Knowing Where and How Criminal Organizations Operate Using Web Content

Michele Coscia
KddLab - ISTI CNR
Via G. Moruzzi 1, Pisa
michele.coscia@isti.cnr.it

Viridiana Rios
Department of Government - Harvard University
1737 Cambridge St, Cambridge, MA, US
vrios@fas.harvard.edu

ABSTRACT

We develop a framework that uses Web content to obtain quantitative information about a phenomenon that would otherwise require the operation of large scale, expensive intelligence exercises. Exploiting indexed reliable sources such as online newspapers and blogs, we use unambiguous query terms to characterize a complex evolving phenomena and solve a security policy problem: identifying the areas of operation and *modus operandi* of criminal organizations, in particular, Mexican drug trafficking organizations over the last two decades. We validate our methodology by comparing information that is known with certainty with the one we extracted using our framework. We show that our framework is able to use information available on the web to efficiently extract implicit knowledge about criminal organizations. In the scenario of Mexican drug trafficking, our findings provide evidence that criminal organizations are more strategic and operate in more differentiated ways than current academic literature thought.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering

General Terms

Economics, Human Factors

Keywords

query, data retrieval, knowledge discovery process

1. INTRODUCTION

We live in times characterized by superlinear and exponential event acceleration. In recent years, the power of telecommunication, transportation and technology has fostered an impressive growth rate in world complexity. The number of Web pages has increased from 11.5 billion in 2005 [1] to at least 25.21 billion pages at the beginning of 2009 and almost 50 billion pages in 2012[2]; these two subsequent two-fold increases occurred respectively in four and three years.

Information complexity critically affects the ability of security agencies to collect intelligence information by making it more costly. To bring the benefits of tracking complex phenomena to those lacking the resources to conduct large-scale intelligence collection we develop a tool that uses the vast amount of knowledge present on the Web to obtain quantitative information about criminal activities. Exploiting some already indexed reliable sources such as online newspapers and blogs, we develop a mechanism that uses unambiguous query terms to identify the areas of operation of criminal organizations and their characteristics. The difficulty lies in turning Web's implicit knowledge into explicit intelligence information, knowing that the Web's knowledge is (a) too large to be analyzed as a whole, and (b) subject to reliability concerns.

In this paper, we prove that our framework is not only inexpensive and relatively easy to use, but also provides an effective way to obtain intelligence data that is useful for real-world applications. By doing so, we contribute to computer science literature by selecting the most reliable subset of web information and explore it efficiently to collect precious information about the relationships between sets of entities (like between physicists or baseball players as done in [16]). We describe this framework and we call it MOGO (**M**aking **O**rder using **G**oogle as an **O**racle). We also contribute to social sciences literature, we prove MOGO's usefulness, we apply it to identify the municipalities in which Mexico's drug trafficking organizations operate, yearly between 1990 and 2010. With more than 51,000 victims of drug-related violence from 2007 to 2011, it is safe to say that no other country in Latin America has a higher need for research on criminal behavior. We provide the first empirical data available about this complex problem, one that has not been properly studied due to a lack of public data on where and when Mexican drug trafficking organizations operate.

The structure of the paper is as follows. In Section 2 we review previous works about our crawling approach and related studies about organized crime in Mexico. Our data retrieval framework is described in Section 3. Section 4 reports the statistics about the raw data we downloaded from

[1] http://www.divms.uiowa.edu/~asignori/papers/the-indexable-web-is-more-than-11.5-billion-pages/
[2] http://www.worldwidewebsize.com/

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

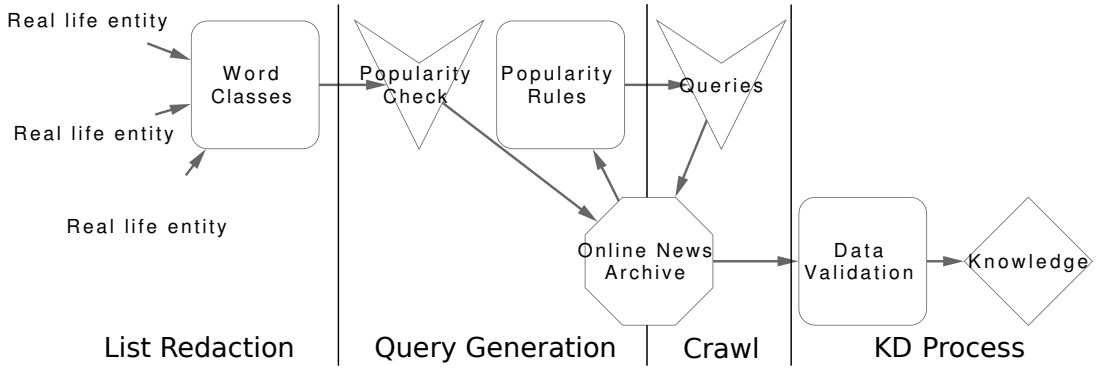


Figure 1: The workflow of MOGO's framework.

Google News. We explain how we clean data in Section 5 and how to extract useful knowledge from it in Section 6. Section 7 concludes the paper.

2. RELATED WORKS

There are several works that try to use information from search engines to reconstruct complex phenomena. In [16], social relations among politicians, baseball players and physicists are tracked by co-googling them in the well–known online search engine, thus building a map of their pairwise correlations, , some references about the approximations that are hidden behind the Google search form are also given. Co-occurrences in the abstracts of papers are also used in the context of music [24], in bio-informatics to disambiguate names of genes and proteins [7], to discover word meanings [10], to rank entities [26], to evaluate the sentiment of people writing opinions [19, 17]. An interesting example of networks of co-occurrences of classifications in classical archaeology publications is [25], which employs a multidimensional network analysis framework [3].

Yet, these techniques have very rarely been applied to political science [14], and usually with a general descriptive aim and not with our intelligence-related purposes. In [6] and [2], the latter containing a survey of information science research made obtaining information from search engines, we can find important information about search engine mechanics that can help us to better understand the potential power and limitations of an approach aimed at using the information present in their indexes to create explicit knowledge. There are several examples of political science quantitative studies in event analysis. An example of such a system is provided in [15]. Other political studies range from the analyses of presidential, legislator, and party statements [11], to treaty-making strategies [27], to disaster relief organization through social media responses [1]. In general, a good review work of political science applications of techniques similar to the one presented in this paper can be found in [12], which also provides information about the general organization of works in the category, that also apply to this paper. None of the reported methods take advantage of the freely available information present in the web from reliable sources like the newspapers indexed by Google News.

As our paper focuses on the Mexico's drug trafficking industry, we provide some literature references to back up our findings. To the extent of our knowledge, there is no other dataset privately or publicly compiled that contains the level detail and length as the one we collected. Private efforts like Stratfor[3] and Guerrero [13] have provided information on the territories of operation of drug trafficking organizations but only at the state level and without time variation. Mexican secret intelligence office (CISEN) has information at the municipal level but it is not available for research purposes and does not provide information for years before 2006 [4].

3. WORKFLOW

In this section we present the workflow of our general framework. We begin by defining our terminology. We named our framework MOGO. In MOGO, an **actor** is a real world entity that is an active or passive part of the phenomenon we want to study. Actors can be of different **types**. For example, since we study the Mexican drug traffic, we have two types of actors: the traffickers (active) and the municipalities (passive). An **actor list** is the list of the different actors of the same type (i.e. the list of traffickers and the list of municipalities). Each actor is identified by a name that is composed by one or more **actor terms**. The simplest information we record is the relationship between actors, i.e. a **couple**: any combination of two actors from different types. The medium we use to get this information is a **query**. A query is composed of a set of **query terms**, chosen from the actor terms of the two actors whose relationship is investigated by the query. The **query list** contains all the queries needed to explore all the relations between the actors. Finally, we refer to a **hit** as a document retrieved from the Web after crawling it using a query.

We designed MOGO to work in three steps. First, we define the types of actors we will study and create actor lists. Then, we combine the various lists into a non-ambiguous set of queries. Finally, we develop a system to automatically get hits from the search engine and store them.

Figure 1 represents the high-level logic of MOGO. We first operate a classification of the actor terms (Section 3.1). Once we have a representation classification, we make a prelimi-

[3]Link needed.

[4]We estimate that about 63 % of CISEN's data had at some point been covered by Google news. This estimate comes from comparing a dataset of personal communications between traffickers that we collected from the web to the same dataset collected by CISEN. Out of a total of 1421 communications collected by CISEN, 888 were reported at Google News. We took this as a reference of the amount of CISEN's information that is available at the web

nary invocation of our oracle (the online news archive) to check which are the actor terms that lead to the least noise. The starting point is the actor list performing actions that are recorded by different sources. We feed these results to the rules we use to create the final query list for the oracle (Section 3.2). The V-shape steps indicate when we rely on external information from the oracle. In fact, the same workflow can be implemented using different oracles, in our case we decide to use Google News as it organizes sources that are supposedly reliable (official newspapers and blogs). We then query the oracle with our crawler (Section 3.3). Finally, we use the raw data provided by the oracle, feeding a standard knowledge discovery process (Sections 4, 5 and 6). We represent the KD process giving particular importance to the validation part, as we follow the approach described in [12].

In the following subsections our framework is developed in detail. Each part of the framework for our case is tackled in a different subsection. The redaction of the actor lists is described in 3.1; the generation of the rules to create the query list is presented in 3.2; finally we describe our search engine crawler in 3.3.

3.1 List Redaction

To generate the query list, we first need the different actor lists for each actor type that we are interested in studying. We could also have one list, if we are interested in a simple co-occurrence investigation. For our case study, two different actor lists were required: a list of municipalities and a list of drug traffickers (both individuals and organizations). Some of the strategies that we used to create these lists are general and work for tackling ambiguity and syntactic issues in search engines.

For both lists, we developed a strategy to deal with special characters. We are dealing with Spanish-speaking newspapers, thus with non-standard ascii characters. Accentuated words (e.g. Apatzingán) were searched with and without accents. Spanish characters like "ñ" were searched as such (e.g. Acuña).

For the municipality list, the general aim was to generate at least one query per municipality as existent in Mexico's 2010 Census figures (there are 2,456 of them, organized in 31 states and the federal district of Ciudad de México). To do so, we first labeled each single actor term in the dataset in four different classes. Class "Generic" is a standard stopword list for Spanish language. Class "Common" includes popular proper names. Class "Unique" includes all the actor terms that are not Generic or Common and that occur only one time in our dataset. Class "State Unique" includes all the actor terms that are not Generic, Common or Unique, but occur only one time in a particular Mexican state, so they are unambiguous if we also specify the state. In Section 3.2 we show how each of these classes were used to generate the query list.

For the lists of traffickers and criminal organizations, we did not create classes. We needed to query the full name of the trafficker even if his name is completely included in the Common class. If this was the case, to avoid ambiguity we added "narcotrafico" to filter out all people with the same name but not related to the drug trafficking industry. Also, each trafficker name was associated with a particular nickname, as it is common in the criminal world to refer to each other without using official names.

The final lists include 2,449 locations. We have 176 actor terms associated with traffickers or drug trafficking organizations classified, according to their affiliation, in 13 criminal organizations and a residual category. So each trafficker is associated with a criminal organization and a criminal organization is simply a list of traffickers.

3.2 Query Rules Generation

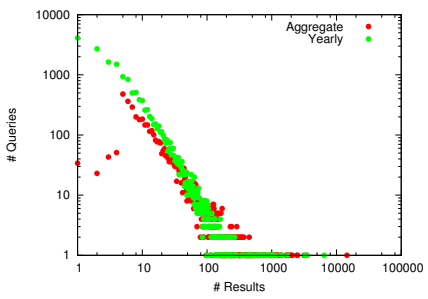
Once we defined the actor list for each type, we generated the query list from them. We needed to have at least one associated query per couple. Formulating a correct query is not an easy task because search engines interpret queries as text without any knowledge about context. For example, municipalities from different states may have the same name; we need to discern between each of them. Additionally, we need to distinguish between a municipality called Valencia the Spanish city of Valencia.

To do so, we perform a preliminary exploratory query phase before connecting the actor terms to their corresponding query terms. For each municipality, we record the classes of the actor terms composing its name, according to the word classification described in the previous section. Then, we apply a cascade of rules. We now provide the list of rules used in our case study. Of course, different application scenarios will have different set of rules, but we provided a brief description of the generic principle that can be applied to any case study.

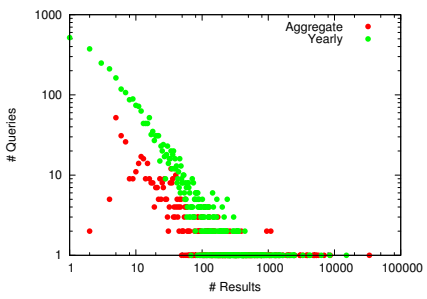
- If a municipality contains a Unique, the corresponding query term is the exact actor term, as it is non-ambiguous.
- If a municipality has a State Unique, the query term is the actor term of the municipality plus the name of its state.
- If a municipality completely contains the name of another municipality (e.g. "Valle de Chalco" completely contains "Chalco"), the query term is the contained municipality extracting the extra characters of the larger-name municipality (e.g. query <Chalco NOR "Valle de"> for Chalco, and for "Valle de Chalco" for Valle de Chalco).
- If a municipality has the same name of a state, the query term is the actor term plus ("municipio de" OR "ciudad de").
- If several states contain municipalities with the same name, the query terms for such municipalities are <municipality name AND state NOT other states), where other states refer to states that have similarly named municipalities as the one which is being queried.

For each trafficker term we applied this set of rules:

- We perform several searches for each trafficker with these different query term schemes: <"first name AND father's last name AND mother's last name">, <"father's last name AND mother's last name">, <"father's last name AND mother's last name"> AND <alias>, <"first name AND father's last name AND mother's last name"> AND <alias>.
- If a trafficker's actor terms are composed, then the corresponding query terms are all its possible composed



(a)



(b)

Trafficker	Total Hits
Sinaloa	57,988
Zetas	37,992
Golfo	28,837
Beltrán Leyva	27,824
Tijuana	17,987
Familia	15,448
Juarez	11,817
Teo	4,505
Azul	2,656
Barbie	2,418
Others	8,961

(c)

Figure 2: (a) The hits distributions aggregated as (trafficker, municipality) and dis-aggregated as (trafficker, municipality, year); (b) Hits distribution per municipality and dis-aggregated as (municipality, year); (c) Hit distribution per trafficker.

and decomposed combinations (e.g. Juan José Esparragoza was queried as <"Juan AND José AND Esparragoza" AND "El Azul">, <"Juan AND Esparragoza" AND "El Azul"> and <"José Esparragoza" AND "El Azul">).

- Aliases are queried along with Commons.
- Query alias along with criminal organization to which trafficker belongs (e.g. "El Chapo" AND "Cartel de Sinaloa")
- If a trafficker has more than one alias, query each alias separately.
- If an alias has an alternative meaning, query it as <alias AND (narcotráfico OR cartel)>

Both for municipalities and for traffickers, these rules generated several alternatives. All the alternatives that end up being composed only by combinations of actor terms labeled as Common or Generic were eliminated. The others were all queried.

For each query term we created a popularity score which is its number of hits. For municipalities we also defined a "Popularity per capita" score measured by dividing the popularity of the query by the population of the municipality the query represents. Then the following rules were applied:

- Reject all queries with Popularity $\geq 100,000$ (for queries referring to municipalities), $\geq 10,000$ (for cartel leaders or criminal organizations) or ≥ 500 (for trafficker lieutenant).
- Reject queries with Popularity per capita > 3.
- Reject all queries with Popularity (for traffickers) or Popularity per capita (for municipality) $\geq 500\ \sigma$, i.e. the one whose standard deviation from the average Popularity per capita of all queries was considered too high.

The rationale behind these rules is founded on the consideration that huge deviations from the average in popularity are caused by actors whose actor terms are shared with actors not included in our targeted actor list. Our assumption is that it is the existence of ambiguity in the actor terms what is unfairly driving actor's popularity up.

When after applying these rules, each couple had more than one possible query, we select the one leading to the largest popularity. We now have our query list, one query per couple, to feed our API crawler.

3.3 The Crawler

The crawler is very simple and implementations of it can be done in different programming languages. The steps of the crawler are:

1. Get one query from the query list;
2. Create the URL of the Google API invocation, by escaping special characters in the query and collecting the API keys;
3. Connect to the Google API system and request the page, obtaining a JSON;
4. Parse the JSON answer. Since Google news provides also a timeline, we can get the results for the years 1991-2010 with just one API invocation;
5. Get the number of hits from the parsed results, store them and wait a courtesy time interval to not overflow the service with too many requests.

Our implementation of the crawler was created entirely in Python. There are not concerns about its efficient time and memory implementation. For time, the main source of inefficiency is the network connection from the computer running MOGO and Google's server. As for memory, JSON objects are very small and there is no need to keep in memory more than one of them.

4. DATA STATISTICS

We now provide some statistics about the data retrieved for our case study.

First, in Figure 2 we present the distribution of hits per query. We depicted the results per couple both aggregate over the entire time window and disaggregate per year. We see a fat tail distribution in both cases because each year, there are many municipalities with a very low drug-related criminal activity and some "hubs" where criminal activity is concentrated that record between 1,000 and 10,000 hits. In the aggregate distribution, we observe also that the number

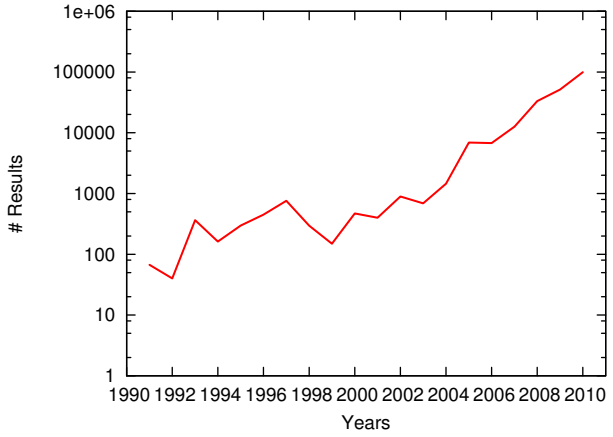


Figure 3: The results distributions per year.

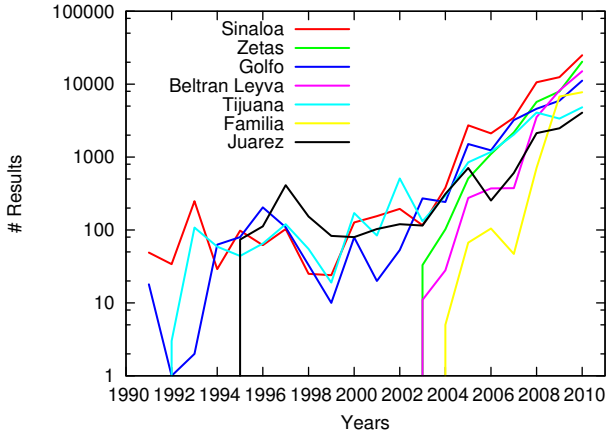


Figure 4: The results distributions per trafficker per year, for the most popular seven traffickers.

of couples with few hits are less than expected. This indicates that the pattern of activity is not constant. Municipalities that do not show any activity in one year often start to experience activities in our 20-year interval. To confirm this hypothesis, we plot in Figure 2b the same hits distributions aggregating over drug trafficking organization. Each dot is now either a municipality (red dots) or a municipality in a year (green dots). The same pattern emerges.

Second, in Table 2c we report trafficking organizations sorted by popularity (we report only the top 10 over the 13 trafficking organizations, aggregating the remaining as "Others"). Given the few data points, we do not observe a fat tail distribution, but the number of hits is heavily unbalanced. The top 4 trafficking organizations (out of 13) account for more than 70 % of the total amount of hits.

These pictures suggest that we are observing a complex system. A complex system is a system composed of different parts that expresses at the global level properties that are not present in any single part taken alone. The exponential growth, and other typical complex system characteristics such as the presence of uneven (power-law) distributions in the number of hits [18], are hint confirming this observation. These findings suggest that to understand Mexico's drug trafficking industry it is not useful to study each single traffic organization, territory and law enforcement organization, i.e. the parts of the system, but rather how these parts interact as a whole (in other words, one organization with each other organization).

So far we have looked at a static perspective over our data. Next, we present some views about the temporal evolution of the observed phenomenon. In Figure 3 for each year we report the number of hits for all queries. This picture is backing the fact that the phenotype of Mexico drug traffic is growing beyond control (please notice the logarithmic y axis). We expect some downward bias for years before 2006, while Google News was still in beta and when the collection of articles in years previous to 2006 may have been incomplete. However, if there were no superlinear growth, one would expect the line to be stable after 2006. Instead, it is exactly from 2006 onward that we witness the most incredible growth, jumping one order of magnitude (from 10,000 to 100,000 articles) in just four years.

In Figure 4 we disaggregate the previous figure by drug trafficking organization (for clarity purposes, this time we select the top 7 organizations reported in the majority of documents). Some organizations appear slightly before their foundation. Zetas, for example, have a low initial popularity (below ten hits) but grow impressively, catching up with older trafficking organizations in just a few years. In general, also old trafficking organizations such as Sinaloa and Golfo, grow superlinearly. These hubs and their dynamic behaviors are one of the main characteristics of Mexican drug traffic as complex systems, a self balancing and organizing organism that cannot be understood by tackling each organization singularly.

5. DATA CLEANING

Extracting information from online search engines, although using reliable sources like newspapers, is an operation that introduces data loss, incompleteness and noise. Thus, retrieving hits connecting two different actors does not necessarily mean that the two actors are involved in an actual relationship in the real world. We developed a way to clean and validate our hits in order to transform them into meaningful results that describe reality. To clean, we normalized the total number of hits we are getting using a hyper-geometric cumulative distribution function (Section 5.1). To validate, we compared information from other cases, cases in which information is known with certainty, with the one we extracted using MOGO. In particular, we used MOGO to identify the territories of operation of Mexican state governors with the expectation of finding them operating only in the states where they rule (Section 5.2).

5.1 Hyper-geometric Distribution Normalization

Our goal is to overlap a trafficking organization t_i and a municipality m_i in order to identify municipalities where drug trafficking organizations actually operate, cutting out noise.

We can draw a parallel between the null model we are trying to identify and the extraction of a labeled ball from a bin. The municipalities would be the bins, and the trafficking organizations would be the labelled balls. Thus, the number of balls we have for each t_i is equal to the total number of

hits that all the queries related to t_i returned, a figure that we represent as $\mathcal{T}_i$. The question we are asking is whether the number of times a ball is seen in a bin is larger or smaller than what it would be expected by chance, given both the number of balls and bins[5]. Since the total number of balls we have is limited and equal to the number of times a t_i appears overall, the null model we are looking for is a hypergeometric probability distribution.

The hyper-geometric distribution is a probability distribution describing the probability of extracting $\bar{t_i}$ times in $\bar{m_i}$ attempts a ball labelled with t_i, given that the total number of such balls is $\mathcal{T}_i$, from a total set of balls equal to $\mathcal{M}$ [22]. The corresponding probability mass function PMF for t_i is defined as follows:

$$PMF(t_i = \bar{t_i}) = \frac{\binom{\mathcal{T}_i}{\bar{t_i}}\binom{\mathcal{M}-\mathcal{T}_i}{\bar{m_i}-\bar{t_i}}}{\binom{\mathcal{M}}{\bar{m_i}}}$$

As an example, assume $\mathcal{T}_i = 5$, i.e. we obtained 5 hits for t_i, that is recording a total of 50 total hits ($\mathcal{M} = 50$). Considering municipality m_i, with a total number of 10 hits ($\bar{m_i} = 10$), we find that among those 10 hits, 4 were related to t_i ($\bar{t_i} = 4$). The corresponding probability of this happening is equal to $\frac{\binom{5}{4}\binom{45}{6}}{\binom{50}{10}}$ or approximately 0.00396.

Our final results, what we actually use as our final real-world information, is not the exact probability of obtaining $\bar{t_i}$ results for t_i in m_i but rather the probability of obtaining $\bar{t_i}$ or less results, that is the cumulative distribution function (CDF) is defined as:

$$CDF(t_i = \bar{t_i}) = \sum_{a=0}^{\bar{t_i}} PMF(t_i = a),$$

which takes values from 0 to 1.

The reason is that the probability mass function is not monotonic, while the cumulative distribution function is. If t_i is very popular, it is difficult to find few documents referring to it in a popular municipality. The probability grows and peaks to the expected value of pure chance. Then it starts becoming lower, as it is difficult that all documents related to t_i appear in the same municipality. With a non monotonic function, we cannot have a simple rule stating "If the function value is low, the relation is significant", because low values can be generated both by particularly strong or particularly weak relations. In other words, $PMF(t_i = n) < PMF(t_i = n+1)$ does not hold. Instead, being the cumulative distribution function monotonic (i.e. the following relationship holds: $CDF(t_i = n) < CDF(t_i = n+1)$), we can say that the higher the value, the stronger the relation (until the theoretical maximum of 1).

5.2 Validation

We cannot present any result of MOGO as accurate unless we validate it against a reliable ground truth. Since the criminal organizations operate secretly, we do not have any reliable knowledge about their areas of operation. For this reason, we choose to test MOGO in a slightly different problem for which we have an evident and reliable knowledge base. We applied it to the study of the municipality relations with Mexican politicians, i.e. we substituted drug trafficking organizations with Mexican state governors. If MOGO is returning real relationships between actors and places, each state governor should be mainly found to operate in the municipalities of the state she is governing.

In Figure 5 we depict the relations between the municipalities and three governors[6]. We highlighted in red all the municipalities for which the chosen governor has a CDF equal to or higher than 0.95. We also highlighted the state governed by the politician by enlarging the thickness of its borders. In Figure 5(a) we highlighted the relationships between municipalities and the governor of Chiapas, in Figure 5(b) the governor of Chihuahua and finally in Figure 5(c) the governor of Durango. We can see that in all the three cases the governors are related mostly, if not entirely, to the municipalities of their state. Furthermore, the amount of municipalities they are related which are outside their state is very low.

Our conclusion is that the relationships extracted with MOGO are an accurate depiction of the explicit relations between someone, or something, operating on the territory and the territory itself. We are now able to provide some examples of the usefulness of the knowledge extracted in the description of the Mexican drug war.

6. KNOWLEDGE EXTRACTION

6.1 Trafficker's Activities

The results of the search allowed us to provide the research community with information on the behaviour of 13 trafficking organizations in Mexico, particularly about their municipalities of operation, their migration patterns and their market strategies for a period of 19 years (1991 - 2010).

The disaggregation up to the municipal level allowed us to challenge the widespread assumption that drug traffickers control vast regions of Mexico's territory by dividing the country in oligopolistic markets [23]. Instead, we show that traffickers select their areas of operation with finer detail (Figure 6a). The Sinaloa Cartel, a drug trafficking organization that was previously thought to operate in the entire state of Sinaloa [20], only operates in 14 of the 18 municipalities of the state; the same case goes to Juárez Cartel that only operates in 28 of 66 municipalities at its home state (Chihuahua), and to La Familia that only operates in 69 of 115 Michoacán.

Actually, according to our results, drug trafficking organizations only operate in 713 of 2,441 municipalities in Mexico. Large areas the country completely lack of the presence of a drug trafficking organizations. Our data changes our understanding of criminal territoriality, showing that drug trafficking organizations pick their areas of operation quite selectively. They concentrate in areas that are closer to ports of entry to the US, large cities within Mexico, and highways that connect cultivation areas or maritime ports to the US-Mexico border (Figure 6b).

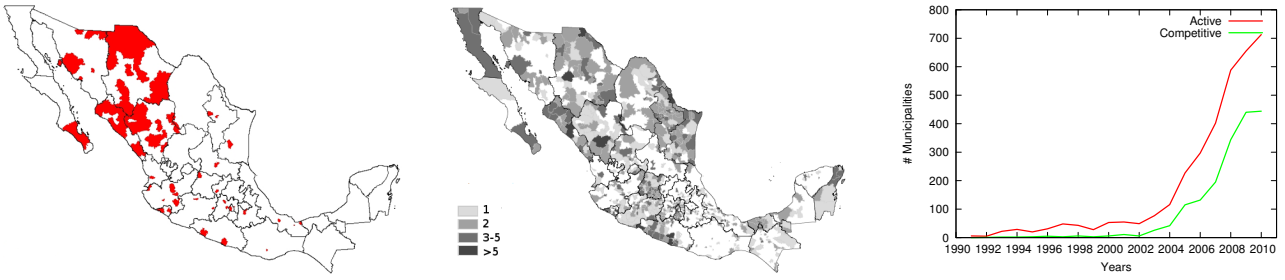
This insight matches evidence of criminal activity tracked by the Mexican government in the form of drug-related violence. The Trans-Border Institute analyzed official figures of homicides caused directly or indirectly by the activities

[5]We assume each t_i is independent

[6]At the time in which we performed the validation searches, i.e. November 2011

(a) Governor of Chiapas. (b) Governor of Chihuahua. (c) Governor of Durango.

Figure 5: The politician-municipality significant relations. The corresponding state has been highlighted with a ticker black border.



(a) Sinaloa-active municipalities. (b) The total active municipalities in Mexico. (c) Active and Competitive municipalities.

Figure 6: Some evidences about the traffickers' activity patterns.

of drug trafficking organizations in Mexico[7] and concluded that less than 44 % of the municipalities in 2011 had experienced drug-related homicides, a large increase from the 15 % municipalities presenting this type of violence in 2007 but still away from arguing that criminal organizations operate in all of the 2,500 municipalities of Mexico.

Our results also provide the first systematic evidence of changes in the territoriality of criminal organizations over time. Rather than providing cross-sectional data at a particular point in time as [13], we show panel data for almost two decades (Figure 6c). The increase in the number of municipalities with drug traffic activity is evident. Furthermore, our information provides the first portrait of the market structure of the illegal drug trafficking within Mexico and of its changes over time. Mexico's organized crime is not the oligopoly the theoretical literature of organized crime and private protection rackets assumes; rather, drug trafficking organizations share territories frequently. As of 2010, 444 (62 %) of all municipalities with trafficking operations had more than one criminal organization operating simultaneously, a significant increase from a decade ago when only 6 (11 %) were competitive (Figure 6c). The market structure of organized crime in Mexico is increasingly competitive; many criminal groups operate in the same territories, sharing accesses to highways and ports of entry to the US.

Because our results were designed to provide information about criminal operations by organization, we are able to show the inner behavior of criminal organizations at a level of detail that was previously unknown. We depart the assumption that all organized groups are the same, and that their motivations can be understood under the same logic, and instead provide evidence of important operational differences between them. Criminal groups differ quite substantially in their territorial extension, the strategies that follow to expand, and their patterns of migration.

[7]http://justiceinmexico.files.wordpress.com/2010/07/2012-tbi-drugviolence.pdf

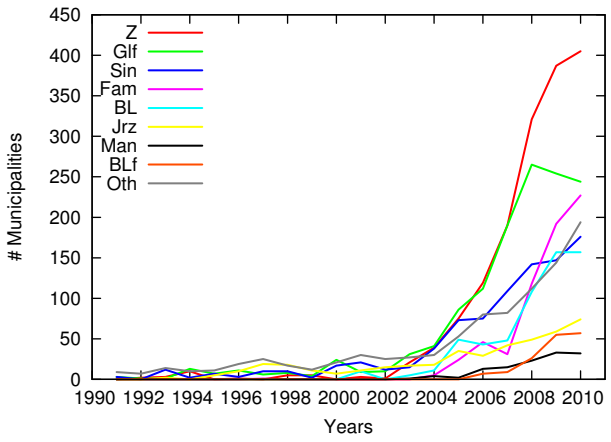


Figure 7: Different activity patterns per trafficker.

The timing, velocity and degree of expansion are significantly different between different organizations (Figure 7a, please note that this figure is different from Figure 3 as the latter included raw results before the cleaning stage). The most expansionary drug trafficking organizations start-

ed their spread in 2003 (Golfo, Zetas), a second group took off in 2005 (Sinaloa, Beltran-Leyva (BL) and Familia), and a final one, which included most of the criminal organizations in Mexico took place in 2007[8]. The most rapid expanders, organizations like La Familia, three-folded the municipalities in which they operated in just two years; slower expanders, like Sinaloa, took six years to expand similarly. Before 2004, all trafficking organizations operated in less than 50 municipalities. By 2010, Zetas had expanded their presence to more than 400, while others like Juárez Cartel have never operated in more than 80 municipalities. The number of municipalities in which each trafficking organization operated as of 2010 and the year in which they stated operating is shown in Table 1 (first two columns).

6.2 Identifying criminal organizations' phenotypes

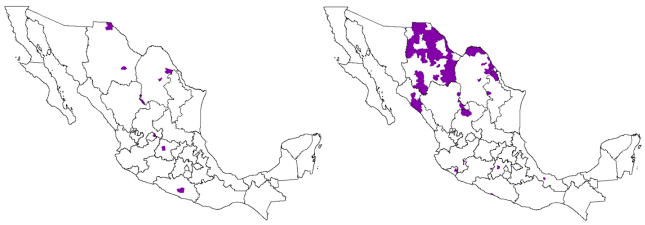
Using the information provided by MOGO we can classify Mexico's drug trafficking organizations according to their market strategies.

Table 1 reports identified characteristics of the Mexican trafficking organizations. Two of them (number of municipalities in 2010 and year of appearances, first and second columns) were already discussed in the previous subsection. The remaining five are: average number of municipalities in which the trafficking organization starts to operate in each year (third column), average number of municipalities abandoned by the organization in each year (fourth column), average number of years in which the organization operates in a municipality (fifth column), competitive and exploratory indexes (sixth and seventh column). We present each of them and explain how we use them to cluster organizations in different homogeneous cartel types (k column).

While some organizations operate in many municipalities simultaneously, others have significantly smaller areas of operation. Zetas appear in an average of 42.2 municipalities every year, while Mana only in three. Second, the average number of years in which an organization operates consecutively in a municipality goes from 3.01 for Golfo, to 1.63 for Barbie. Third, trafficking organizations also tend to abandon markets with quite high variance. While organizations like Sinaloa abandon about 16.95 municipalities on average, the BL faction only abandons 2.15. Finally, we also identified two more dimensions: the propensity to explore new municipalities, and their preferences towards engaging in competitive behavior. We measured propensity to explore as the number of standardized municipalities in which a given drug trafficking organization was the first to ever operate, and preferences towards engaging in competitive behavior as the standardized number of municipalities in which a given trafficking organization was sharing a municipality with another one. While organizations like Zetas have a high tendency to engage in competitive behavior and in exploration of new territories, others like Tijuana and Juárez tend to avoid competition and exploration. The case of BL is particularly interesting, showing a tendency to explore new territories without engaging in competition.

We used all the different characteristics of drug trafficking organizations to classify them according to their modus operandi in four clusters. We create a matrix with the seven features from Table 1. We scale and normalize each feature to avoid that one or some of the features may dominate over the others. Then, we perform several runs of the k-means algorithm for varying ks. For each k we calculate the within groups sum of squares, to determine the appropriate number of clusters, that in our case is $k = 4$. We then return the four main classes of drug cartels according to our data.

[8]This last expansion coincides with increases in prosecution launched after the arrival of a new political administration in December of 2006 at the national level.



(a) 2000. (b) 2010.

Figure 8: The migration pattern for Juarez cartel.

The four classes are: 1 = "Traditional", 2 = "New", 3 = "Competitive" and 4 = "Expansionary Competitive". The results show quite differentiated strategies (Table 1) between each of them. A first cluster, integrated by Juarez, Tijuana, and Sinaloa is integrated by "Traditional" trafficking organizations that have operated in Mexico for the longest time. These organizations have a tendency towards being not competitive, typically being the first to operate in a particular territory. They operate in a large number of municipalities but also have a high rate of turnover.

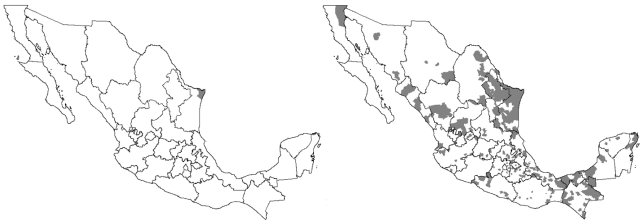
A second cluster, composed of la Mana, fractions of Sinaloa and BL, Barbie and a residual category, are the "New" organizations. On average, they emerged in 2007, more than ten years after the first cluster. They operate in a very reduced number of municipalities. They are not competitive and do not have a very developed tendency towards exploring new territories. This means their market strategy consists of operating in municipalities that had once been controlled by other criminal organizations but had been abandoned.

A third category is composed of BL and Familia, two criminal organizations that are relatively new (created on average in 2004), operate in many municipalities, and have strongly competitive tendencies. We called these organizations "Competitive" because they do not explore new territories but rather operate in places where another organization is already operating.

This tendency towards invading territories that are already taken is even stronger for the fourth cluster, integrated by Zetas and Golf organizations. We called these organizations "Expansionary competitive" because they are not only the most competitive but also the ones with the largest tendencies to explore new territories. In other words, they not only try to invade others' territories but also are the first to colonize new markets and to operate in areas where drug trafficking organizations had never been present before. In general, this last cluster is the one with the largest criminal organizations, operating on average on 324 municipalities (as of 2012) and spreading to an average of 38.87 new municipalities every year. Yet, it is also important to mention that their mobility is also the largest: they abandon an average of 22 municipalities per year, lasting only an average of 2.86 years in each one of them.

Trafficking Org	2010 Territories	Start Year	Territories	Abandoned	Years operated	Competitive	Exploratory	k
Juárez	74	1997	13.85	10.15	2.78	-0.67	-0.03	1
Tijuana	39	1997	10.1	8.15	2.74	-0.96	-0.21	1
Sinaloa	176	1993	25.6	16.95	2.84	0.18	0.64	1
Barbie	66	2006	5.75	2.45	1.63	-0.48	-0.72	2
Mana	32	2006	3.8	2.2	2.15	-0.82	-0.73	2
Sinaloa faction	53	2008	5.15	2.5	1.96	-0.67	-0.70	2
BL faction	57	2008	5	2.15	1.79	-0.52	-0.75	2
Other	24	2008	2.15	0.95	1.38	-0.99	-0.75	2
BL	157	2004	18.65	10.8	2.08	0.81	-0.36	3
Fam	227	2005	18.75	7.4	2.09	0.95	0.01	3
Golfo	244	1994	35.55	23.5	3.01	1.25	1.07	4
Zetas	405	2003	42.2	21.95	2.71	1.94	2.55	4

Table 1: The main features extracted for each drug trafficking organization.



(a) 2000. (b) 2010.

Figure 9: The migration pattern for Zetas cartel.

These differences match qualitative analysis done by ethnographers and specialist in Mexican criminal organizations. Journalistic accounts of Tijuana Cartel [4] and Juárez Cartel [8] had previously identified these organizations as localized, with few interests in moving towards other territories and being pro-actively competitive. Sinaloa cartel had been portrayed as a cartel with larger incentives to invade others although still mostly confined to the north of the country [20]. Both Golfo [21] and Zetas [9] cartels had recurrently been described as the most invasive and aggressive criminal organizations in terms of market expansion. Our findings also match qualitative evidence collected by experts on La Familia, BL and Barbie organizations.

Finally, patterns of migration also differ. While some tend to expand in their neighboring territories, others expand more broadly around the country (Figures 8 and 9). Notoriously different patterns can be seen in Zetas (Figure 9), a drug trafficking organization that tends to migrate more randomly, extending over the whole territory and a group of more localized criminal organizations like Juárez Cartel (Figure 8).

Overall, our exercise extracted significant knowledge about Mexico's drug trafficking industry, and about the inner behaviours of different criminal organizations. The collected data further advances our knowledge of criminal organizations and provides quantitative evidence of criminal behaviour that was previously only qualitatively described.

7. CONCLUSION

In this paper we presented a simple framework, called MOGO, to generate low cost intelligence information. MOGO uses the vast amount of knowledge present on the Web to obtain quantitative information about a phenomenon that would otherwise require the operation of large scale, expensive intelligence exercises. Based on a simple three step process (list definition, query generation, and crawling), MOGO is able to create a knowledge by exploiting indexed reliable sources such as online newspapers and blogs.

As our first approach, we use this mechanism to understand Mexican drug trafficking organizations and identifying their market strategies, their preferred areas of operation, and the way in which these have evolved over the last two decades. Information on these aspects had never been collected before. Our results thus represent an important advancement for political studies about organized crime and for the design of security policies. We showed that criminal organizations, rather than being similar and operating under identical mechanics, differ significantly in their strategies and market orientations. We identified four types of Mexican criminal organizations: traditional, new, competitive and expansionary competitive. Traditional organizations operate in municipalities that they have controled for a long time, on average since 1995. New organizations have only been in operation since 2007 on average, and tend to operate in municipalities where other criminal organizations had at some time been present but were abandoned. Competitive organizations are those that operate in territories are controlled by other organizations. Finally, expansionary competitive are those not only operate in territories that were already taken but also explore new territories, expanding their operations to areas in which drug trafficking organizations had never operated before. Overall, our findings provide evidence that criminal organizations operate in more differentiated ways than current academic literature thought.

To test how accurate MOGO is extracting knowledge we used it to identify the areas of operation of known individuals, particularly governors of Mexico. In the validation section we showed that MOGO perfectly identifies the areas of operation of governors assigning each of them to the state that they rule. This paper opens the path for much future work. Most immediately, the knowledge extracted by MOGO will be used by to identify patterns of criminal violence within Mexico by linking different types of drug trafficking organization with degrees of violence. Yet, in the near future we will apply MOGO to extract information about different problems. For example, identifying the areas of operation of different political groups, of particular individuals, or public figures, and insurgency groups. In terms of com-

puter science future developments, the most important one lies in the improvement of MOGO's framework. By improving the query list generation rules and the data validation phase, and in parallel eliminating the usage of an oracle by directly crawling our set of reliable newspapers, we will make MOGO a framework able to provide better and more accurate results. We also plan to use the article's textual data for semantic analysis [5].

Acknowledgements. Michele Coscia is a recipient of the Google Europe Fellowship in Social Computing, and this research is supported in part by this Google Fellowship. Viridiana Rios was supported by Mexico's Office of the Executive, the Center for US-Mexico Studies at the University of California in San Diego, and the Program in Inequality and Criminal Justice at Harvard Kennedy School. Special thanks to Cesar Hidalgo, Gary King, Peter Bol and Ricardo Hausmann for valuable feedback. Thanks to Brad Holland for editing.

8. REFERENCES

[1] Mohammad˜Ali Abbasi, Shamanth Kumar, Jose Augusto˜Andrade Filho, and Huan Liu. Lessons learned in using social media for disaster relief - asu crisis response game. In *SBP*, pages 282–289, 2012.

[2] Judit Bar-Ilan. The use of web search engines in information science research. *Annual Review of Information Science and Technology*, 38:231–88, 2004.

[3] Michele Berlingerio, Michele Coscia, Fosca Giannotti, Anna Monreale, and Dino Pedreschi. Foundations of multidimensional network analysis. In *ASONAM*, pages 485–489, 2011.

[4] Jesus Blancornelas. El cartel: los arellano félix. In *Plaza y Janés*, 2002.

[5] Markus Bundschus, Anna Bauer-Mehren, Volker Tresp, Laura˜Inés Furlong, and Hans-Peter Kriegel. Digging for knowledge with information extraction: a case study on human gene-disease associations. In *CIKM*, pages 1845–1848, 2010.

[6] Stefan Büttcher, Charles L.˜A. Clarke, and Gordon˜V. Cormack. *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press, 2010.

[7] A.˜M. Cohen, W.˜R. Hersh, C.˜Dubay, and K.˜Spackman. Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. *BMC bioinformatics [electronic resource].*, 6(1), April 2005.

[8] Francisco Cruz. El cartel de juarez. In *Editorial Planeta*, 2009.

[9] Jorge Fernandez˜Menendez and Victor Ronquillo. De los maras a los zetas: los secretos del narcotráfico, de colombia a chicago. In *Ediorial Grijalbo*, 2006.

[10] Olivier Ferret. Discovering word senses from a network of lexical cooccurrences. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[11] Justin Grimmer. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35, 2010.

[12] Justin Grimmer and Brandon˜M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Open Scholar*, 2012, to appear.

[13] Eduardo Guerrero. Security, drugs and violence in mexico: A survey. In *7th Norh American Forum*, 2011.

[14] R.˜Hausmann, C.˜Hidalgo, S.˜Bustos, M.˜Coscia, S.˜Chung, J.˜Jimenez, A.˜Simoes, and M.˜Yildirim. *The Atlas of Economic Complexity*. Puritan Press, 2011.

[15] Gary King and Will Lowe. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, 57:617–642, 2003.

[16] Sang˜Hoon Lee, Pan-Jun Kim, Yong-Yeol Ahn, and Hawoong Jeong. Googling social interactions: Web search engine based social network construction. *PLoS ONE*, 5(7):e11233, 07 2010.

[17] Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. In *Mining Text Data*, pages 415–463. 2012.

[18] M.˜E.˜J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351, December 2005.

[19] Brendan O'Connor, Ramnath Balasubramanyan, Bryan˜R. Routledge, and Noah˜A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2010.

[20] Diego˜Enrique Osorno. El cartel de sinaloa: Un historia del uso politico del narco. In *Random House Mondadori*, 2009.

[21] Ricardo Ravelo. Osiel. vida y tragedia de un capo. In *Ediorial Grijalbo*, 2009.

[22] John˜A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, April 2001.

[23] R.˜Rodríguez Castañeda. El méxico narco. In *Editorial Planeta*, 2010.

[24] Markus Schedl, Tim Pohle, Peter Knees, and Gerhard Widmer. Exploring the music similarity space on the web. *ACM Trans. Inf. Syst.*, 29(3):14:1–14:24, July 2011.

[25] Maximillian Schich and Michele Coscia. Exploring co-occurrence on a meso and global level using network analysis and rule mining. *Minign and Learning with Graphs, KDD Workshop*, 2011.

[26] M.˜V. Simkin and V.˜P. Roychowdhury. Theory of Aces: Fame by chance or merit? *eprint arXiv:cond-mat/0310049*, October 2003.

[27] Arthur Spirling. Us treaty-making with american indians. *American Journal of Political Science*, 56(1):84–97, 2012.