public, our analysis of the textual material must contend with several constraints. First, this material should not be considered as a set of inert *utterances,* but rather as a set of *statements*. In other words, the analysis should seek to capture the movement through which comment authors come to interpret the occurrence, in connection with the other contributors. In line with earlier work (Parasie and Cointet, 2012), we rely here on the theoretical framework developed by Luc Boltanski in his seminal work on the analysis of public speaking (Boltanski, Schiltz and Darré, 1984). Drawing on semiotics, he conceptualized speaking as the construction of relationships between several "actants". In this case, the actants are the author of the comment; the murder victim; the culprits; and the public whose compassion is sought or that is called upon to witness an injustice. We thus conceptualize any comment posted on the "Homicide Report" as a way of connecting these actants to one another.

But the analysis should also not blindly rely on algorithms; sociologists must take responsibility for the way in which they interpret them. We thus spent considerable time familiarizing ourselves with the algorithms' particularities in order to define suitable criteria for coding the comments. The following six criteria are binary variables guided by the pragmatic framework outlined in the previous paragraph:

1. **[*Contributor-victim relationship*]** This variable relates to the relationship between the comment author and the victim. In their comment, the author mentions a personal link with the person murdered (or not). This link can be conveyed in various forms, depending on whether the author addresses the victim directly ("you were special"), manifests a family or friendship tie ("he was a dear friend"), recalls memories shared with the victim ("we had so many good laughs"), or shows acquaintance in another way ("I lost someone very special").
2. [***Affection for the victim or their relatives***] This variable corresponds to situations where the comment author shows their affection towards the deceased ("rest in peace", "you will not be forgotten", "You're Forever in our Hearts"), or towards the relatives of the deceased by conveying their condolences ("my prayers go to the family").

3. **[*Moral evaluation of individuals*]** This variable identifies the presence of a judgement about the moral value of the victim, their relatives, or the suspects. It may be a positive judgment ("he was a loving and devoted 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") or a negative judgment ("this guy was evil"; "He was a known notorious gangster"; "a 15 year old punk"). In the latter case, the comment author simply assigns individual responsibility, blaming the victim or other individuals (a parent, a police officer, etc.) for their behaviour. Here, the homicide is exclusively considered as the consequence of individual choices.

4. **[*Search for collective responsibility*]** The last variable corresponds to situations where the comment author identifies more general responsibility, involving collective entities beyond the individuals involved in the homicide. We organized this variable into three different sub-variables, capturing three distinct forms of generalization:

    a. **[*Public issues*]** This sub-variable identifies comments which associate the homicide with one or several more general problems (urban violence, gang culture, police brutality, the school system crisis, etc.). The comment author does not necessarily assign blame or put forward solutions, but considers the particular occurrence as part of a series of occurrences, reflecting a more general problem ("with all the violence in our society", "these killing fields neighborhoods", "plenty of murders like this", "We as a society have to come to accept that certain young teens are damaged goods").

    b. **[*Institutions*]** This sub-variable identifies comments in which the responsibility of an institution is invoked (the government, the courts, the police, the school system, churches, etc.). These institutions can either be denounced as the source of the violence ("the failure of the religious leaders") or identified as a possible solution ("start DEMANDING action by your elected representatives").

    c. **[*Social/ethnic* groups]** This third sub-variable identifies comments in which social or ethnic groups are mentioned ("black on black, latino on latino crime", "low income residents", "You white people are funny as hell").

Categorization through supervised learning

Due to the variety of objects they mobilize and the multiple individual styles of writing involved, each of these variables refers to heterogeneous lexical and syntactic forms. This made it very difficult to follow a strictly lexicometric approach, since it is virtually impossible to list all the expressions corresponding to each variable. However, though more satisfactory in principle, the resources required to perform strictly human

coding were too great[5]. We therefore opted for supervised learning algorithms to categorize all messages on the platform according to these different variables.

Methods to classify a textual corpus with supervised learning have so far been used very little in the social sciences (Hillard *et al.*, 2008; Burscher *et al.*, 2015). There is no stabilized process for assessing the quality, from a social science perspective, of the categorization of a textual corpus through supervised learning. Assessing its quality is especially tricky, for machine learning algorithms can be somewhat opaque, insofar as it is often impossible to clearly identify the criteria used by the classifier to run the algorithms. The learning algorithm we used was no exception, as it was a neural network. In practical terms, we used the *Prodigy* software[6], which has the benefit of offering both an annotation interface and an inference engine to build a text classifier. *Prodigy*, which is already used by information science and political science scholars (Liang *et al.*, 2018), combines a linguistic analysis module (including morphosyntactic analysis, semantic vectors, and a semantic parser), a user interface and an active learning module[7] coupled with the interface to allow for learning a category on the corpus.

We trained a different classifier for each of our six variables, observing the following three steps:

1. We first drew a list of about ten expressions that we thought were likely to be good markers for each category (for example, the terms "nephew" or "mother" regarding the contributor-victim relationship). *Prodigy* subsequently drew on these lists for the selection of comments to submit to the annotator, particularly at the start of the process. This sample was not random, as it was designed to enable the software to learn more quickly and efficiently to identify comments that fit any one of the variables.
2. We – the two authors of this article – thus manually coded about 800 unique comments for each variable (which allowed us to generate inter-

[5] The 28,364 comments in our corpus count nearly 1.8 million words and 9.4 million characters (each comment counts an average 62 words). By comparison, the most widely distributed edition of the Bible contains less than 0.8 million words and less than 3.5 million characters.

[6] https://support.prodi.gy/, accessed on 10 April 2019.

[7] Active learning refers to the way in which the learning corpus is deliberately biased so as to provide the neural network with examples for which the annotation is likely to make the classifier converge as quickly as possible in its task. Active learning particularly allows for balancing out the distribution of the examples provided to the machine when a category of documents is rare in a corpus.

coder confidence-building measurements, see Appendix 1). Based on this human coding, the software then inferred a classifier using 75% of the tagged comments as a training set and 25% as a testing set. We are then able to produce a confusion matrix and measure the classifier accuracy, recall and associated F-score. As these measurements were satisfactory (for each of the categories learned, *Prodigy* calculated an F-score greater than 0.75 see Appendix 1), we applied each model to the entire corpus to build a coding of the whole corpus of comments.

3. To capitalize on the possibilities offered by the software, in addition to the classifier performance measurements calculated on an inherently biased evaluation sample, we also assessed the quality of the classifier on a third set of comments that were drawn randomly from the corpus (after excluding comments from the training set of course). We thus carried out this ex-post evaluation on 300 random comments in the corpus and compared the categorizations produced by the neural network with those performed manually (see Appendix 2).

Five sets of comments

Let us examine the main forms of statements found throughout the corpus. First, the corpus appears to be split into two main groups: comments in which the author signals a personal tie with the deceased (58% of the corpus); and comments in which no tie between the author and the deceased is expressed (42% of the corpus).

Within the first group, one set of comments is characterized by the author expressing their affection for the deceased (and/or their relatives), but without taking a stand about their moral value. These are "love-centred tributes" (32.6% of the corpus), in which the author expresses unconditional love for the victim. The moral evaluation of the victim is not relevant here, as in the comment below:

> Olga, I can't believe you are gone so quick, ... I will always cheris our wonderful memories in school .. .. .. i love you and will miss you always.. you'll forever be missed .. may god bless ur kids and ur family in this hard times ..... May u rest in heavenly peace .....
> Always
> Letty .... [Letty M., 14 May 2010; homicide of Olga Martinez]

Among the comments expressing a personal tie with the deceased, we can then distinguish "moral-centred tributes" (22% of the corpus). These are comments focused on the moral qualities of the deceased. Here the author not only shows their love for the deceased, but also emphasizes their

qualities – he was a "wonderful son", a "loving father", a "strong woman", etc.:

> One of the best young man I ever known a good father and a wonderful brother a lovin son Sean never meet a person he didn't like his heart was as large as his love for all who ever meet him our hearts are broken our lost is great help us fine justice so we can have closure and peace a life taken from us to soon. [Sherree, 9 August 2015; homicide of Sean Sylvester]

In the second group of the corpus, where the author does not signal any tie with the victim, we identified three distinct sets of comments. First, in "distant tributes" (15.2% of the corpus), the contributors express their affection for the deceased and their relatives, sympathizing with their pain without knowing them personally:

> Christopher sounds like a wonderful, promising gifted and giving human being. So sad for his family. I hope those responsive for this senseless taking of his life will be prosecuted to the full extent of the law. My prayers are with Christopher's family every day. Of there is a fund in Christopher's memory, please let me know. [Linda Willson, 17 June 2015; homicide of Christopher Jermaine Handy]

Another set consists of "distant moral evaluations of individuals" (14.3% of the corpus). Here, the authors refer to the moral value of the victim, of their family or friends, or of the individuals involved in the homicide. Since the authors did not know the victim personally, their moral evaluation focuses on their activities – were they a gang member? Were they involved in illegal activities? Or, on the contrary, were they an innocent victim who was in the wrong place at the wrong time? These comments have the particularity of not linking a murder to a set of systemic elements (such as poverty in certain neighbourhoods, social policies or discrimination against minorities), focusing solely on moral explanations.

> ".... Robert was a Great person besides the Gangbanging, he was a Highly Respected person & a great person."
> In what type of culture? In the normal world where people don't commit violent crimes, his behavior is seen as being sociopathic. A respectable person who has a family stays away from this lifestyle so as to provide for them.
> ".... & for ppl to just to talk so negative about him & his Family is just not Right."
> So what are we supposed to say about an individual who goes on a shootout in a car in a residential neighborhood?
> ".... & no Robert didnt own a Gun."

> And how do you explain he had a weapon on him and it was loaded?
> "…. Noone is perfect …"
> And only a few people adopt a violent lifestyle and act like its normal. That statement alone is a common phrase used by gang enablers to justify the indefensible.
> "…. instead of trying to play God."
> It wasnt any of us who decided to shoot at another person AND PLAY GOD!!!!! [Syscom3, 29 October 2010; homicide of Robert Earl Gipson]

The comment above exemplifies this type of "distant moral evaluation". The author attacks the victim's relatives who attributed a set of personal qualities to him, pointing out the profoundly immoral nature of the life the victim led. The interpretation provided here is strictly individual: the deceased had chosen a "violent lifestyle", which made him responsible for his own death and for all the harm caused.

Finally, the last set of comments encompasses "searches for collective responsibility" (14.4% of comments). We define these as comments that show no connection between the author and the victim, and which mention collective entities – social groups, institutions – or consider the occurrence as a public problem and not just a problem of individual morality. These comments are often much longer and less directly related to the details of the particular homicide with which they are associated:

> I don't want any more of my tax dollars going to these failed social programs. I would love to see welfare turned into "workfare", you show up and work for one day doing whatever job the city needs at the time, sweeping streets, painting over graffiti, picking up trash, you name it, and at the end of the day you get a day's pay, come back the next day and do it again.
> The other thing I would love to see is our prisoners get put to work. First by farming to feed themselves, next by doing all the jobs that illegal's are hired to do. Kill two birds in one stone, the prisoners pay for their stay in prison by producing a product and learning work ethic, and provide labor at a reduced cost to farmers and businesses.
> I have a strong suspicion that this will not be posted due to it's non-politically correctness. [*Eye Opener,* 31 December 2010; homicide of Cesar Guerrero]

These five groups account for 87.2% of the comments published on "The Homicide Report" between 2010 and 2016 (some comments may naturally fall outside these major groupings, which partially overlap, and combine discourses of closeness and search for collective responsibility, for example). They outline specific – and conflicting – ways of giving meaning to the multiple occurrences shared on the platform. Based on the identification of

these forms of statements, we now shed light on the coexistence of three forms of public, each of which emerged following distinct processes. Let us study each of these in turn.

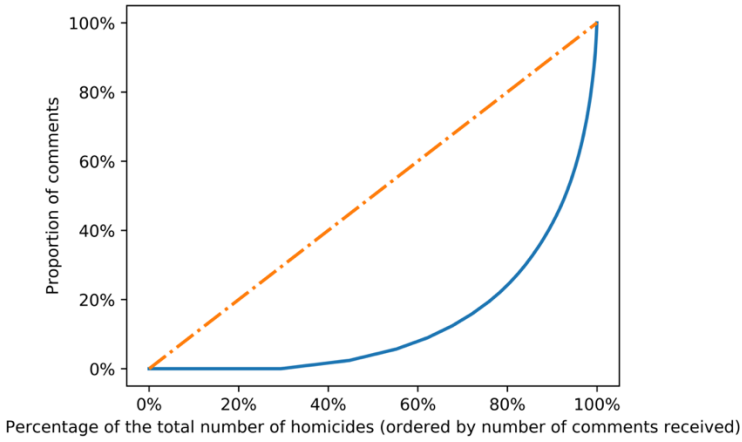
Gathering around a minority of chosen occurrences

A first way of "forming a public", characteristic of mass media, consists in gathering a large number of individuals around a small number of occurrences selected for their editorial value. Building on the work of Gabriel Tarde, this is what we call the "invisible Coliseum". As we will now see, the users of the *Los Angeles Times* platform aggregate in a way that bears resemblances with this type of public formation. Even though they can access any occurrence, contributors focus their attention on a small number of occurrences, which are selected according to a set of shared criteria.

The chart below shows the distribution of comments across all homicides committed over the period (Figure 5). If all occurrences were to receive the same number of comments, the distribution would follow the dotted line. Instead, it forms the black line, which indicates a high concentration on a small number of occurrences – 80% of comments are on less than one quarter of the homicides. Certainly, the selection made by users who published comments is not as restricted as that made by journalists in the print edition of the *Los Angeles Times*. We calculated that 100% of the newspaper's homicide stories were about only 10% of the homicides that actually occurred[8]. But even though it is slightly less pronounced, a process of selection of occurrences by Internet users is indeed at play.

Figure 5. Distribution of the number of comments received by homicides

[8] On the LATimes.com website, we manually checked wether or not each of the homicides committed over the period had been the subject of an article in the newspaper.



This selection, however, is not just the result of independent individual choices. On the contrary, it is based on a set of relatively shared criteria. This is another similarity with the way traditional media publics converge around occurrences: journalists and Internet users may use profoundly different criteria, but their selection criteria are consistent.

Let us first look at the criteria used by *Los Angeles Times* journalists to cover a homicide in the print edition of the newspaper. A statistical regression over the first three years of our corpus (see Appendix 3, Column A) indicates that they preferred to cover murders that occurred in neighbourhoods considered safer, and involving older victims. They systematically excluded homicides that took place in the street, and gave greater coverage to those that occurred following a burglary or involved police forces. These results support research conducted in the United States over several decades (Roshier, 1973; Katz, 1987): journalists prioritize exceptional occurrences that are not related to gun violence, which mainly affects young black and Hispanic men in poor neighbourhoods[9].

The homicides that draw comments from Internet users are very different from those of interest to journalists. Our statistical regression (see Appendix 3, Column B) indicates that Internet users comment primarily on homicides that take place in poor neighbourhoods, and involve young black men killed by firearms. Thus they are not interested in the murders of older people in wealthier neighbourhoods. Internet users are therefore primarily interested in those incidents that are systematically neglected by traditional media, and

[9] Over the period studied, homicides involving young black and hispanic males killed by firearms accounted for 65% of all homicides.

which make up the overwhelming majority of homicides committed in the Californian metropolis – those affecting young black men, often in connection with gangs.

Journalists' and Internet users' selection processes, however, are not entirely different. First, they share a common criterion – the police's involvement in the homicide[10] –, even if Internet users give it far greater importance. Second, a homicide being covered by the newspaper increases its chances of being commented on by Internet users.

The rapid growth of this type of platform therefore does not signal the disappearance of the forms of emergence of publics associated with mass media. Several current studies even show that these ways of converging around cultural content are not systematically disrupted by online technology (Beuscart, Beauvisage and Maillard, 2012). Where journalists no longer filter occurrences, Internet users themselves implement an occurrence selection process. The "invisible Coliseum" is still relevant today, even if Internet users now have greater influence over the criteria governing the selection of occurrences.

HOW DOES COLLECTIVE INQUIRY MATERIALIZE?

A second type of public formation involves individuals who investigate occurrences in order to find solutions to a problem that affects them collectively. Here, the issue is urban violence, which affects most major U.S. cities, especially Los Angeles, where 700 people are murdered every year. In the United States, this problem informs structured debates – around the social, economic and racial relegation of certain neighbourhoods, the relationship to violence in those neighbourhoods, the role of the police in dealing with minorities, etc. – and is addressed by specific forms of urban policy (Donzelot *et al.,* 2003: 323-359). Thus the platform "The Homicide Report" is also visited by individuals who are interested in the occurrences not because they affect them personally, but rather because they provide an opportunity to discuss the problem of urban violence. We will see that these individuals compare a very large number of homicides, seeking to collectively identify more general explanatory frameworks for the problem of violence based on a cognitive and political repertoire.

Let us consider the visitors we call "superposters", who comment on many

[10] Over the period, 7.2% of homicides involved the police.

different homicides. Applying the threshold of 10 homicides commented on, we have a population of 76 authors (whose aliases we checked manually to ensure that they did not on face value refer to several contributors) who alone account for 16% of the comments. These posters took an interest in many homicides; *Syscom3*, for example, published over 1,000 comments about 449 homicides. It is possible that these superposters were seeking to take advantage of the possibilities of comparison offered by the platform, to try to rapidly identify interesting murders. In addition to the considerable number of homicides they commented on, these contributors also stand out from other users of the platform through their discourse. Not only do they demonstrate no personal connection with the victim, but they refer almost exclusively to entities removed from the deceased, as they talk mainly about institutions, public problems, social and ethnic groups, and US society. The chart below shows that superposters employ very different forms of argumentation from occasional posters (Figure 6).

Figure 6. Radar chart of participation by contributor type

Note: the blue polygon outlines the discursive profile of superposters. It links six points corresponding to the proportions of their comments that were coded in each variable. Thus, 30% of comments from superposters mention institutions, while only 10% of comments from occasional posters include such references.

Fisher exact test: *p < .05; **p < .01; ***p < .001

It thus appears that comments by superposters fall more often within the "search for collective responsibility" and "distant moral evaluation" groups, while those of local contributors fit within the "love-centred tributes" and "moral-centred tributes" groups.

The analysis of a sample of comments showed that these superposters were often pursuing a specific political agenda. This is the case of *Syscom3*, whose rhetoric consists in accusing the victim of being responsible for their fate by having chosen an irresponsible, immoral and violent existence:

> Why should anyone like this guy? He's a primary reason these neighborhoods are falling to pieces. He's like the harbinger of doom and crime.
> I would expect his family to say the usual things about him. But for the rest of you why? He went looking to start trouble and he knew exactly what he was doing. He knew there would be a violent reaction to the vandalism he was doing.
> How many 10's of thousands of dollars was spent cleaning up his mess, that he did upon other peoples property? This state is bankrupt and we get to spend money on his thoughtless activities!
> heavan, do you support the activities of miscreants and barbarians like this guy? Why? Are you so immersed in the violent and sociopathic world he was a part of, that to you, this is normal and acceptable behavior? [Syscom3, 15 April 2010; homicide of Jose Castillo]

On the opposite end of the political spectrum, *Jag* commented to point out the Los Angeles Police Department's discrimination against minorities:

> The under privilege people, which are for the most part minorities, are the ones who get the short end of the stick. Many have to take deals in order to avoid serving more time for a crime they did not commit. I avoid playing the race card, but when a white homeless man gets killed by cops in fullerton, some cops are actually held accountable. I'm all for justice, but when a cop kills a hispanic or black justice is not served. Syscom3, you like stats and I was wondering if you happen to have the demographic breakdown of people who are charged with a crime and are actually found guilty. I'm almost certain the White people get more of a break than minorities. It's sad but true. [Jag, 17 November 2012; homicide of Amondo Casillas]

As this comment shows, some superposters address each other directly. Since they have been on the platform for a long time, a new homicide is an opportunity for them to pursue a discussion that has been ongoing for months or even years. From one occurrence to the next, there is a high degree of continuity in their arguments. They defend the same clear-cut

positions, which oppose “cop supporters” and “police bashers”. Evidently, these positions are closely associated with forms of activism that originate outside the platform. This “collective of inquiry” fits well within the sociological tradition that emphasizes the ability of social movements to develop “cognitive frameworks” offering coherent ways of interpreting disparate occurrences (Benford and Snow, 2000).

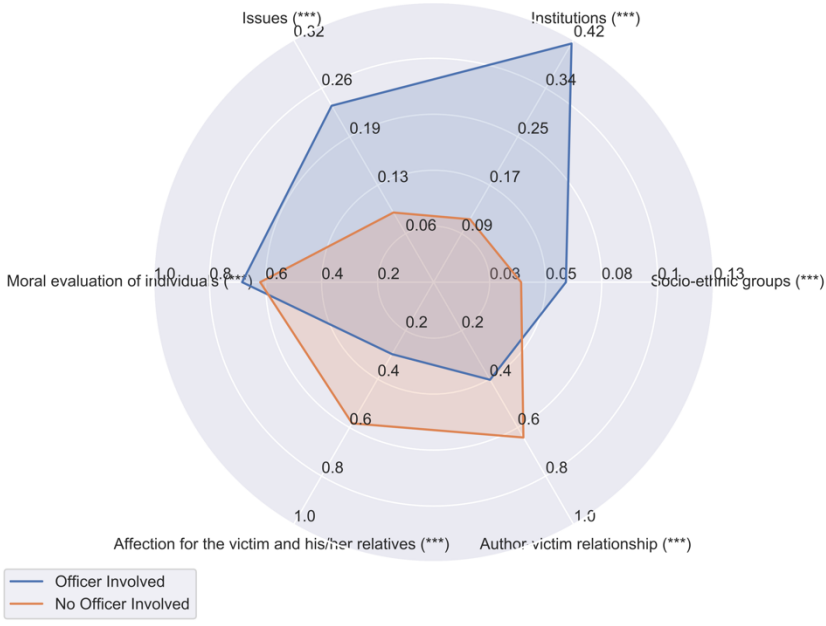
When a news organization stops selecting the occurrences it shows its public, a second way of forming publics emerges which, although very different from the first, has also existed for a long time.

UNDER WHAT CONDITIONS DO RESIDENTS FORM A PUBLIC?

A final way of forming a public, which we have called the “multitude of residents”, is usually associated with the dissolution of public space. The case is made that as individuals come to take an interest only in occurrences that affect them personally, they no longer share the same concerns and lose sight of the common good. This argument, which has never been empirically tested, must be qualified. The case of the “Homicide Report” platform shows that under certain conditions, Internet users can come to share interpretations oriented towards the public good, even as they aggregate around occurrences that affect them personally. These conditions relate to both the particularities of the homicide and the interactions between superposters and occasional posters.

Let us consider occasional posters only, in other words contributors who commented on fewer than ten homicides, 67% of whom expressed a personal relationship with the victim (Figure 6). First, it appears that when the occurrence presented certain characteristics, these occasional posters commented differently, providing more interpretations that involved institutions and social or ethnic groups, referring for example to public problems. This was the case for homicides that directly involved the police – in other words, when a police officer was responsible for the victim’s death. The chart below (Figure 7) shows that where the victim was beaten to death or shot by the police, occasional posters’ comments were more geared towards a “search for collective responsibility”. They mentioned institutions far more and interpreted the homicide as a public problem, with greater references to skin colour and discrimination issues. As ethnographic research conducted in the central areas of Los Angeles suggests, we can assume that these interpretations are informed by a set of norms shared by the black population of these neighbourhoods (Costa Vargas, 2006).

Figure 7. Radar chart of occasional posters' involvement, based on whether or not the police was involved in the homicide

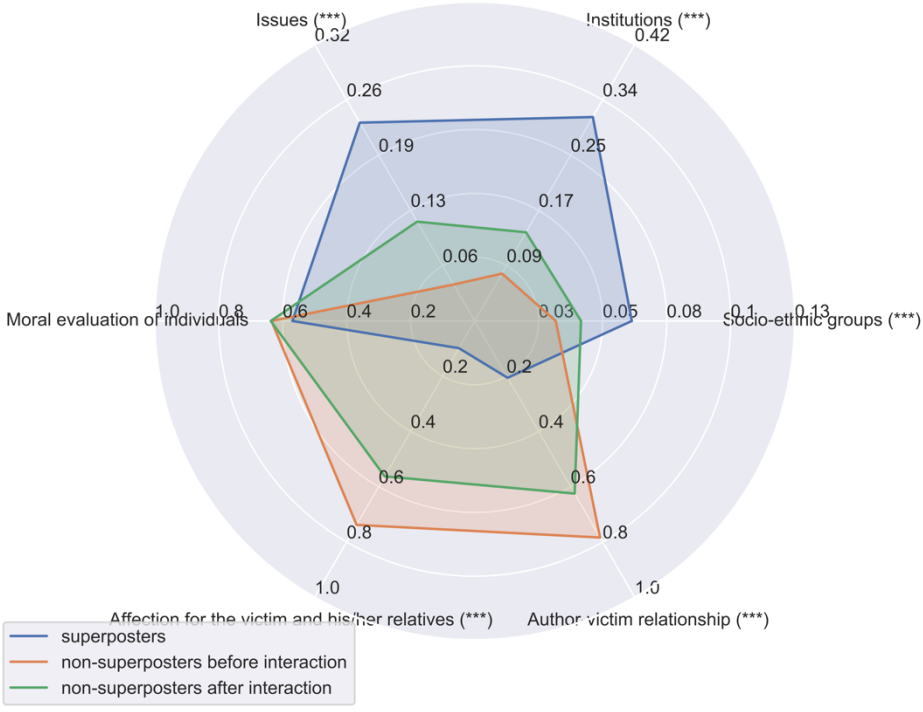


Note: *the dark grey polygon outlines the discursive profile of the comments made by occasional posters, in the case where a police officer was responsible for the death. It connects the dots corresponding to the proportion of these comments that were coded in each variable. For example, 42% of the comments made by* occasional *posters mentioned institutions when a police officer was responsible for the death, compared to 10% when the police was not responsible for the death.*

Fisher's exact test: *p < .05; **p < .01; ***p < .001

A second condition that increased public-oriented discourse in comments by occasional posters was the interactions they had with superposters. As we have seen, the latter overwhelmingly focused on "distant moral evaluation" and the "search for collective responsibility". Their statements were often violent for local contributors, both because they provided general interpretations and because they conveyed harsh judgements of the deceased. Yet quantitative analysis shows that when superposters started commenting on a homicide, this had a significant effect on the participation of local contributors (Figure 9).

Figure 8. Radar chart showing how occasional posters make sense of homicides before and after interaction with superposters



Note: *the grey polygon at the bottom outlines the discursive profile of comments by* occasional *posters where no superposter had yet commented on the homicide. The light grey polygon in the centre outlines the discursive profile of the comments made by* occasional *contributors after at least one superposter had commented.*

Fisher's exact test: *p < .05; **p < .01; ***p < .001

Once at least one superposter had posted a comment on the homicide, occasional posters changed their discourse. They made fewer mentions of their connection to the victim, expressed their love for the victim less, and referred to public problems, institutions and social groups. In other words, the presence of superposters forced them to consider the occurrence in political terms. As Luc Boltanski found when he studied the tensions between the regimes of love and justice (Boltanski, 1990), this shift in framing was experienced as particularly violent by occasional posters who had lost a loved one.

When certain conditions are met, surrounding certain characteristics of the occurrence and interaction with superposters, occasional posters' discourse is more public-oriented. Contrary to what is often hastily posited, the "multitudes of residents" are therefore likely to form a public, that is, to share judgements that are not limited to the private sphere.

CONCLUSION

This study on the emergence of the publics of online platforms that share information in the form of occurrences – in this case crimes, but they could just as well be pollution measurements or school rankings – offers two key insights. First, it shows that when media organizations use digital technology to stop selecting newsworthy stories, the public does not suddenly disappear. To start off, a large number of Internet users select the stories that affect them personally, and consider them only from a proximity perspective. More importantly, however, other ways of forming a public, more closely associated with traditional media, can also be found. More specifically, a significant proportion of Internet users come to share more general interpretations focused around morality or public problems – either because they are activists themselves, or because they interact with activists, or yet because they are affected by problems that spark their indignation. In a way, this finding is consistent with contemporary research, which shows that online publics are undergoing transformation, rather than changing in nature.

Second, this study offers an alternative to research investigating online news publics based solely on audience measurements. While the use of textual traces left by Internet users imposes a limitation – it precludes any insight about the much wider public viewing the information without posting comments –, it does afford access to interpretations which can be processed quantitatively. The originality of the computational method proposed in this

article lies in combining a prior theorization of the textual material (using a pragmatic or actantial framework) with the coding of this material through supervised learning. Coupled with a sociological validation procedure, the aim of this method is to reduce the gap between quantitative audience surveys, which cover a wide range of individuals, but fail to capture interpretations, and more qualitative studies, which capture interpretations but on a very local scale.

As we have seen throughout this article, the sociological exploitation of digital traces forces researchers to find strategies to make up for the lack of information surrounding internet users and the opacity of algorithmic processing. But provided that a research protocol is defined and implemented, which we hope to have contributed to with this article, this type of study allows for investigation of the way in which publics emerge, without falling into the pitfalls of reification that can come with the use of computational research methods.

REFERENCES

BAKSHY E., MESSING S., ADAMIC L. (2015), "Exposure to ideologically diverse news and opinion on Facebook", *Science*, Vol. 348, no. 6239, pp. 1130-1132.

BARNHURST K., MUTZ D. (1997), "American Journalism and the Decline of Event-Centered Reporting", *Journal of Communication*, Vol. 47, no. 4, pp. 27-53.

BENFORD R. D., SNOW D. A. (2000), "Framing processes and social movements: An overview and assessment", *Annual Review of Sociology*, Vol. 26, no. 1, pp. 611-639.

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

BERTHAUT J., DARRAS É., LAURENS S. (2009), "Pourquoi les faits-divers stigmatisent-ils ? L'hypothèse de la discrimination indirecte", *Réseaux*, Vol. 5, no. 157-158, pp. 89-124.

BEUSCART J.-S., BEAUVISAGE T., MAILLARD S. (2012), "La fin de la télévision ? Recomposition et synchronisation des audiences de la télévision de rattrapage", *Réseaux* Vol. 5, no. 175, pp. 43-82.

BEUSCART J.-S., DAGIRAL É., PARASIE S. (2016), *Sociologie d'internet*, Paris, Armand Colin.

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

BOLTANSKI L., GODET M.-N. (1995), "Messages d'amour sur le téléphone du dimanche", *Politix*, Vol. 8, no. 31, pp. 30-76.

BOLTANSKI L., SCHILTZ M.-A., DARRÉ Y. (1984), "La dénonciation", *Actes de la recherche en sciences sociales*, Vol. 51, no. 1, pp. 3-40.

BURSCHER B., VLIEGENTHART R., DE VREESE C. H. (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, no. 1, pp. 122-131.

CÉFAÏ D., PASQUIER D. (eds.) (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, pp. 47-64.

CHATEAURAYNAUD F., TORNY D. (2005), *Les sombres précurseurs. Une socio- logie de l'alerte et du risque*, Paris, Ehess.

CLAVERIE É. (1994), "Procès, Affaire, Cause. Voltaire et l'innovation critique", *Politix*, Vol. 7, no. 26, pp. 76-85.

COINTET J.-P., PARASIE S. (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, no. 3, pp. 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, translated by Joëlle Zask.

DONZELOT J., MÉVEL C., WYVEKENS A. (2003), *Faire société. La politique de la ville aux États-Unis et en France*, Paris, Seuil.

FLAXMAN S., GOEL S., RAO J. M. (2016), "Filter bubbles, echo chambers, and online news consumption", *Public Opinion Quarterly*, Vol. 80, pp. 298-320.

FLETCHER R., KLEIS NIELSEN R. (2017), "Are news audiences increasingly frag- mented? A cross-national comparative analysis of cross-platform news audience frag- mentation and duplication", *Journal of Communication*, Vol. 67, no. 4, pp. 476-498.

GRAHAM T., WRIGHT S. (2014), "Discursive equality and everyday talk online: the impact of 'superparticipants'", *Journal of Computer-Mediated Communication*, Vol. 19, no. 3, pp. 625-642.

HALL S. (1994), "Codage/décodage", *Réseaux*, no. 68, pp. 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, no. 4, pp. 31-46.

KATZ J. (1987), "What makes crime 'news' ?", *Media, Culture and Society*, Vol. 9, pp. 47-75.

LAZER D., PENTLAND A., ADAMIC L. *et al*. (2009), "Computational social science", *Science*, Vol. 323, no. 5915, pp. 721-722.

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)*, pp. 88-93.

LIM M. (2012), "Clicks, cabs, and coffee houses: Social media and oppositional move- ments in Egypt, 2004-2011", *Journal of Communication*, Vol. 62, no. 2, pp. 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., SHAW D. L. (1972), "The agenda-setting function of mass media", *The Public Opinion Quaterly*, Vol. 36, no. 2, pp. 176-187.

MOLOTCH H., LESTER M. (1996), "Informer : une conduite délibérée de l'usage stratégique des événements", *Réseaux*, no. 75, pp. 23-41.

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, no. 1, pp. 45-70.

PARASIE S., DAGIRAL E. (2013a), "Data-driven journalism and the public good: 'Computer-assisted-reporters' and 'programmer-journalists' in Chicago", *New Media & Society*, Vol. 15, no. 6, pp. 853-871.

PARASIE S., DAGIRAL E. (2013b), "Des journalistes enfin libérés de leurs sources ? Promesse et réalité du 'journalisme de données'", *Sur le journalisme*, Vol. 2, no. 1, pp. 52-63.

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

PRIOR M. (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 and D. Pasquier (eds.), *Les sens du public*. *Publics politiques, publics média- tiques*, 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. Available at http://www.journalism.co.uk/news/how-the-homicide-report-tells-the-true-story-of-la-s-violent-crime/s2/a555713/, accessed on 10 April 2019.

RODERICK, K. (2013), "Homicide Report Gets New Life at the LA Times", *LA Observed*. Available at http://www.laobserved.com/archive/2013/03/homicide_report_gets_new.php, accessed on 10 April 2019.

ROSHIER R. (1973), "The selection of crime news by the press", in S. Cohen and J. Young (eds.), *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, no. 1, pp. 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) [1901], *L'opinion et la foule*, Paris, PUF.

VENTURINI T., CARDON D., COINTET J.-P. (2014), "Méthodes digitales. Présentation", *Réseaux*, Vol. 6, no. 188, pp. 9-21.

YOUNG M. L., HERMIDA A. (2015), "From Mr. And Mrs. Outlier to central tendencies. Computational journalism and crime reporting at the *Los Angeles Times"*, *Digital Journalism*, Vol. 3, no. 3, pp. 381-397.

ZASK J. (2008), "Le public chez Dewey : une union sociale plurielle", *Tracés*, Vol. 2, no. 15, pp. 169-189.

Appendix 1. Ex-ante evaluation of the classifiers

For each variable, the software uses a sample of the comments coded by us, called a test sample, to evaluate the quality of the inferred statistical model that will subsequently be used to code the entire corpus. These comments are deliberately excluded from the learning sample that is actually used to train the classifier. The labels of the test sample are thus compared with those predicted by the model generated at the end of the training. The "confusion tables" below summarize the results of this double coding (machine and human) for each variable. A set of usual metrics are then produced, which assess the quality of the ex-ante coding performed by the machine.

1. Contributor-victim relationship		machine		Total
		no	yes	
human	no	34	11	45
	yes	6	74	80
Total		40	85	125

Precision: 0.87
Recall: 0.92
F score: 0.90
Baseline: 0.64
Accuracy: 0.86

2. Affection towards the victim or their relatives		machine		Total
		no	yes	
human	no	74	8	82
	yes	9	130	139
Total		83	138	221

Precision: 0.94
Recall: 0.94
F score: 0.94
Baseline: 0.63
Accuracy: 0.92

3. Moral evaluation of individuals		machine		Total
		no	yes	
human	no	46	19	65
	yes	6	42	48
Total		52	61	113

Precision: 0.69
Recall: 0.87
F score: 0.77
Baseline: 0.58
Accuracy:0.78

4a. Public problems		machine		Total
		no	yes	
human	no	150	3	153
	yes	15	28	43
Total		165	31	196

Precision: 0.90
Recall: 0.65
F score: 0.76
Baseline: 0.78
Accuracy: 0.91

4b. Institutions		machine		Total
		no	yes	
human	no	147	6	153
	yes	20	47	67
Total		167	53	220

Precision: 0.89
Recall: 0.70
F score: 0.78
Baseline: 0.70
Accuracy: 0.88

4c. Social/ethnic groups		machine		Total
		no	yes	
human	no	138	1	139
	yes	10	17	27
Total		148	18	166

Precision: 0.94
Recall: 0.63
F score: 0.76
Baseline: 0.84
Accuracy: 0.93

Appendix 2. Ex-post evaluation of the classifiers

For each variable, the model generated at the end of learning was tested on a random sample of 300 comments from which we exclude the training set. A new human coding was performed, which was then compared to the coding performed by the machine. The "confusion tables" below summarize the results of this double coding for each variable. A set of standard metrics are then produced, which assess the quality of the machine coding. This ex-post control seemed necessary to us due to the voluntarily biased nature of the annotation corpus, which is a consequence of the active nature of our annotation procedure. In fact, our ex-post evaluation on a completely random corpus of comments (after exclusion of the learning corpus used) shows that the performance of the classifiers is slightly lower than the ex-ante evaluation while still satisfactory.

1. Contributor-victim relationship		machine		Total
		no	yes	
human	no	98	42	140
	yes	13	154	167
Total		111	196	307

Precision: 0.78
Recall: 0.92
F score: 0.85
Baseline: 0.54
Accuracy: 0.82

2. Affection for the victim or their relatives		machine		Total
		no	yes	
human	no	101	18	119
	yes	28	160	188
Total		129	178	307

Precision: 0.9
Recall: 0.85
F score: 0.87
Baseline: 0.62
Accuracy:0.85

3. Moral evaluation of individuals		machine		Total
		no	yes	
human	no	88	36	124
	yes	59	124	183
Total		147	160	307

Precision: 0.77
Recall: 0.68
F score: 0.72
Baseline: 0.6
Accuracy: 0.69

4a. Public problems		machine		Total
		no	yes	
human	no	270	6	276
	yes	13	17	30
Total		283	23	306

Precision: 0.74
Recall: 0.57
F score: 0.64
Baseline: 0.9
Accuracy: 0.94

4b. Institutions		machine		Total
		no	yes	
human	no	245	7	252
	yes	18	36	54
Total		263	43	306

Precision: 0.84
Recall: 0.67
F score: 0.74
Baseline: 0.82
Accuracy: 0.92

4c. Social/ethnic groups		machine		Total
		no	yes	
human	no	292	3	295
	yes	4	7	11
Total		296	10	306

Precision: 0.7
Recall: 0.64
F score: 0.66
Baseline: 0.96
Accuracy:0.98

Appendix 3. Linear regression of the number of comments and of coverage in the print edition of the *Los Angeles Times* based on the characteristics of the victim, the homicide and the neighbourhood (Ordinary least squares method)

	A	B	C	D
	Coverage in the *Los Angeles Times*' printed edition	Number of comments per month (all users)	Number of comments per month (superposters only)	Number of comments per month (occasional posters only)
Age of the victim				
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)
Gender of the victim				
Female	0.1129	-0.1109	0.0263	-1.4866
	(0.1056)	(0.5557)	(0.1571)	(1.1657)
Male	0.0176	-0.1789	0.0053	-0.9258
	(0.1063)	(0.5592)	(0.1581)	(1.1732)
Ethnicity of the victim				
Black	0.0417	0.3713**	0.0609	-0.2240
	(0.0306)	(0.1613)	(0.0456)	(0.3384)
Hispanic	0.0259	0.2220	0.0796*	-0.0864
	(0.0295)	(0.1550)	(0.0438)	(0.3251)
Asian	0.0582	-0.2214	-0.0362	0.2762
	(0.0456)	(0.2399)	(0.0678)	(0.5034)
White	0.0528	0.1262	0.0335	0.1791
	(0.0326)	(0.1719)	(0.0486)	(0.3606)
Other	-0.0650	-0.3539	-0.0841	0.2216
	(0.1242)	(0.6534)	(0.1847)	(1.3708)
Circumstances of the homicide				
Domestic violence	-0.0781***	-0.3502**	-0.0794*	0.1260
	(0.0302)	(0.1589)	(0.0449)	(0.3334)
Drive-by	-0.0734***	0.1245	-0.0065	0.3493
	(0.0259)	(0.1368)	(0.0387)	(0.2869)
Fight	-0.0624***	-0.1293	-0.0292	-0.3473
	(0.0224)	(0.1181)	(0.0334)	(0.2477)
Police officer involved	0.0918***	0.7587***	0.3902***	-0.3681
	(0.0274)	(0.1446)	(0.0409)	(0.3033)
Party	0.0702	-0.0640	-0.1376	-0.5746
	(0.0619)	(0.3255)	(0.0920)	(0.6828)
Robbery	0.1576***	-0.0687	0.0455	0.3996
	(0.0394)	(0.2082)	(0.0589)	(0.4368)
Walk-up	-0.0535**	0.0078	-0.0814***	0.1509
	(0.0209)	(0.1102)	(0.0312)	(0.2313)

shooting				
Cause of homicide				
Blunt force	-0.0010	-0.0053	0.0253	-0.6884**
	(0.0316)	(0.1660)	(0.0469)	(0.3482)
Gunshot	0.0157	0.0020	0.0135	-0.3001
	(0.0191)	(0.1004)	(0.0284)	(0.2106)
Other	0.0949**	-0.0563	0.0025	-0.2447
	(0.0380)	(0.2001)	(0.0566)	(0.4197)
Stabbing	-0.0098	-0.0571	-0.0419	0.2191
	(0.0245)	(0.1290)	(0.0365)	(0.2706)
Strangling	-0.0452	0.2328	-0.0220	-0.3680
	(0.0445)	(0.2341)	(0.0662)	(0.4911)
Mention of the word "gang" in the article	0.0170**	0.1265***	0.0370***	-0.2104**
	(0.0086)	(0.0454)	(0.0128)	(0.0953)
No coverage in the *Los Angeles Times*' printed edition	–	-0.2152***	-0.0753***	0.2471*
		(0.0644)	(0.0182)	(0.1351)
Main ethnicity of the neighborhood's residents				
Asian	-0.0611	0.1699	0.0788	-0.9299**
	(0.0422)	(0.2223)	(0.0629)	(0.4664)
Black	0.0032	0.1030	-0.0567	0.3914
	(0.0241)	(0.1270)	(0.0359)	(0.2664)
Hispanic	-0.0230	0.0058	-0.0115	0.1751
	(0.0177)	(0.0933)	(0.0264)	(0.1956)
Average income of the neighborhood (log)	-0.0160	0.0488	-0.0142	0.2563
	(0.0261)	(0.1372)	(0.0388)	(0.2877)
Neighborhood's homicide rate (log)	-0.0778***	0.0521	0.0268	-0.2769
	(0.0242)	(0.1278)	(0.0361)	(0.2680)
Intercept	0.6506*	0.4707	0.2837	2.2773
	(0.3368)	(1.7716)	(0.5009)	(3.7166)
R-squared	0.07	0.09	0.12	0.04
No. Observations: 1,699				

Note: *p < .05; **p < .01; ***p < .001
Standard errors in parentheses