# Public Perceptions on Organised Crime, Mafia, and Terrorism: A Big Data Analysis based on Twitter and Google Trends

**Panos Kostakos**[1]
University of Oulu, Finland

## Abstract

*Public perceptions enable crime and motivate government policy on law and order; however, there has been limited empirical research on serious crime perceptions in social media. Recently, open source data—and 'big data'—have enabled researchers from different fields to develop cost-effective methods for opinion mining and sentiment analysis. Against this backdrop, the aim of this paper is to apply state-of-the-art tools and techniques for assembly and analysis of open source data. We set out to explore how non-discursive behavioural data can be used as a proxy for studying public perceptions of serious crime. The data collection focused on the following three conversational topics: organised crime, the mafia, and terrorism. Specifically, time series data of users' online search habits (over a ten-year period) were gathered from Google Trends, and cross-sectional network data (N=178,513) were collected from Twitter. The collected data contained a significant amount of structure. Marked similarities and differences in people's habits and perceptions were observable, and these were recorded. The results indicated that 'big data' is a cost-effective method for exploring theoretical and empirical issues vis-à-vis public perceptions of serious crime.*

Keywords: Organised Crime, Terrorism, Mafia, Big Data, Twitter, Google Trends, Social Media, Perceptions.

## Introduction

Culture and public perceptions enable serious crime and motivate policy action; however, there has been limited empirical research on serious crime perceptions in the social media. This study uses data from web-search queries and assorted social media postings to establish perceptions of serious crime. The aim is to explore the feasibility of using disaggregated data from Google Trends and Twitter to gain theoretical insights and to encourage readers to undertake their own analysis. With around forty percent of the world population having an Internet connection (United Nations, 2015), the data being

[1] Postdoctoral Researcher, Center for Ubiquitous Computing, Faculty of Information Technology and Electrical Engineering, University of Oulu, P.O. Box 4500, FIN-90014, Finland. Email: panos.kostakos@oulu.fi

produced by users and machines communicating over data networks is growing at an exponential rate. Also growing rapidly is the digital footprint people leave online. These trends have motivated research in criminology (Wang et al., 2012; Gerber, 2014), economics (Preis et al., 2013; Choi & Varian, 2012; Vosen & Schmidt, 2011), sociology (Kostakos et al., 2013; Anderson et al., 2014; Eagle et al., 2009), linguistics (Michel et al., 2011; Hand 2011; Lieberman et al. 2007), politics (Kallus 2014; Louis & Zorlu, 2012), and healthcare (Seifter et al., 2010; et al., 2015). Research that studies human behaviour through digital footprints is addressing focal social problems and identifying new problems (Kostakos & Ferreira, 2015). We first provide a review of recent work regarding perceptions of serious crime and then proceed to present the data sources. Subsequently, we expand on the research methods and obtained results. The final section draws out preliminary conclusions and highlights the main bottlenecks affecting our approach.

## Perceptions of Organised Crime: From Discourse Data to 'Big Data'

Individual and group perceptions is a conspicuous subject in the literature of organised crime, not only because of the critical works of American and European historians, criminologists, and sociologists who have been trying to debunk the stereotypical image of "organised crime" created by media sensationalism and bureaucratic obsession (Paoli 2003: 3), but also because of the causal impact perceptions have in enabling individual and collective behaviourism that arena (Smith, 1975; Allum et al., 2010). Several well-established empirical studies have demonstrated how criminal groups engage in criminal recruitment and extortion by manipulating ethnic identities and public perceptions (Ianni, 1974; Bovenkerk et al., 2003; Chin, 2000; Bovenkerk, 1998; Paoli, 2003), while others have examined the role of culture in enabling criminal infiltration into the economy, politics, civil society and bureaucracy (Kleemans & de Bunt, 1999, 2008; Chambliss, 1971; Ianni, 1974; Allum, 2006; Schneider and Schneider, 2003).

More recently, there has been renewed interest in measuring the perceptions of a range of actors involved in organised and serious crime. Sarno (2014) reviewed Spanish, German and Dutch newspapers articles between 2000 and 2013 and studied the representations of the Italian Mafia. Similarly, content analyses of newspapers have also been published by Pruss (2014) and Young and Allum (2012). Shen et al. (2013) in their recent study of child trafficking in China adopted a broader approach by focusing on Chinese open media sources. Systematic content analysis of official documents and media sources has been used sparingly in the past. For example, Mcillwain (1997) studied public perceptions of Chinese organised crime in the US by analysing New York City press documents from 1894 to 1908. Likewise, Arsovska and Kostakos (2010) developed a corpus of about 3 million words from EU press releases from 1994 to 2008 and examined EU norm diffusion in the Balkan region. Lastly, Décary-Hétu and Aldridge (2015) have discussed the application of automatic data extraction.

A large body of empirical literature has focused on how the general public perceives serious crime. Mendoza (2015) has reported findings on public perceptions using focus groups in Latin America. Arsovska and Michilli (2015) conducted open-air interviews in New York in their study of the public's perceptions of ethnic Albanian. Correspondingly, Travanglino et al. (2015) surveyed 179 high school students from the Southern Italian Region of Campania and recorded local perceptions of organised crime. Finally, a fair number of articles have looked into the perceptions of experts, victims (Tilley & Hopkins, 2008; Hill 2010; Ouimet & Montmagny-Grenier, 2014; Williams and Levi, 2012; Jerry et

al., 2014), entrepreneurs (Van Dijk, 2007; Daniele & Marani, 2011; Gottschalk, 2013; Sutter et al., 2013), and criminals (Covington & Bloom, 2003; Reynolds, 2011).

Researchers outside the narrow field of organised crime have been increasingly drawn to mining online data sources; Twitter and Google Trends stand out as two important sources of social data. Online data sources provide fresh insights into online sociological research. Liang et al have studied online networks to determine how tobacco-related content can influence users/consumers (Liang et al., 2015). Makin and Morczek (2015) have analysed Google search queries to study perceptions of rape. Google Trends has enabled a stream of research on detecting suicide epidemics, and a recent review article has noted that there is a fast-growing body of published research looking at health related phenomena using data from the same source (Fond et al., 2015). There has also being some research on security-related and law and order phenomena. Gerber (2014) incorporated textual data from Twitter into a crime prediction model, concluding that this model outperforms other standard crime prediction approaches. Wang et al. (2012) analysed twitter feeds from local news agencies using automatic sentiment analysis and predicted hit-and-run incidences and breaking-and-entering crimes. Nation-wide events like mass riots and social unrest have also been observed through the social media lens; proposed models have achieved effective prediction results. For example, Howard et al. (2011) discovered retrospective evidence that mass protests during the Arab Spring of 2010 were correlated with social media activities preceding these events. More recently, Kallus (2014) analysed data from 300,000 open content web sources, data which referred to the 2013 Egyptian coup d'état and concluded that open-source and social media data can predict future events.

Although organised crime is enabled by culture and perceptions, research on people's perception of serious crime remains sporadic and fractured. Traditional tools that measure public attitudes make extensive use of questionnaires and self-reported data, and although crucial sources of information, they are expensive and also time-consuming to generate and reproduce (Sung, 2004; Bryman, 2012). Moreover, field research on crime in general and serious crime in particular involves additional limitations (Van Dijk et al., 1990; Sung, 2004; Chin, 2000). However, there is a growing body of literature outside of the narrow boundaries of our research field, research which is facilitated by rapid advancement in information and communication technologies.

With the increase in Internet use, there has been a substantial surge in user-generated content including keyword search volume, social media networks, geo-tagging, online messaging, and images, to name a few (Krumm et al., 2008; Girardin et al., 2008; Cha et al., 2007). While some studies in the area of organised crime have sought to examine open-source data in a systematic way (Décary-Hétu and Aldridge, 2015; Pruss, 2014; Young and Allum, 2012; Shen et al., 2013; Arsovska and Kostakos, 2010), the vast majority of social media data remains untapped. Thus, it remains an open question as to whether disaggregated social data can help us draw meaningful conclusions which will supplement current knowledge on serious crime perceptions.

## Data Sources: Twitter and Google Trends

The empirical data presented here is drawn from user-generated content. User-generated content is defined as media information that is 'created or produced by the general public, rather than by paid professionals and is primarily distributed via Web 2.0

technologies online' (Daugherty et al., 2011:147). Some prominent websites and services hosting content created and shared by users include YouTube, Twitter, Facebook, and Google. The digital footprint of user-generated content is distinguished between *active* and *passive* content (Girardin, 2008). Active content is composed of self-reported data, created and shared via various popular broadcasting sites, and includes media artifacts like text and images or a combination of these two. Passive content emerges from the interaction between users and machines. This interaction generates "behavioural data" that captures the habits of users in a disaggregated manner (Girardin, 2008; Kostakos et al., 2009). The remainder of this section will review the two main sources of data used in this study: Twitter and Google Trends.

Twitter is a free social networking site that was launched in March 2006, commands more that 200 million users, and handles on average 500 million messages a day. Twitter users have personalized accounts and can post messages (tweets) about any topic within the 280-character limit. Tweets are micro-narratives that encapsulate, in a brief sentence, users' perceptions on a given topic, making it an ideal tool for studying public perceptions. Tweets can contain text, images, videos, and hyperlinks. The novelty of Twitter stems from the fact that users have innovative conversational technology to contribute to public debates by using *hashtags*, *re-tweets*, *mentions* and *replies* (Zubiaga et al., 2014). While social media are challenged by the spread of 'bots' and trolls, hashtags have been critical in enabling norm and information diffusion from Twitter into the public domain, and vice versa. In simple terms, hashtags are being used for tagging and classifying messages and ideas that, in turn, promote specific topics and people. This makes the task of searching for clusters of tweets and information with common topical keywords much easier and reliable. It is also an effective way of identifying debates and trending conversations online. Another innovation of Twitter is that anyone can join a public debate by replying to others and/or by mentioning users and topics in their own public messages. Thus, the combination of tools like *hashtags* and *mentions* fosters a robust and open community for public deliberation. From a researcher's perspective, this technology enables the extraction and preservation of both textual information and relational data. Messages and relational data are extracted from Twitter manually or by using Twitter's Application Programming Interface (API).

**Table 1. Organised Crime Twitter Bank (OCTB): Descriptive indicators**

| Keywords | Sample* | Final Sample | % | Re-tweets | Word count | Users | Replies | Mentions | Hashtags |
|---|---|---|---|---|---|---|---|---|---|
| Terrorism | 392,974 | 317,370 | 87% | 140,251 | 6,041,345 | 147,187 | 17,875 | 37,257 | 13,436 |
| Mafia | 128,248 | 31,795 | 9% | 11,005 | 555,100 | 22,086 | 4,044 | 9,740 | 3,631 |
| OC | 18,496 | 13,320 | 4% | 4,191 | 242,277 | 9,240 | 1,351 | 3,095 | 1,505 |
| Total | 539,718 | 362,485 | 100% | 155,447 | 6,838,722 | 178,513 | 23,270 | 50,092 | 18,572 |

*This value indicates raw number of tweets streamed from the API. High level of noise in the data indicates the generic use of the keyword (i.e. the keyword was found in URL or in username fields but not in the actual text of the tweet).

Google is the largest search engine online with which users can passively interact to search for information. A free service with over four billion users, Google is leading the

search engine infrastructure with over 40,000 search queries on average per second. As of 2012, Google captures 65% of all web search volume worldwide with over 100 billion searches monthly (Sullivan, 2013). Interestingly, 20% of daily-submitted search queries in Google have never been asked before (Farmer, 2013). Thus, Google is a suitable data source for monitoring past, current, and future trends. Google has launched an in-house service called Google Trends[2] that allows users to access search volume trends; it is a real-time weekly and monthly index of the volume of queries that users enter into Google.

## Methods and Results

Twitter's API was used to collect tweets on three conversational topics between June 10 and July 10, 2014. A total of 540,000 tweets containing three alterative spellings of keywords (mafia, organised crime, and terrorism) were initially collected from Twitter's API and stored into a database—the Organised Crime Twitter Bank (OCTB). While philological debate about the definitions of these phenomena is ongoing, the three topics mentioned above have been selected because each one, overall, appeared to involve significant levels of violence, strong cultural residues, and powerful threat narratives that capture the public's imagination (Smith, 1975). The sample was subsequently filtered based on inclusion and exclusion criteria. Tweets that did not contain one of the three keywords in the message-field or whose language was other than English were excluded from the final analysis. The data was reduced to a total of 362,485 entries with the majority of excluded tweets (noise) coming from the keyword mafia (see Table 1). Surprisingly, the number of tweets that contained more than one topic (i.e., crime-terror nexus) was insignificant and therefore excluded from the final analysis.

Significant differences between the three conversational topics were observed. Of the 362,485 instances analysed, over 13,000 tweets mentioned the keyword "organised crime" whereas about 30,000 records mentioned the "mafia" keyword. The sample of tweets related to the keyword "terrorism" was by far the largest sample group, totalling about 317,000 cases. In terms of percentages, tweets that mentioned "terrorism" comprised over 87% of the total data; about 9% of the data were related to "mafia" tweets; and 4% of the messages mentioned "organised crime". The number of Twitter users varied considerably between the three topics. The results, as seen in Table 1, indicate that more than 147,000 users tweeted about "terrorism;" about 22,000 users shared messages about the "mafia;" and over 9,000 users mentioned "organised crime" in their conversations. Similarly, the proportion of replies, mentions, and hashtags per topic are in line with the overall size of the network. Further measures of the validity and reliability of the data were considered and analyzed.

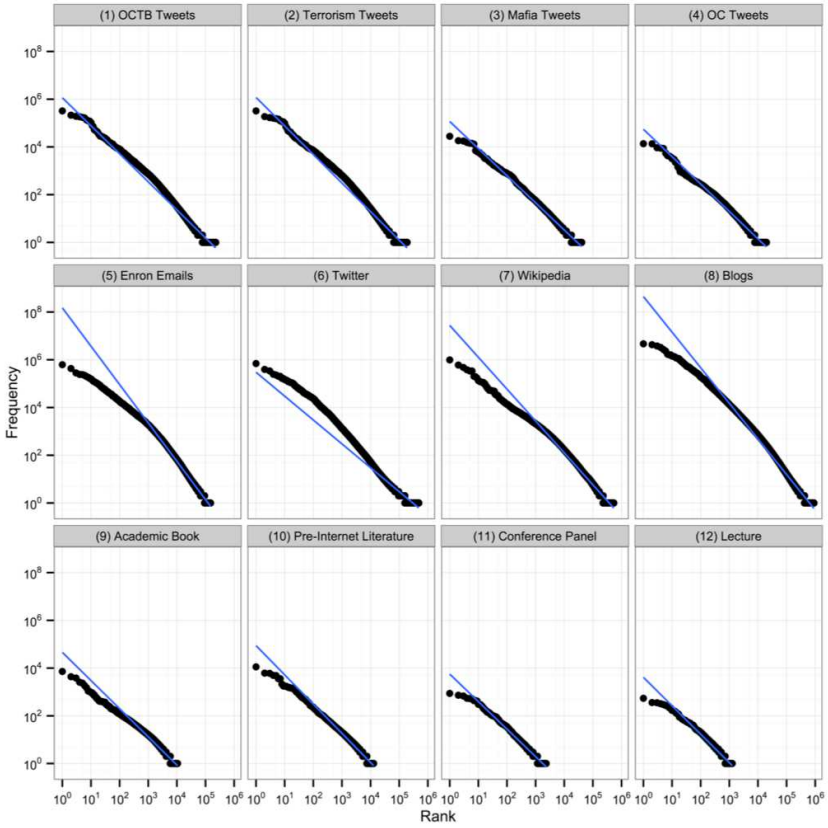
### *1. Internal Consistency and Contextomy*

Given that the social media data presented here are machine-extracted and time-sensitive, two control measures were employed to account for internal consistency and contextomy before proceeding to the final analysis. First, we sought to establish the fact that the crawled data included non-random language inputs. As shown in Table 1, the OCTB is rich in textual data, with the keyword "terrorism" returning over 6 million words, following by well over half a million words for "mafia," and nearly a quarter of a

[2] https://www.google.com/trends

million words for the keyword "organised crime". We used Zipf's Law to determine the internal consistency of the vocabulary distribution in our sample data (Ferrer-i-Cancho & Elvevåg, 2010; Zipf, 1949). In Figure 1 we compare the OCTB with eight different text corpora that include both electronic and non-electron communications. When plotted in log-log scale, the resulting graphs return an upward angle line—a signature of a natural language corpus. Figure 1 (graphs 1-4) shows the plots of word-rank against frequency in log-log coordinates, with the results indicating the normality of our sample. The distribution that emerges seems to follow Zipf's law (Zipf, 1949; Newman, 2005), suggesting that the textual data in the OCTB convey information and not just random noise.

**Figure 1. Distribution of word frequencies in different corpora**[3]

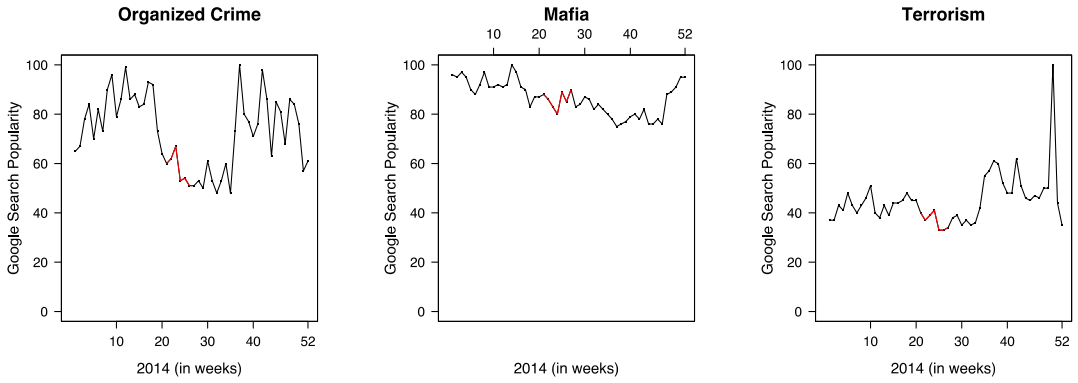



A second benchmark was used to determine the reliability of our data and control for contextomy (Mcglone, 2005). Spikes in social media activity are being generated too often in response to real life events and developments. A news event or a major police operation

[3] (1) Double logarithmic scale of 6,838,722 words recorded in the OCTB. The slope of the curve indicates that the two quantities (word frequency and word rank) are related with a power-law. When plotted in log-log scales, the resulting graph assumes an upward angle line, a signature of a natural language corpus. (2-4) Plots for each individual keyword as found in the OCTB database; (5-8) Various online databases in rich textual information. (9-12) Examples of textual data–oral and written.

might produce a significant stream of social media data. Research has shown that major world events like terrorist attacks and earthquakes very often cause a spike in the social media data-stream, so this should be factored into the analysis (Sakaki et al., 2010). We are interested in identifying whether the social media data we had collected correlated to any major world event. Unfortunately, Twitter does not provide free access to every user's broadcast data. However, using Google Trends we were able to identify the popularity of each keyword searched within a 52-week period. Figure 2 shows the volume of searches conducted in Google for each keyword throughout 2014. The period that the data were being generated by Twitter users does not align to a spike in search volume that might had been triggered by a major real-life event.

**Figure 2. Worldwide Google search popularity for the three topics over the course of one year**[4]




### *2. Discourse production: Broadcasting*

We first examined the broadcasting patterns in each conversational topic in Twitter. Using "discourse production" as an analytical metaphor, we looked at the time series of micro-narratives published in Tweeter. The underlying assumption is that the way textual information is being produced denotes how people perceive the subject in question. This reductionist logic is a rudimentary and easily reproduced metric that measures the frequency of individual narratives (tweets) and allows us to infer results by comparing how different conversational topics perform at the macro level.

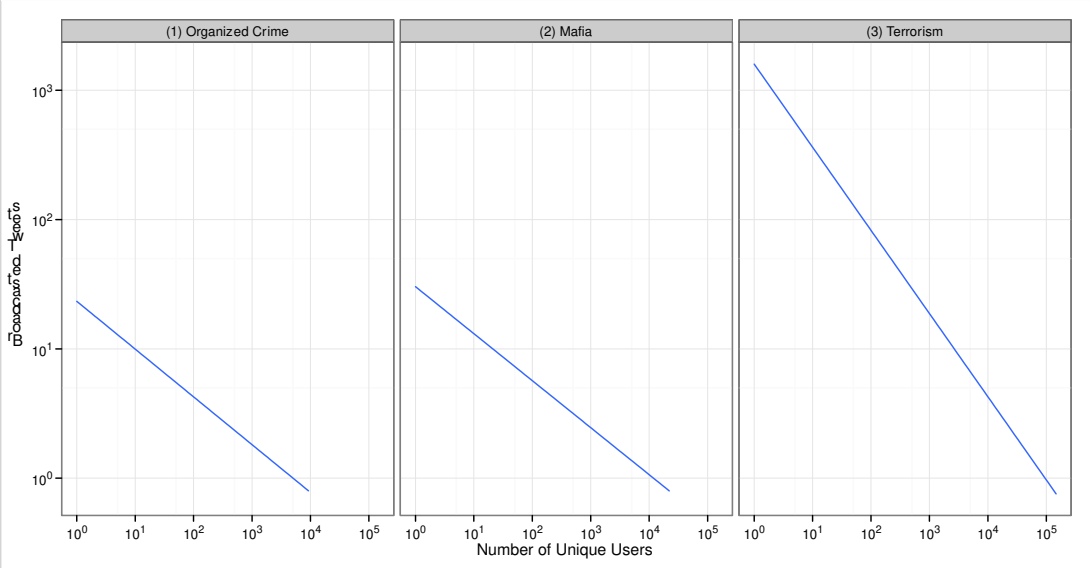
The frequency distribution of the broadcast messages for each keyword is given in Figure 3. A scatter plot of frequencies versus unique users (blue line denotes power law proportionality) shows the cumulative distribution. Our results demonstrate that the broadcasting patterns are heavy-tailed, indicating that the users' behaviours across the three groups are non-normally distributed. Broadcasting tends to exhibit extreme polarity, with very large or very low values likely to arise. For example, when we zoom in the first

[4] Each graph shows the relative popularity of the keyword as expressed through searches conducted by users via the Google search engine. The data are for the year 2014. The x-axis on the graphs measure weeks, and the y-axis measures the value relative to the total searches conducted for each term. Areas marked in red highlight the periods covered by the Twitter data.

graph of Figure 3, we see that 7,968 users published one tweet, comprising a significant part of the published tweets. Next, 976 users published between two and three tweets, and 233 users published between four and nine messages. As the number of messages per user increases, the number of users drops sharply. In particular, 57 users published between 10 and 49 tweets and five users posted in the range of 50 to 149 messages. A single user published over 150 tweets, with the maximum number of tweets per user being 175. Thus, at the low end of the distribution range, we have found that 90% of the sampled population had published 60% of the messages, whereas at the high end of the range, a staggering 4% had published about 25% of the tweets.

This is not a counterintuitive result, for heavy-tailed distributions have been found to be widespread in Internet traffic, income, wealth, books sold, telephone calls received, size of cities, booms and busts in economic cycles, citations of scientific papers, co-authorship and dynamics of earthquakes, to name just a few (Newman, 2005; Andriani & McKelvey, 2009). The human-decision making dynamics behind heavy-tailed distributions are complex and require further analysis. In regard to our problem, the results obtained might be applicable in explaining norm diffusion and moral panics. While heavy-tailed distributions are ubiquitous on the Internet, there still remains the question of why this behaviour emerges in our data and what there is to be learned.

**Figure 3. Heavy-tailed distributions of tweets vs unique number of users**




An interesting finding is the marked differences observed in the power-law fit of the three distribution lines. Although the distribution curves in Figure 3 are hyperbolic, having the characteristic long tail, the third case stands out, as the power-law curve fits well in the data over numerous orders of magnitude. This implies that the broadcasting behaviour of people posting messages about "terrorism" is more predictable. Thus, the relationship between broadcast messages per user as seen in Figure 3, becomes more consistent and self-organised. Unfortunately, due to a lack of data, controlling for size sample was not possible; however, we analysed the profiles of the 100 most active twitter users within each conversational topic and identified possible difference that might explain this divergence.

The subjects discussed in Twitter are vast. User profiles range from personal to institutional to political to corporate to journalistic (Yamaguchi et al., 2010; Zubiaga et al., 2014). Table 2 summarizes the demographic data taken from users who were active in the three conversational topics. The left column is a list of 20 general themes based on a manual review of the 100 most active users from our sample. The three remaining columns report how many times each predefined theme was encountered. Whilst additional measures and benchmarks should be considered when selecting and reviewing the profiles, the results reveal some interesting patterns. First, we note a significant number of personal and art-related profiles that have been very actively involved in the conversation about mafia. Further analysis showed that the keyword "mafia" is being used mainly in the context of the hip-hop community where "Mafia style" branding is thriving (Ogbar, 1999, p. 167). Second, "terrorism" is by far the most institutionalised topic discussed with a significant number of official profiles (including terrorists and military organisation) participating in the online conversation. While further statistical tests are in order, it might be fruitful to explore the observed relationship between broadcasting patterns and the type of user profile.

**Table 2. Major themes found in the top 100 active profiles in the OCTB**

| Theme | Organised Crime | Mafia | Terrorism |
|---|---|---|---|
| Citizens/activists/bloggers | 19 | 10 | 14 |
| Politicians | 14 | 4 | 22 |
| News | 13 | 6 | 24 |
| Police | 13 | 0 | 0 |
| Campaign | 7 | 3 | 9 |
| Journalists | 6 | 2 | 10 |
| Global Governance | 5 | 0 | 1 |
| Academic | 3 | 0 | 0 |
| Government | 3 | 0 | 2 |
| Specialist News | 3 | 0 | 0 |
| Book club | 2 | 5 | 2 |
| Art | 1 | 41 | 1 |
| Prosecutors | 1 | 0 | 0 |
| Religion | 1 | 1 | 1 |
| Corporations | 1 | 2 | 2 |
| NGOs | 4 | 0 | 2 |
| Personal blog | 3 | 26 | 4 |
| Terrorists | 0 | 0 | 2 |
| Army | 0 | 0 | 4 |
| Criminals | 1 | 0 | 0 |

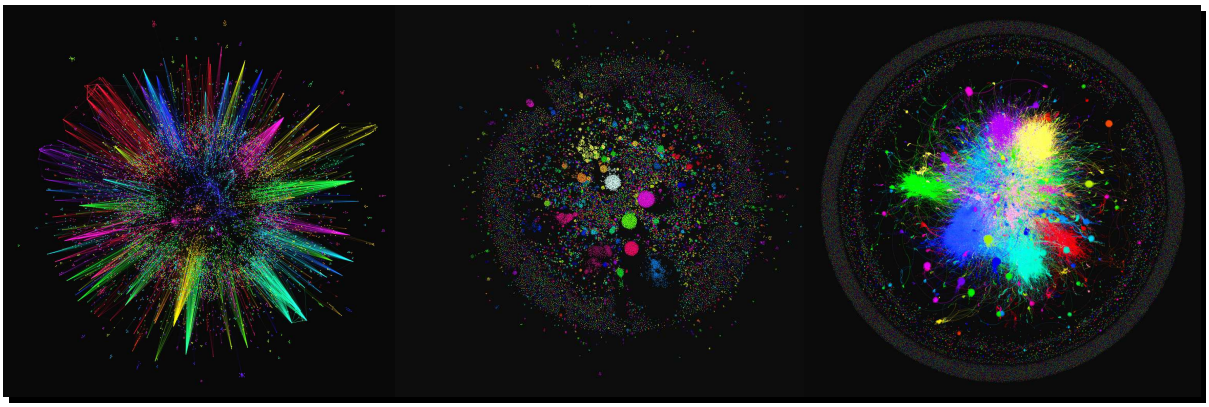
### *3. Discourse Diffusion*

In the previous section, we looked at how disaggregated data on discourse production can be examined. In this section we measure the diffusion of discourse and examine how information within the three topics was shared. The logic behind this is that the way information is being shared between users can tell us a lot about how people perceive the topics in question. Recent research has shown that the network topology in online conversations varies according to the subject and topics discussed. Smith et al. (2014) proposed a typology of conversational archetypes on Twitter and have put forward six archetypes of network clouds that one finds in large scale online conversations:

- *Polarised Crowds:* Polarised discussions are defined by two large, closely connected groups of users discussing contested topics and characterised by the lack of communication between the two densely connected groups.
- *Tight Crowds:* Unlike polarised crowds, tight crowds are highly interconnected users with few isolates.
- *Brand Clusters:* Popular products and services very often become the subject of social media discussion. These topics attract large crowds of disconnected and fragmented users that tend to mention a topic, but for the most part, do not mention each other.
- *Community Clusters:* There are cases in which online discussions produce hubs of small groups around a debate.
- *Broadcast Network:* This type of network is very often encountered around commentary concerning news and events broadcast by major news media and is composed of users who repeat the messages generated by media outlets or individuals.
- *Support Network:* In contrast to the broadcast network, support networks are characterised by hubs that reply to many disconnected users. This network typology is very often encountered in situations where official Twitter accounts handle complaints and customer issues.

We sought to discover how well the data from the OCTB fit into this typology. The obtained results exhibited in Graph 1 depict three community networks based on users' exchange of messages within each one of the three on-going conversations. Each node in the network graph corresponds to a unique user, and a link between two users is drawn if a user's username is mentioned in a tweet. The further from the centre of the network a node appears, the less connected with the rest of the nodes it is. Thus, nodes that appear as white noise in the black background are isolated users who mention the topic, but not each other. Coloured nodes are clusters of users or hubs that are being mentioned by others. The first graph on the left (organised crime) shows what appears to be a "tight crowd" network structure. This network is composed of densely connected hubs in which many people mention one another and a fairly low number of isolated nodes. Next, the second graph (mafia) shows a "brand cluster" with a large number of people mentioning the topic and not each other and a small number of hubs with limited interconnection. Finally, the third network (terrorism) resembles a 'community cluster' with many highly connected medium and large groups and a moderate number of isolated nodes. Although additional tests are in order, the network logic seems to be a promising tool for inferring perceptions from large crowds.

**Graph 1. Visual representation of the three topics (users' mentions)**
*From left to right: social network of interaction between users who mention organised crime, mafia, and terrorism.*



### *4. Discourse consumption*

In this section we look at how information and discourse around specific keywords are being consumed by users. The underpinning hypothesis is that the keyword search queries and habits associated with the way people search for information in Google allows for meaningful conclusions to be inferred with regard to users' perceptions. While discrete observation of the queries users enter into online search engines can appear random, there are repeating and easily identifiable patterns when location and time are taken into account. Using Google Trends, an online tool that provides a time series index of the volume of queries people enter into Google in a given time and location, we analysed the performance of three keywords: organised crime, terrorism, and mafia. Comparisons between the three cases allowed us to observe how people's queries differed over time. The data obtained were subsequently shorted into daily and monthly time series.

In Figure 4 we depict the popularity of each of the three keywords for every day of the week over the course of 90 days. This kind of user data can help us determine how many searches have been carried out for a specific keyword on a given day of the week. The periodic variation observed in daily breakdown shows that high and low popularity values tent to always occur in some days of the week. It can be inferred from Figure 4 that Google searches seem to follow user's habits, workflows, and routines. In our 90-days sample period, search queries for "organised crime" and "terrorism" are decreasing during weekends, whereas searches for "mafia" are decreasing on weekdays and increase during weekends. This observation indicates that "mafia" search queries are correlated with leisure time. This conclusion is in line with our previous observations that "mafia" tweets tent to be related with leisure and entertainment activities, as well as with the overall network topology of the community. Searches for "organised crime" and "terrorism" are more popular on weekdays than during weekends. In particular, we notice that during the first days of the week "organised crime" queries have a high popularity value whereas queries for "terrorism" tend to occur towards the end of the working week. A plausible explanation is that the two keywords are perceived as productivity-related subjects that remain embedded in the workflow of the working week. Results from our twitter analysis

also support this conclusion as the institutionalisation theme remerged also here in the web surfing habits. We now turn to an examination of whether there is month-to-month seasonal variations in the data.

**Figure 4. Day-to-day variations of web query popularity**

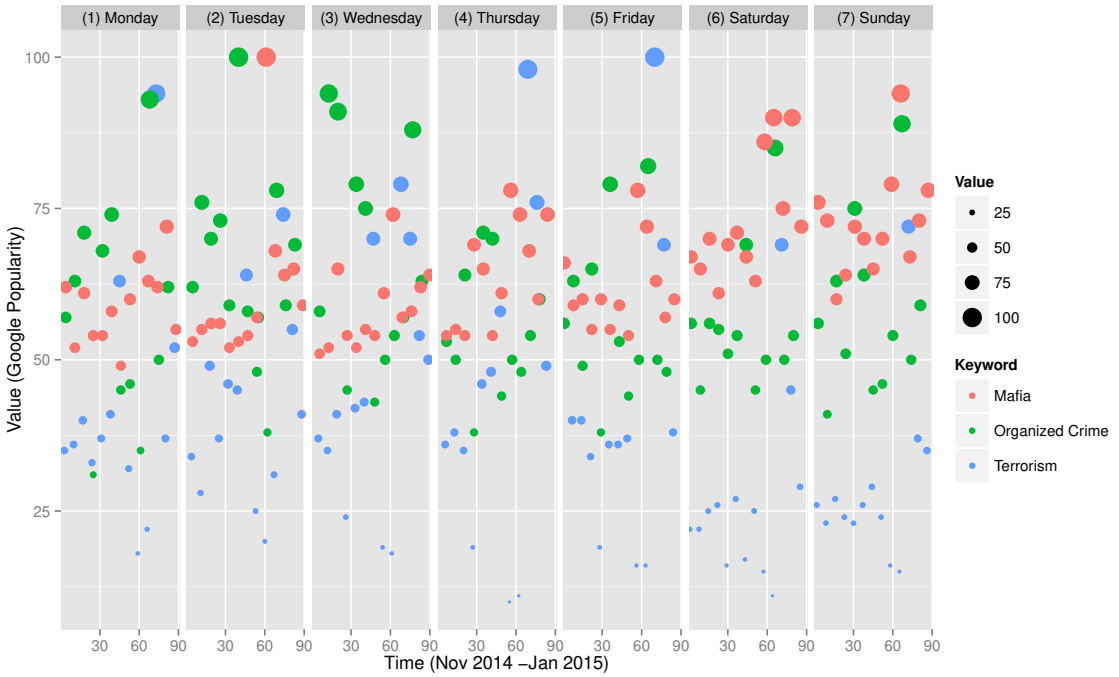
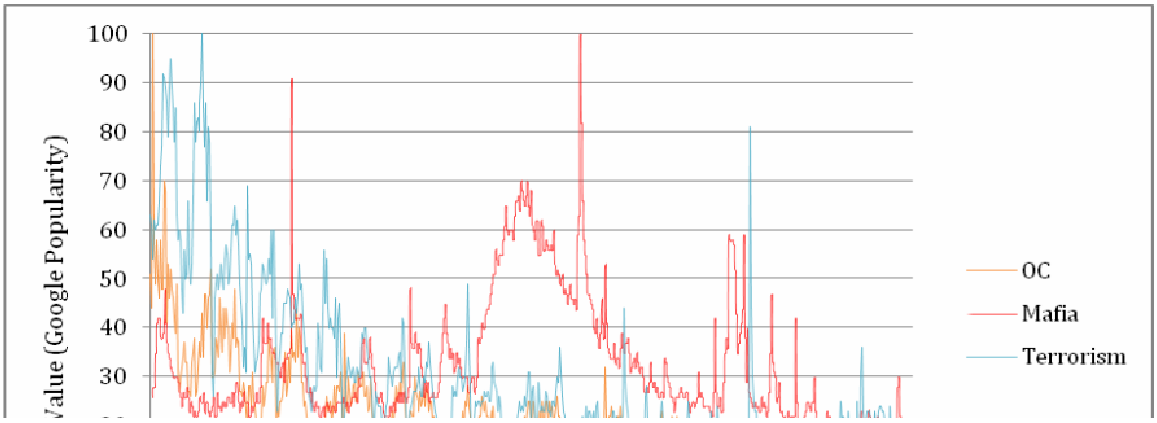



**Figure 5. Month-to-month variations web query popularity**




In figure 5 we present monthly Google Trends data for the three keywords depicted by a blue, red, and green line, respectively, with the index covering a period from 2004 to 2015. The results show that people's searching habits for "organised crime" and

"terrorism" have regularity over the course of time. As already noted in Figure 4, Google search queries for our sample keywords seem to be correlated with leisure and productivity. The monthly data for "organised crime" and "terrorism" results in an overlapping "M" like pattern. The trend line reaches high values every November and April while the downward trends picks up back around January and August. Finally, the monthly data for the keyword "mafia" do not exhibit strong signs of seasonality; therefore, further statistical tests are in order.

## Conclusion

Although not a novel method, 'big data' analytics can shed new light into collective perceptions of organised crime, the mafia, and terrorism. In summary, our findings are consistent with the suggestion that online behaviour data from web searches and social media contain a significant amount of structure that can feasibly provide insights into the way crowds perceive social phenomena. Numerous conclusions can be drawn from the preceding analysis.

First, automatic data extraction from Twitter was found to score well in terms of vocabulary consistency. Freely available online tools were used to control for contextomy. Second, a significant number of "mafia" tweets are linked back to the hip-hop community, and the number of tweets mentioning more than one keyword was insignificant. Third, broadcast messages were found to demonstrate extreme polarity. The number of users varies exponentially with the number of tweets per user, resulting in a typical long-tail distribution. Thus, we have found a large number of users with very few tweets and a small number of users who have posted a large number of tweets during the period examined in our sample. Overall, Twitter feeds about "terrorism" show a consistent power-law structure. The profiles of the most active users were manually scanned and categorised. The profiles of users sharing content about "terrorism" were also found to have a more institutionalised nature. Likewise, profiles linked to "mafia" tweets were predominantly linked to entertainment and leisure activities. These findings can be further explored to offer theories about how norms, moral panics, and perceptions spread across communication networks on the given topics. Fourth, the network topology was evaluated to account for how discourse is being shared and diffused within networks. It was found that the conversational structure of organised crime feeds resembles a tight-crowd network structure; mafia feeds resembles brand clusters; and terrorism feeds resemble community clusters. Finally, we evaluated discourse consumption as evident in web search volume—data from Google Trends have provided valuable insights. We proposed that search queries for "organised crime" and "terrorism" are mostly conducted on weekdays whereas "mafia" related web queries are popular during weekends. This seasonality is also evident in month-by-month web surfing habits. We have established that web search query variations for "organised crime" and "terrorism" are consistent and overlapping, while no seasonality was found for "mafia".

There are several problems and bottlenecks associated with the disaggregated social data used in this study. First, the social media data gathered from Twitter represent a static snapshot of the communication patterns between users over a rather limited time span. Further effort is required in order to collect longitudinal and balanced samples. Additionally, relational data and comments from other social media sites like YouTube may also provide valuable sources of information. Second, the primary focus of this study

was to establish the feasibility of a big data approach, so the social media data were randomly collected. However, social media data can be used to capture sentiments expressed by users in response to real-life events. Third, the analysis presented here lacks statistical rigor; therefore, further effort should be devoted to establish statistically significant correlations between the examined variables. Fourth, the extraction of disaggregated data should be automatized in order to avoid bias. Presently, the extraction of information from Twitter profiles was conducted manually; however, Natural Language Processing could be employed to enhance the analysis. Fifth, while both Twitter and Google Trends are rich in textual data, the present study explored only non-discursive aggregated trends. An avenue for further research can thus be the exploration of the linguistic features found in the texts produced by users. Finally, data from Google Trends and Twitter provide generous geographical information. Thus, web search queries and tweets can be further analyzed in terms of the location of users.

## Acknowledgement

*This work is (partially) funded by the European Commission grant 770469-CUTLER and 645706-GRAGE.*

## References

Allum, F. (2006). *Camorristi, politicians, and businessmen: the transformation of organized crime in post-war Naples*. Leeds: Northern Universities Press.

Allum, F. Longo, F., Irrera, D., & Kostakos, P. (eds). (2010). *Defining and defying organized crime: discourse, perceptions and reality*. London: Routledge.

Anderson, A., Goel, S., Huber, G., Malhotra, N., & Watts, D. J. (2014). Political Ideology and Racial Preferences in Online Dating. *Sociological Science*, *1*, 128-40.

Andriani, P., & McKelvey, B. (2009). Perspective-from Gaussian to Paretian thinking: causes and implications of power laws in organizations. *Organization Science, 20*(6), 1053-1071.

Arsovska, J. & Kostakos P. (2010). The social perception of organized crime in the Balkans: a world of diverging views? In F. Allum, F. Longo, D. Irrera, & P. Kostakos (Eds.), *Defining and defying organized crime: discourse, perceptions and reality* (pp. 113-131). London, Routledge.

Arsovska, J., & Michilli, A. (2015). Perceptions of Ethnic Albanians in New York City and the Role of Stereotypes in Fostering Social Exclusion and Criminality. The *European Review of Organised Crime, 2*(1), 24-48.

Bovenkerk, F. (1998). Organized crime and ethnic minorities: is there a link? *Transnational Organized Crime, 4*(3), 109-126.

Bovenkerk, F. Siegel, D., & Zaitch, D. (2003). Organized crime and ethnic reputation manipulation. *Crime, Law and Social Change*, *39*(1), 23-38.

Bryman, A. (2012). *Social research methods*. Oxford: Oxford University Press.

Cha, M., Kwak, H., Rodriguez, P., Yeol Ahnt, Y., & Moon, S. (2007). I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, 1-14.

Chambliss, W. J. (1971). Vice, corruption, bureaucracy, and power. *Wis. L. Rev*, *4*, 1150-1173.

Chin, K. L. (2000). *Chinatown gangs: Extortion, enterprise, and ethnicity*. Oxford: Oxford University Press.

Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. Economic Record, *88*(s1), 2-9.

Covington, S., & Bloom, B. (2003). Gendered justice: Women in the criminal justice system. In B. Bloom (Ed.). *Gendered Justice: Addressing Female Offenders* (pp. 3-23). Durham, NC: Carolina Academic Press.

Daniele, V., & Marani, U. (2011). Organized crime, the quality of local institutions and FDI in Italy: A panel data analysis. *European Journal of Political Economy, 27*(1), 132-142.

Daugherty, T. Eastin, M. S., Bright, L. F., & Chu S. C. (2011). Expectancy-Value: Identifying Relationships Associated with Consuming User-Generated Content. In Burns, N. M., Daugherty, T., & Eastin, M. (eds). *Handbook of Research on Digital Media and Advertising: User-Generated Content Consumption* (pp. 146-160). Hershey, PA, USA: IGI Global.

Décary-Hétu, D., & Aldridg, J. (2015). Sifting through the Net: Monitoring of Online Offenders by Researchers. *The European Review of Organised Crime, 2*(2), 122-141.

Eagle, N., Pentland, A., Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences, 106*(36), 15274-15278.

Ferrer-i-Cancho, R., & Elvevåg, B. (2010). Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS ONE, 5*(3), e9411.

Fond, G., Gaman, A., Brunel, L., Haffen, E., & Llorca, P. M. (2015). Google Trends®: Ready for real-time suicide prevention or just a Zeta-Jones effect? An exploratory study. *Psychiatry Research, 228*(3), 913-917.

Kostakos, V., & Ferreira, D. (2015). The Rise of Ubiquitous Instrumentation. *Frontiers in ICT*, October 25. Retrieved from http://journal.frontiersin.org/article/10.3389/fict.2015.00003/full (accessed 20 October 2015).

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation *Decision Support Systems*, *61*, 115-125.

Girardin, F., Calabrese, F., Fiore, F. D., Ratti, C., Blat, J. (2008). Digital footprinting: Uncovering tourists with user-generated content. *Pervasive Computing, IEEE*, 7(4), 78-85.

Gottschalk, P. (2013). Limits to Corporate Social Responsibility: The Case of Gjensidige Insurance Company and Hells Angels Motorcycle Club. *Corporate Reputation Review, 16*(3), 177-186.

Hand, E. (2011). Culturomics: Word play. *Nature, 474*(7352), 436-440.

Hill, D. (2010). A critical mass of corruption: why some football leagues have more match-fixing than others. *International Journal of Sports Marketing & Sponsorship, 11*(3), 221-235.

Howard, P. N., Duffy, A., Freelon, D., Hussain, M. M., Mari, W., Mazaid, M. (2011). Opening closed regimes: what was the role of social media during the Arab Spring? Project on Information Technology and Political Islam. Department of Communication, University of Washington.

Ianni, F. (1974). *Black Mafia: Ethnic succession in organized crime*. New York: Simon and Schuster.

Jerry, J., Steven, S. & Ralph, R. (2014). Assessing the success factors of organized crime groups: Intelligence challenges for strategic thinking. *Policing: an international journal of police strategies & management*, *37*(1), 206-227.

Kallus, N. (2014). Predicting crowd behavior with big public data. Proceedings of the companion publication of the 23rd international conference on World wide web companion, 625-30.

Kleemans, E. R., & Van de Bunt, H. G. (1999). The social embeddedness of organized crime. *Transnational Organized Crime, 5*(1), 19-36.

Kleemans, E. R., & Van de Bunt, H. G. (2008). Organised crime, occupations and opportunity. *Global Crime, 9*(3), 185-197.

Kostakos, V., Juntunen, T., Goncalves, J., Hosio, S. and Ojala, T. (2013). Where am I? Location archetype keyword extraction from urban mobility patterns. *PLoS ONE, 8*(5), e63980.

Kostakos, V., Nicolai, T., Yoneki, E., O'Neill, E., Kenn, H., & Crowcroft, J. (2009). Understanding and measuring the urban pervasive infrastructure. Personal and *Ubiquitous Computing, 13*(5), 355-364.

Krumm, J., Davies, N., & Narayanaswami, C. (2008). User-generated content. *IEEE Pervasive Computing, 4*, 10-11.

Liang, Y., Zheng, X., Zeng, D. D., Zhou, X., Leischow, S. J., & Chung, W. (2015). Characterizing Social Interaction in Tobacco-Oriented Social Networks: An Empirical Analysis. *Scientific Reports, 5*, Article number: 10060.

Lieberman, E., Michel, J., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature, 449*(7163), 713-716.

Liu, Y., Kostakos, V., Li, H. (2015). Climatic Effects on Planning Behavior, *PLoS ONE, 10*(6), e0131954.

Makin, D. A., & Morczek, A. L. (2015). The Dark Side of Internet Searches: A Macro Level Assessment of Rape Culture. *International Journal of Cyber Criminology, 9*(1), 1.

McGlone, M. S. (2005). Contextomy: The art of quoting out of context. *Media, Culture & Society, 27*(4), 511-522.

Mcillwain, J. S. (1997). From tong war to organized crime: revising the historical perception of violence in Chinatown. *Justice Quarterly, 14*(1), 25-52.

Mendoza, A. A., (2015). Nociones de justicia, legalidad y legitimidad de las normas entre jóvenes de cincopaíses de América Latina. *Sociedade e Estado, 30*(1), 75-97.

Michel, J-B, Shen, Y., Aiden, A., Veres, V. and Gray, M., The Google Books Team, Pickett, J., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M., & Lieberman, A. E. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science, 14*, 176-82.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics, 46*(5), 323-351.

Obgar, J. (1999). Slouching toward Bork: The Culture Wars and Self-Criticism in Hip-Hop Music. *Journal of Black Studies, 30*(2), 164-183.

Ouimet, M., & Montmagny-Grenier, C. (2014). Homicide and Violence—International and Cross-National Research. The Construct Validity of the Results Generated by the World Homicide Survey. *International Criminal Justice Review, 24*(3), 222-234.

Paoli, L. (2003). Mafia brotherhoods: Organized crime, Italian style. Oxford: Oxford University Press.

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific reports*, *3*, 1684.

Pruss, S. B. (2014). The German Medias Portrayal of Ethnic Organised Crime and Its Implications. *The European Review of Organised Crime, 1*(2), 97-118.

Reynolds, D. (2011). Manipulating perceived risk to deter and disrupt counterfeiters. *Journal of Financial Crime, 18*(1), 105-118.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. Proceedings of the 19th international conference on World Wide Web, 851-60.

Sarno, F. (2014). Italian mafias in Europe: between perception and reality. A comparison of press articles in Spain, Germany and the Netherlands. *Trends in Organized Crime, 17*(4), 313-341.

Schneider, P. T., & Schneider, J.C. (2003). *Reversible destiny: Mafia, antimafia, and the struggle for Palermo*. California: University of California Press.

Seifter, A., Schwarzwalder, A., Geis, K., & Aucott, J. (2010). The utility of Google Trends for epidemiological research: Lyme disease as an example. *Geospatial health, 4*(2), 135-137.

Shen, A., Antonopoulos, G. A., & Papanicolaou, G. (2013). Chinas stolen children: internal child trafficking in the Peoples Republic of China. *Trends in organized crime, 16*(1), 31-48.

Smith, M.A., Rainie, L., Shneiderman, B., & Himelboim, I. (2014). *Mapping twitter topic networks: From polarized crowds to community clusters*. Washington: Pew Research Center.

Smith, D. (1975).The Mafia Mystique. London: Hutchinson.

Louis, C., & Zorlu, G. (2012). Can Twitter predict disease outbreaks? *BMJ*, 344:e2353.

Sullivan, D. (2013). Google still world's most popular search engine by far, but share of unique searchers dips slightly. Search Engine Land, February 11. Retrieved from http://searchengineland.com/google-worlds-most-popular-search-engine-148089.

Sung, H-E (2004). State failure, economic failure, and predatory organized crime: A comparative analysis. *Journal of Research in Crime and Delinquency, 41*(2), 111-129.

Sutter, C. J., Webb, J. W., Kistruck, G. M., & Bailey, A. V. (2013). Entrepreneurs' responses to semi-formal illegitimate institutional arrangements. *Journal of Business Venturing, 28*(6), 743-758.

Tilley, N., & Hopkins, M. (2008). Organized crime and local businesses. *Criminology and Criminal Justice, 8*(4), 443-459.

Travaglino, G. A., Abrams, D., Randsley de Moura, G., & Russo, G. (2015). That is how we do it around here: Levels of identification, masculine honor, and social activism against organized crime in the south of Italy. *European Journal of Social Psychology, 45*(3), 342-348.

United Nations (2015). United Nations News Centre - UN projects 40% of world will be online by year end, 4.4 billion will remain unconnected. October 26, 2015, Retrieved from http://www.un.org/apps/news/story.asp?NewsID=46207.

Van Dijk, V. J. (2007). Mafia markers: assessing organized crime and its impact upon societies. *Trends in organized crime, 10*(4), 39-56.

Van Dijk, J. J. M., Mayhew, P., & Killias, M. (1990). *Experiences of crime across the world: key findings from the 1989 international crime survey*. Boston: Kluwer Law and Taxation Publishers.

Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: survey-based indicators vs. Google trends. Journal of Forecasting, *30*(6), 565-578.

Wang, X., Brown, D. E., & Gerber, M. S. (2012). Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information. 2012 IEEE International Conference on Intelligence and Security Informatics (ISI), 36-41.

Williams, M., & Levi, M. (2012). Perceptions of the eCrime controllers: Modelling the influence of cooperation and data source factors. *Security Journal, 28,* 252–27.

Young, A. B., & Allum, F. (2012). A comparative study of British and German press articles on organised crime (1999-2009). *Crime, Law and Social Change, 58*(2), 139-157.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge: Addison-Wesley Press.

Farmer, D. (2013). Google Search scratches its brain 500 million times a day, Cnet, May 13, Retrieved from: http://www.cnet.com/news/google-search-scratches-its-brain-500-million-times-a-day.

Yamaguchi, Y., Takahashi, T., Amagasa, T., & Kitagawa, H. (2010). TURank: Twitter User Ranking Based on User-Tweet Graph Analysis. Web Information Systems Engineering – WISE 2010, 11th International Conference Proceedings, 240-253.

Zubiaga, A., Spina, D., Martínez, R., & Fresno, V. (2014). Real-time classification of twitter trends. *Journal of the Association for Information Science and Technology, 66*(3), 462-473.