Study of Social Networks : modeling and Analysis

Yonathan Portilla

THÈSE

Présentée pour obtenir le grade de Docteur en Sciences de l'Avignon Université France

SPÉCIALITÉ : Informatique

École Doctorale 536: *Agrosciences et Sciences*

Laboratoire d'Informatique d'Avignon

Etude des Réseaux Sociaux - Modélisation et Analyse

par

Yonathan Portilla

Soutenue publiquement le 20 mai 2019 devant un jury composé de :

Jocelyne Elias	Professeur à l'Universié Paris Descartes	Rapporteur
Jean-Marc Francony	Directeur de recherche à l'Universié de Grenoble	Rapporteur
Patrice Bellot	Professeur à l'Universié Aix-Marseille - Polytech'Marseille	Examinateur
Tijani Chahed	Professeur à Telecom Sud-Paris	Examinateur
Eitan Altman	Directeur de recherche à l'INRIA Sophia-Antipolis	Directeur de thèse
Rachid El-Azouzi	Maitre de conférence à l'Université d'Avignon	Co-directeur

Laboratoire LIA, Avignon

INRIA, Sophia Antipolis

THESIS

A thesis submitted in partial fulfillment for the degree of Doctor of Sciences in Avignon University France

In Computer Science

Doctoral School 536: *Agrosciences et Sciences*

Laboratoire d'Informatique d'Avignon

Study of Social Networks - Modeling and Analysis

by

Yonathan Portilla

Committee

Jocelyne Elias	Associate Professor at Paris Descartes University	Reviewer
Jean-Marc Francony	Director of Research at Grenoble University	Reviewer
Patrice Bellot	Associate professor at Aix-Marseille University - Polytech'Marseille	Examiner
Tijani Chahed	Professor at Telecom Sud-Paris	Examiner
Eitan Altman	Director of research at l'INRIA Sophia-Antipolis	Supervisor
Rachid El-Azouzi	Senior Lecturer at Avignon University	Co-supervisor

Laboratoire LIA, Avignon

INRIA, Sophia Antipolis

Dedicated to my Family...

Résumé

Actuellement les réseaux sociaux, se focalisent sur le partage et échange des opinions, vidéos, photos, musique, actualités et autres informations, un de ses objectifs c'est d'établir des liens directes et indirectes avec les utilisateurs. Les réseaux sociaux permettent aussi de promouvoir des produits, des personnes (leur image politique ou artistique) ou des marques influentes.

Les réseaux sociaux changent rapidement, pour cette raison on cherche a voir l'évolution de ces outils de partage, et aussi voir comment les réseaux sociaux changent avec le temps.

On a l'opportunité d'étudier les événements qui se produisent dans les réseaux sociaux grâce à la quantité des données qui se produisent. Dans le marché actuel il y a des outils qui permettent l'analyse des réseaux sociaux, mais la plus part est payante, et les outils 100% gratuits disparaissent avec le temps. Pour cette raison nous avons décidé de produire des outils informatiques capables d'extraire et analyser les données des réseaux sociaux étudiés.

Cet étude commence avec l'état de l'art, ou on décrit le contexte du problème abordé, les travaux qui sont à l'origine de cette étude et un résumé des contributions faites au cours de la thèse que nous présentons brièvement dans le reste du résumé.

i. D'abord nous nous focalisons dans l'empreinte géo-linguistique et l'évolution du langage en Twitter. L'accès au contenu des messages envoyés par un groupe des abonnées d'un réseau social peut être utilisé pour identifier et quantifier certaines spécificités d'un groupe. La spécificité peut représenter le niveau d'intérêt pour un événement ou un produit, ou la popularité d'une idée, un hit musicale ou une figure politique. La spécificité peut aussi représenter la façon comment le langage est utilisé et transformé, la façon comment les mots sont écrits et la manière comme apparaissent des nouvelles règles de grammaire.

ii. Ensuite nous étudions l'évolution du phénomène culturel appelé mème dans les réseaux sociaux. Les mèmes ont été définies par R. Dowkins comme un phénomène culturel qui

se propage a travers de formes non génétiques. Nous examinons trois des plus populaires mèmes de l'internet et nous examinons leur impact dans la société dans les Pays Méditerranéens. Nous utilisons pour les analyses Google Trends, Topsy (un outil pour mesurer la popularité des mots sur Twitter) et YouTube pour quantifier l'impact des mèmes dans la société du Méditerranée.

iii. Après cela nous étudions le graphe de recommandations de YouTube basées sur les mesures et les outils stochastiques. Nous confirmons que les listes de recommandations influencent les vues d'une vidéo. Nous nous focalisons sur le système de recommandations qui boostent la popularité des vidéos. Nous construisons en premier un graphe qui capture le système de recommandations dans YouTube et nous étudions la relation entre le nombre de vues d'une vidéo et la moyenne du nombre de vues d'une vidéo dans sa liste de recommandation.

iv. Pour conclure nous décrivons les outils disponibles en ligne et les outils que nous avons développés pendant l'écriture de la thèse. Les outils en ligne Topsy, Trendistic et Google Trends nous ont permit d'analyser des plateformes comme YouTube et Twitter. On a produit aussi des outils basés sur les API's : dans Twitter nous avons utilisé la fonction Streaming pour télécharger et analyser les tweets , avec l'API de Topsy nous avons étudié l'évolution de la langue et l'utilisation des mots , et les API's de YouTube nous ont permis de décrire la façon dont se comportent les listes de recommandations et la popularité des vidéos.

Abstract

Currently social networks focus on the sharing and exchange of opinions, videos, photos, music, news and others informations, one of its objectives is to establish direct and indirect links with users. Social networks also promote products, people (their political or artistic image) or influential brands.

Social networks are changing rapidly, so we're looking to see the evolution of these sharing tools, and see how social networks change over time.

We have the opportunity to study the events that occur in social networks thanks to the amount of data they produce. In the current market there are tools that allow the analysis of social networks, but most tools are not free, and 100% free tools disappear over time. For this reason we decided to produce computer tools able to extract and analyse the data of the social networks studied.

This study begins with the state of the art, where we describe the context of the problem, the work that led to this study and a summary of the contributions made during the thesis that we present briefly in the rest of the abstract.

i. First we focus on the geo-linguistic fingerprint and language evolution in Twitter. Access to content of messages sent by a group of subscribers of a social network may be used to identify and quantify some features of a group. The feature can represent the level of interest in an event or product, or the popularity of an idea, or of a musical hit, or of a political figure. The feature can also represent how language is used and transformed, how words are written and how new grammatical rules appear.

ii. Then we study the evolution of the cultural phenomenon called meme in social networks. Memes were defined by R. Dowkins as a cultural phenomenon that spreads through non-genetic forms. We examine three of the most popular memes of the internet and examine their impact on society in the Mediterranean countries. We use for analysing Google

Trends, Topsy (a tool to measure the popularity of words on Twitter) and YouTube to quantify the impact of memes in the Mediterranean society.

iii. After that we study the YouTube recommendation graph based on measurements and stochastic tools. We confirm that recommendation lists influence the views of a video. We focus on the recommendation system that boosts the popularity of videos. We build first a graph that captures the recommendation system in YouTube and we study the relationship between the number of views of a video and the average number of views of a video in its recommendation list.

iv. To conclude we describe the online tools available and the tools that we developed during the thesis. The online tools Topsy, Trendistic and Google Trends allowed us to analyse platforms like YouTube and Twitter. We also produced tools based on API's: in Twitter we used the Streaming function to download and analyse tweets, with the Topsy API we studied the evolution of the language and the use of words, and the YouTube's APIs allowed us to describe the behaviour on the lists of recommendations and the popularity of videos.

Acknowledgements

First of all, I would like to thank all members of the PhD committee for agreed to evaluate my work.

I warmly thank my supervisor Prof. Eitan Altman, for the constant support, help and for trust in my capabilities.

I warmly thank my co-supervisor Rachid El-Azuzi, for his support.

I would also like to thank all who contributed to this research, *La Maison de la Recherche*, University of Avignon, Inria's, CERI, the Agorantic S.F.R. (Structure Fédérative de Recherche on digital societies and culture) cofounded by UAPV, INRIA and the CNRS.

I also need to thank all the people I met, who made me be better in my professional and personal development.

My family for the support, understanding and love.

Thanks all, for trust and support my work.

Contents

Chapter 1

Introduction and state of the art

1.1 Context and motivation

Online social networks are currently a source of sharing opinions, videos, photos, music, news and others informations. Different communities with distinct areas of interest may use the same online social network so as to share issues of common interest with users with similar areas of interest. In order to share contents of interest, the online social network promotes the establishment of direct or indirect links with users. The social network allows to market products and ideas, and to acquire influence. The latter can be in the form of establishing a brand image or sometimes also a political image.

Online social networks are changing quickly, for this we are interested in seeing the evolution of these sharing tools, and also see how social networks change over time.

We have the opportunity to study the events that occur within social networks thanks to the amount of data they produce, and to the computer tools that are available to extract the information that allows us to analyze, study and to establish after several scientific experiments, changes in the characteristics of the social networks studied. All this to try to explain the behavior of people, optimize the process of displaying the information produced, or just know the preferences of users, with the following aims 1) of advertising products that meet the expectations of users. This is the case of Facebook and YouTube; 2) selling data that allows to the opinion mining. This is the case of Twitter.

There are social networks that are specialised gather users according to users' interest, like Instagram that does with photos and YouTube that focuses on sharing videos, we can also find social networks that try to evolve according to the preferences of users and that add features

from users like Facebook and others like Twitter that focuses on micro-critics, finally we have a multitude of options depending on the choice of the user. But in social networks we can find different types of users and classify them according to their interests and / or as they intervene in the evolutionary dynamics of social networks.

1.1.1 Social Media

Novel social media are capable of deeply influencing communication and human interaction in ways that have impact on the whole society. Such paradigms have drastically changed our interaction patterns in a disruptive way compared to the pre-social networks era. They caused surprising effects including 1) the promotion of dialogue in society which could facilitate democratization and social change; 2) the creation of novel communication paradigms through platform-specific interaction patterns, and 3) wider access to culture beyond geographic and national borders. The sheer scale and rapid dynamics of communicative practices in the digital domain poses major challenges for studying and monitoring the formation and evolution of communities in the digital domain.

Social media's main enabler is the possibility to share information, ideas and create new content. Individuals can join online groups able to influence economics, politics and opinion dynamics at the local, national and transnational level. For example, social media platforms, most notably Twitter and Facebook, played a key role in mobilizing and coordinating protests during the so-called Arab Spring (Breuer and Farquhar, 2012; Salmon, 2016). In European countries it is now customary for citizens to engage in several forms of social media during political campaigns through platforms such as Facebook and Twitter. Social media is also able to influence opinion formation, social mobilization and decision makers, thus complementing traditional mechanisms such as national media and public consultations. Music and liberal arts have also benefited from social media; more and more artists and art enthusiasts are now participating in the creation and diffusion of art through online communities.

1.1.2 Evolution of languages and Culture

Unlike many other social networks whose business model is mainly based on offering advertisements, Twitter makes money by selling content: the content of a large portion of transmitted messages is sold to interested companies. A small portion of around 1% of all tweets is made available for free. The fact that such a huge amount of messages is made available makes Twitter attractive as a tool for opinion mining (learning about opinions) in a large population. Twitter can serve as an alternative to opinion polls for market analysis not only in the context of selling goods but also for opinion trends analysis such as election campaigns (O'Connor et al., 2010;

Tumasjan et al., 2010). Twitter allows to access some information for free through different APIs (Application Programming Interface).

We propose a simple methodology that can give some information on the geographical area in which a given word is used. It is based on the fact that different geographical areas may have different activity pattern of tweets during the time of the day. More precisely, we make use of the fact that the amount of messages generated by subscribers at a given location changes during the time of the day in a periodic way which may differ from one region to another. We show that words and spelling evolve in a way that allows to display the geographic or social identity of the person writing the message. of political parties and public figures.

In previous work (Nowak M. A., 2002; Christina Pawlowitsch and Ritt, 2011) the evolution of language was assumed to be due to preference for simplicity and to preference to more informative words.

We apply this method to track how words evolve in different ways within different geographical areas that have the same language in common.

1.1.3 Data Mining with Twitter

Twitter has become a platform for socially media diffusion, we can diffuse an image a video or an opinion in a short text (140 characters in the beginning and 280 in september 2017 (Monde.fr and AFP, 2017)) . The business model of Twitter is based on selling content: Transmitted messages are sold to interested companies. This may be used for detecting opinions and their evolution. Twitter serves as an alternative to opinion polls for market analysis not only in the context of selling goods but also for opinion trends analysis such as election campaigns. A small portion of around 1% of all tweets is made publicly available for free.

We developed a tool that uses an API of Twitter to take the stream of 1% of all tweets, these are divided in two parts, the first part includes geolocalized streams which consist of around 10% of tweets, and the others part consist of tweets without the localization.

1.1.4 Data Mining with YouTube

YouTube has become a key international platform for socially enabled media diffusion. This platform allows not only to share videos, but also to create interaction between users (friends, creators). It has become the most attractive and popular media diffusion with a huge quantity of user generated content and none of its competitors have achieved the same success. Based on statistics available from the website Alexa.com, more than 30% of global internet users visit

youtube.com per day. Other statistics from http://www.youtube.com/t/press establishes: “over 800 million unique users visit YouTube each month” and “72 hours of video are uploaded to YouTube every minute”. All YouTube videos are available to the general public and include valuable data about the video. In order to help users to find interesting videos from this huge number of videos, YouTube provides several features such as a search engine, front-page highlighting, and recommendation system. In this section, we present a measurement study performed on the dataset of videos crawled from the YouTube website. We summarize our study on YouTube as follows: We have developed two datasets related to YouTube as a part of the european project CONGAS.

Public Dataset: This dataset contains some data extracted using a tool that we developed for that purpose which uses data made available publicly to any user. This data consists of the time evolution of popularity (number of views) of videos. The information of videos which includes the basic information about the video, such as title, upload time, and total view count, and related video list can be retrieved through YouTube Data APIs We use for this section the YouTube Data API.

For retrieving the view statistics data can only be retrieved through HTML because it is not supported by the API. for this part we develop a tool able to download the statistic for a video in JSON format including views, watch-time, shares and subscribers if the video has the option stats enabled.

Recommendation Dataset: We have developed this second dataset which concerns data on properties of the recommendation graph of YouTube. This graph is defined as follows: Nodes corresponds to video clips. A directed link between two videos, say from A to B, means that B is on the recommendation list of video A. The data presented here is obtained by analysis of data which is publicly available through YouTube. Each node has a weight which represents the total number of views that the video has received since it was uploaded to YouTube. The list of recommendations of a video appears at the right of YouTube site. At the left hand side its title, numbers of views.

Image print screen

1.1.5 Communities

Community structure of social networks heavily relies on the algorithms used. For example, recommender systems can narrow users’ options to a limited community of items or information. develops tools to investigate the effect of recommender algorithms on the evolution of communities, and characterizes influence of recommender systems on the diversity of opinions

measured across the groups, and the role of users bridging different communities. One can classify users of socials networks according to their geographic and social background. We study the correlation of this classification with the language of corresponding tweets.

1.1.6 Globalized Culture through social media

We target cultural phenomena arising from the creation of a massive number of variants of a given video, e.g., the "Harlem Shake" video was reproduced by millions of people (Altman and Portilla, 2015). The biologist Richard Dawkins called such phenomena "meme". The way memes propagate suggests epidemic models can be used to describe this process, using a methodology similar to those previously proposed for the evolution of languages. Because the creation of art is often a process involving collaboration among several contributors, the novelty in recent years is the (Richier et al., 2014) influence that entire online communities of authors and followers may have on the generation of art.

1.2 Contributions of the thesis

In this section we describe the main contributions of this thesis. The main objectives aimed by this thesis are

- Study geodiversity of words from a common language
- Devellopments of sofware tools for that
- Geolocalize the source of new words in Twitter
- Study evolution of languages in Twitter
- Study the evolution of cultural phenomena (memes) through social networks.
- Markov modeling of recommendation system of YouTube

The manuscript is organised as follows :

Chapter 2 focuses on geo-linguistic fingerprint and the evolution of languages in Twitter. Access to content of messages sent by some given group of subscribers of a social network may be used to identify (and quantify) some features of that group. The feature can stand for the level of interest in some event or product, or for the popularity of some idea, or of a musical hit or of a political figure. The feature can also stand for the way the written language is used and transformed, the way words are spelled and the way new grammatical rules appear.

In this work we have two targets. The first is to identify the characteristics of subscribers by geographical location and by language. We develop a methodology that allows one to perform such a study using freely available statistical tools which makes use of a part of all tweets which Twitter makes available for free over the Internet. The methodology is based on the fact that the geographical area can be predicted according to the pattern of activity of the tweets during certain hours of the day. The second objective is to present our findings. We analyze the different ways of writing a word, and the expressions between communities that can be located even if they have a common language. We also show you how to use the periodogram of tweets to analyze the popularity of political or public figures.

In chapter 3 we study the evolution of cultural phenomena (memes) through social networks Memes were defined by R. Dowkins as cultural phenomena that propagate through non-genetic forms. We examine three of the most popular Memes on the Internet and examine its impact on society in the Mediterranean countries. We use google trends as well as Topsy for analyzing tweets in order to quantify the impact of the Memes on the Mediterranean societies.

We obtain quite different results with the different tools we use, which we attempt to explain based on some propagation characteristic of each one of the Memes. Our analysis shows the extent at which these Memes cross borders and thus contribute to the creation of a globalized culture. We end the chapter by identifying some of the impacts of the globalization of culture.

In chapter 4 we study YouTube recommendation graph based on measurements and stochastic tools. Indeed, the list of YouTube recommendations influences the views of a video. We focus on the recommendation system in boosting the popularity of videos. We first construct a graph that captures the recommendation system in YouTube and study empirically the relationship between the number of views of a video and the average number of views of the videos in its recommendation list. Next, we consider a random walker in the list of recommendations, this is a random user who navigates through videos in such a way that the video that chooses to see is randomly selected between the videos in the list of recommendations of the previous video what did you see. We study the stability properties of this random process and show that the trajectory obtained does not contain cycles if the number of videos in the recommendations list is small (which happens if the computer screen is small).

In chapter 5 we describe the tools that I devoleped during my thesis. These tools allow to analyse several platforms of contents such as Twitter, YouTube, Goole Trend and Topsy. The development of these tools takes a huge time since all these platforms change constantly their technology. In this chapter we also explain how these tolls allow us to study the social networks and its evolution. Furthermore, analyzing and discovering the operation of some platforms allow to identify some of the factors associated with the evolution of popularity and user behavior. This has allowed us to describe and explain the results and tests carried out throughout our research.

In chapter 6, we summarize the results of the previous chapters about the social networks analysis on YouTube and Twitter. Then, we discuss about the future works of the thesis. These future works include improve the analysis of opinion in micro-regions on Twitter and use metadata for analysis.

Chapter 2

Geo-linguistic fingerprint and the evolution of languages in Twitter

2.1 Introduction

The evolution of social and geographical communities can be traced through the analysis of language on social media. Different geographical areas provide differentiation of the same language, other reasons for differentiation may be due to social class division. The language used by an individual (e.g., their use of certain words or phrases in a post) will reveal something about their experiences, worldview and identity. At the same time, in the era of globalization online communities influence language evolution (C. Danescu-Niculescu-Mizil, 2013) because communities may adopt different languages, depending on how new words are accepted or not in the globalization process either by national or cross-lingual communities.

Using very large number of tweets we are able to determine the geographic source of words using a periodogram approach that we develops. This approach makes uses of the fact that most people of the same region had similar activity periods.

Related work. The idea of using activity periodogram of communications in order to characterize user behavior has already been studied in the context of mobile communications at (Bulut and Szymanski, 2015). Yet in that context it is not possible to track the use of specific words. In contrast, using the periodogram in the context of Twitter adds the possibility for localisation of the source of words which are frequently used. On the description of and reasons for changes and evolution of spelliings of words in social media such as SMS, chats and Twitter, the reader is referred to (Stephan Gouws and Hovy, 2011; mei Sun, 2010; Fernández and Seemann, 2009).

Reference (Danescu-Niculescu-Mizil et al., 2013) studies the user reactions to linguistic changes in online communities. A short conference version of this chapter appeared at (Altman and Portilla, 2012). This chapter expands on all existing material in (Altman and Portilla, 2012). It futher adds new concepts (e.g. the quality of geolocalisation in Section 2.3), it adds completely new applications to political science (sections 2.11-2.12) and examines refinements of the geolocalisation approach (Section 2.13). In (Morchid et al., 2015) we have studied an alternative approach for geo-localisation based on similarity measures of tweets that are not geo-localized with other tweets in which prior localisation information is available.

Structure. The chapter begins with a section that introduces the periodogram of daily Twitter activity. Sections 2.3-2.9 then focus on the application of Twitter periodogram to the evolution of words, of expressions and of spelling of words. Section 2.10 briefly compares findings on the language over Twitter to those obtained through otherh media, and Sections 2.11-2.12 apply the periodogram approach for geo-localisation to politicians, to public figures and to parties. A refinement of our approach is proposed in Section 2.13. We end with a concluding section. A preliminary shorter version of this chapter appeared in (Altman and Portilla, 2012).

2.2 Periodograms of daily activity: a geo-linguistic fingerprint

We begin by introducing our key approach for geolocalisation of tweets. Figure 2.1 displays the frequency of appearance of the words "to, the, el, y, a, i" over a period of a month. The frequency each of these words has a periodic behavior where the period corresponds to one day. We also observe that the words "To", "I" and "The" have a very similar wave form, and so do the words "El" and "Y". The word "A" has a distinct wave form differenet from the other two. The first group contains words that are frequently used in English, where as the second group corresponds to words that appear frequently in Spanish. The word "A" appears frequently in many languages (e.g. English, Spanish, French). The word *Y* appears also in French but its frequency there is much smaller. We conclude that words that are typical to one specific language have a common pattern, which we call a "fingerprint" of the language.

Figure 2.2 presents a typical german finger print. The same daily period is seen to be common to three different words that are very common in German.

Next we give an example of several spanish words, see Figure 2.3. There are two exceptions. We included an English word, "the", which is the most popular word in the figure. It appeared in around 9% of all tweets. It indeed has a completely different periodic pattern. The second exception is the word "La" which frequently appears not only in Spanish but also in French and Italian. Nevnerthless, unlike "and", we observe much reseemblence to the pattern of Spanish

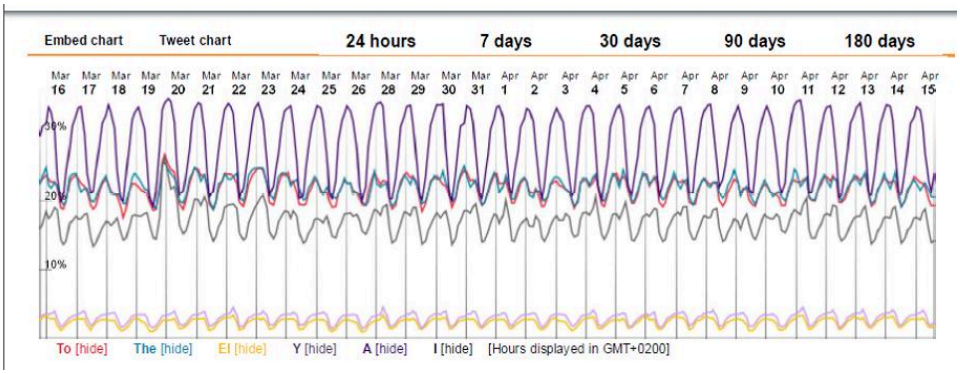


Figure 2.1: *The frequency of appearance of the words "to, the, el, y, a, i"*

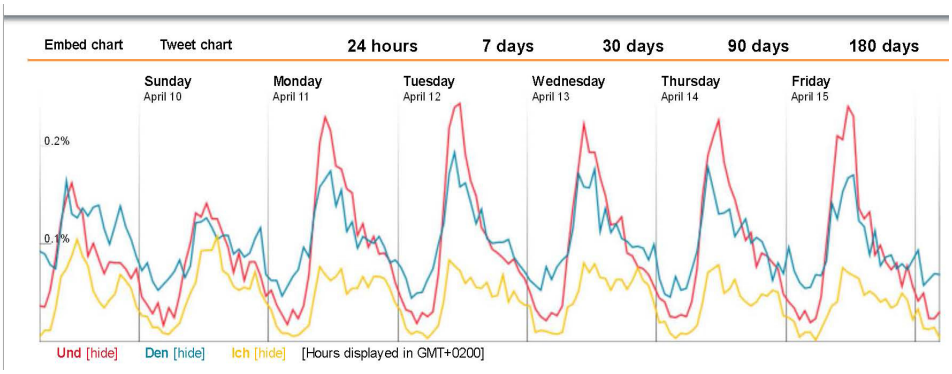


Figure 2.2: *The frequency of appearance of several german words during 7 days. A very clear common periodic daily pattern appears.*

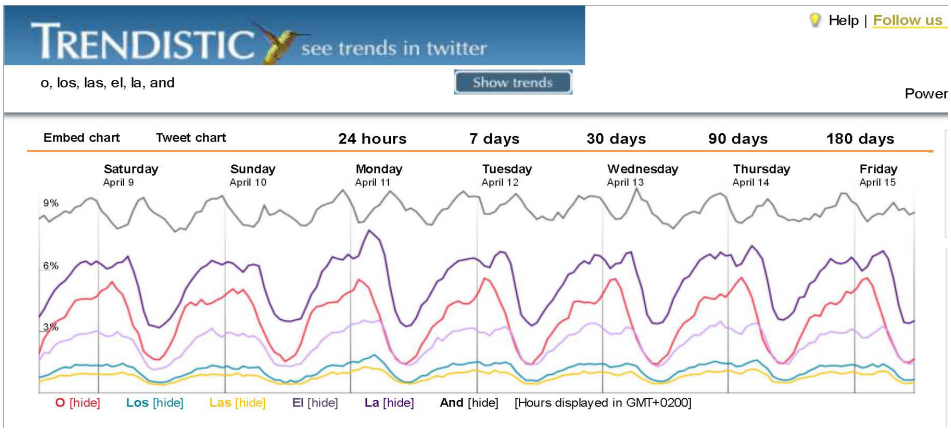


Figure 2.3: *The frequency of appearance of several spanish words during 7 days. A very clear common periodic daily pattern appears and is compared to non-spanish words*

words. A possible explanation could be that there are significantly more tweets in Spanish than in French and Italian. Therefore the periodogram of "La" is closer to the spanish even if spanish and french words had quite different periods.

The reason that each language has its own fingerprint could be

- The fact that each language has its own geographic distribution, and thus a different time-zone distribution.
- The habits related to working hours, eating hours etc may differ from one community to another, and these habits may imply different distribution of tweeting times.

Can we check which of the above is more pertinent? Observe in Figure 2.4 the frequency of appearance of the words "une", "della", "der". These three words correspond to articles in French, Italian and German. We see that the periodic frequency pattern of the three words is very similar. These three languages are mainly spoken in Europe, and the time zones in which they are spoken are the same. It thus seems that different words in different languages would give a similar fingerprints in the trend graph if they are mainly used in geographical areas that have the same time zone (or neighboring time zones).

We next compare the Spanish word "Todas" with the French word "et". The Spanish one is seen to be shifted with respect to the French word by around 6 hours. For example its lowest activity during the day appears around 6 hours later than that of the French word. This suggests that most tweets in spanish originate in Latin America which has a time difference of 6 hours or more with respect to France.

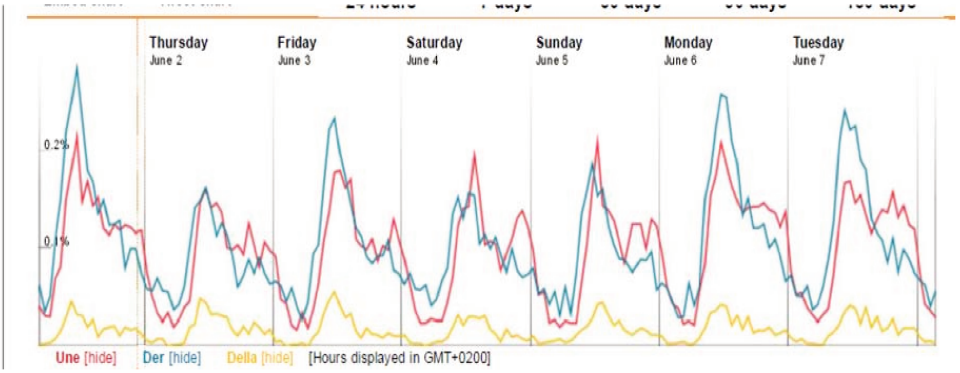


Figure 2.4: *The frequency of appearance of the words "une", "della", "der"*

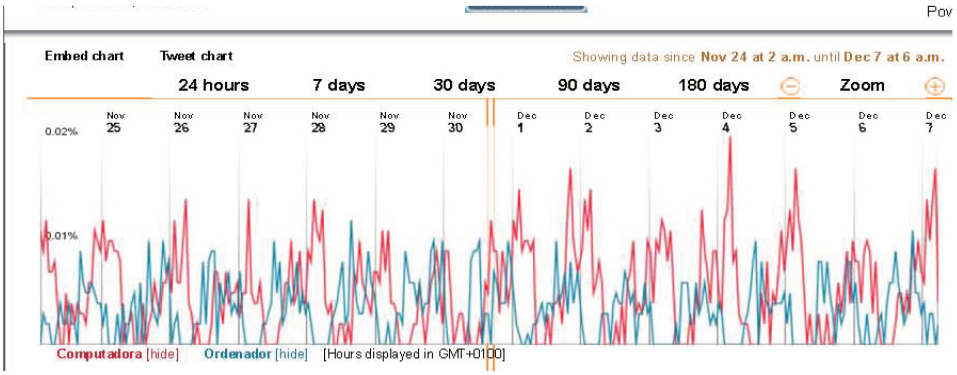


Figure 2.5: *The frequency of appearance of a word in Spanish from Latin America and that from Spain*

2.3 Differences in a language spoken in various geographical areas

This Section explains how to use Twitter periodogram to locate the tweets' source.[1] Daily periodograms can be made more selective so as to restrict to a subregion in which a language is spoken. As an example, we compare tweets with the Spanish words "computadora" and "ordenador". Both mean "computer", but the first is used in Latin America the second in Spain. The corresponding periodograms appear in Figure 2.5. We see that the term used in Spain has its minimal appearence around 8 hours before the Latin American one.

We note that the average daily number of tweets in which "ordenador" appears is around 2/3 the one corresponding to "computadoras". Does this suggest that the fraction of spanish tweets originatinng from Spain is close to that originating from Latin America? To answer this question, we may wish to compare also other words, or in contrast, to see how the relative frequencies behave in other contexts. When comparing the number of appearence of these words over the

[1] A preliminary shorter version of the beginning of this Section appeared in (Altman and Portilla, 2012). In this chapter we further introduce for the first time a quality measure for geo-localisation of tweets.

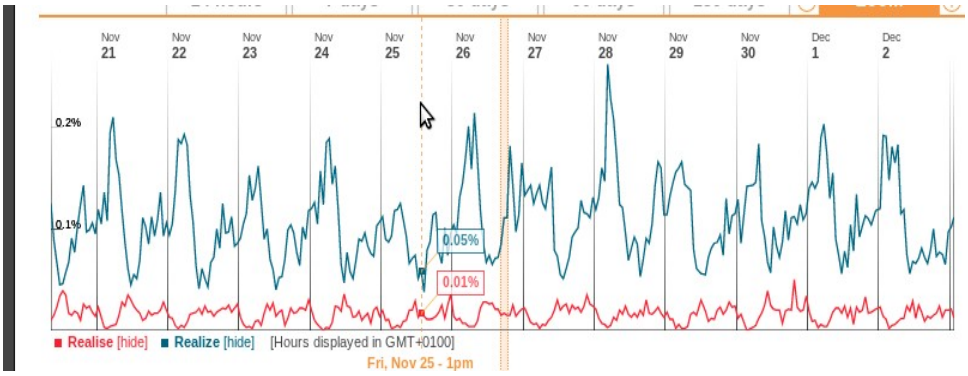


Figure 2.6: *The frequency of appearance of the words "realise" and "realize"*

whole Internet, by using fightgoogle, we obtained (on Dec. 7, 2011) the figures: 4,580,000 for "computadora", and 7,150,000 for "ordenador".

Next we shall differentiate between the periodograms of the American versus the British versions of English. We do so by comparing the fraction of tweets containing "realize" (American version) and "realise" (British version) as a function of time, as is seen in Figure 2.6.

We see clearly that the minimum daily tweeting activity of the American word occurs around 6 hours later than the Britih one.

Our approach for geolocalisation is based on the fact that people's behavior as a function of time is quite similar across different geographical zones. In particular, we observed low activity periods late at tight, i.e. at 24h00 - 6h00. By identifying the time shift between our own period of low tweeting activity and that of a set of tweets that contain specific keywords, we can identify the difference in the time zone and hence the difference in the longitude.

The keywords have to be chosend in a way that uniquely define theh longitude. Indeed, if the same keyword is tweeted from various locations in different time zones then there may not be clear low activity periods. This is the case with the words in Figure 2.1. We do not observe there inactivity periods but rather oscillations during the tweeting activity.

This motivates us to introduce the following criterion for the **quality of geo-localisation** of tweets. We shall consider the ratio of tweet rates (RTR) between the peak activity period and the lowest activity period as a quality measure for the geo-localisation. The highest the ratio is, the better is the geo-localisation quality.

Note: the tweeting hours in all the figures in this chapter are measured with respect to our own location which in terms of time zones is given by the UTC (Central European Time Zone). It is one hour later than GMT (Greenwich Meaen Time).

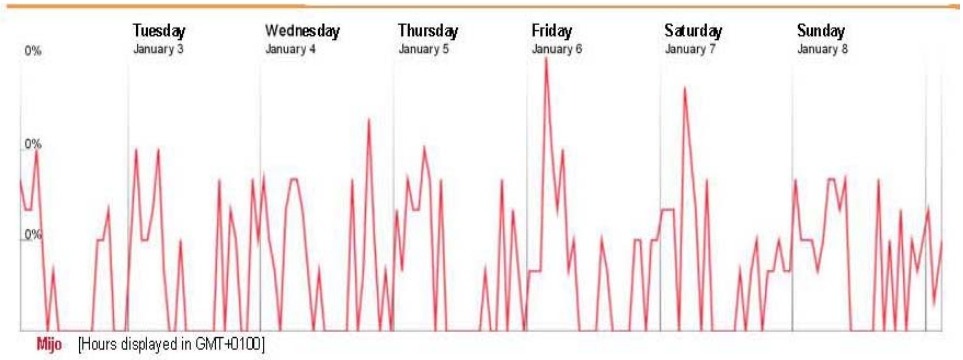


Figure 2.7: *The frequency of appearance of "mijo"*

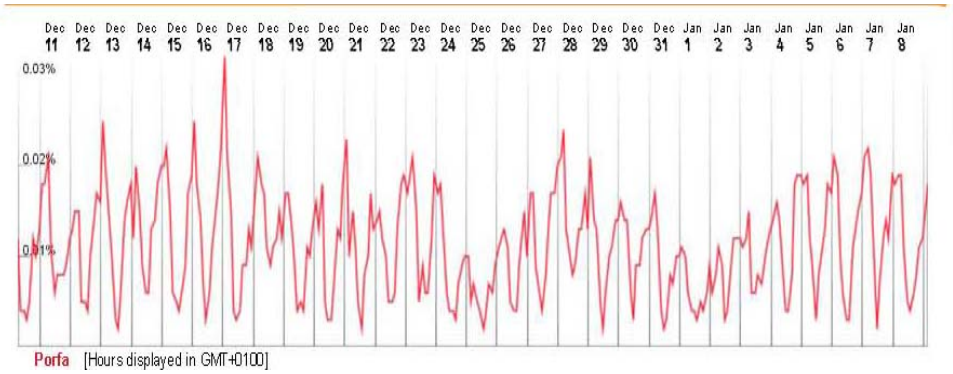


Figure 2.8: *The frequency of appearance of "porfa"*

2.4 Twinglish and other languages

We focus in this section on some spellings or forms of writing words that are typical for social media (Stephan Gouws and Hovy, 2011; mei Sun, 2010; Fernández and Seemann, 2009). We highlight some geographic aspects relalted to these.[2]

"My son" in Spanish appears in Twitter often as "mijo" which is an abreviation of the two words "mi hijo". Figure 2.7 shows the daily pattern of the use of the word. All apearences of the word which we observed were indeed in Spanish. There is a clear inactivity period that corresponds to around 8am in French time. We conclde that the term "mijo" probably originates from the west part of Latin America. Similar behavior characterizes the word "porfa" whose periodogram is given in Figure 2.8. This is a way of shortening the word "please" in Spanish, which is written as "por favor".

In contrast, the word "xk" which means in Twitter-Spanish "because" or "why" (Fernández and

[2] A preliminary shorter version of the this Section as well as the next five sections appeared in (Altman and Portilla, 2012).

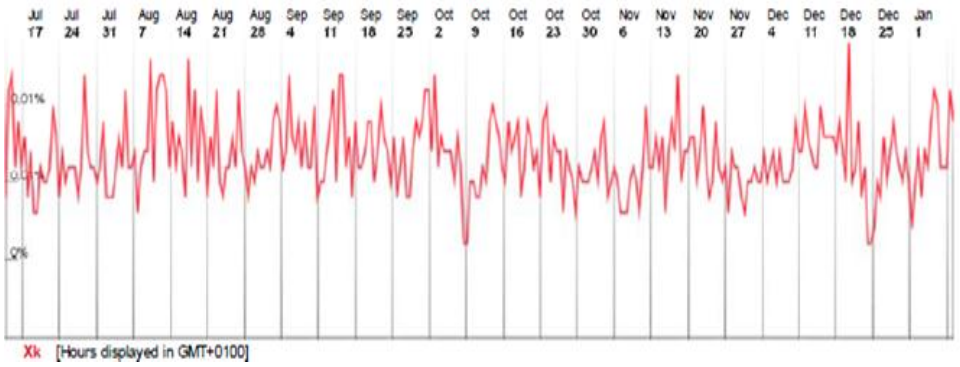


Figure 2.9: *The frequency of appearance of "xk"*

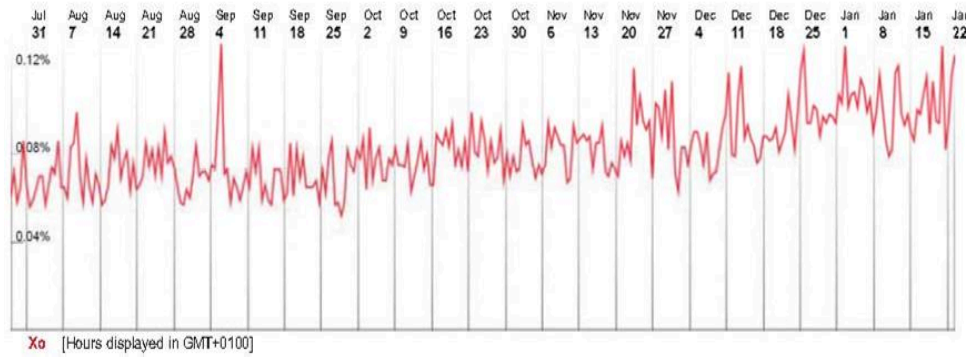


Figure 2.10: *The frequency of appearance of "xo"*

Seemann, 2009) and is pronounced "porque" has no inactivity periods, see Figure 2.9. "xk" is much less localized and is probably used both in latin America and in Spain. Note that the translation of "por" using "x" is due to the interpretation of x as multiplication, which is pronounced as "por".

We next observe the evolution of the word "xo".

The online urban dictionary http://www.urbandictionary.com/ says that x means kiss and o means hugs. xoxo then means "kisses and hugs". We found out that in Spanish "xo" is also used to say "pero" ("but" in English), where the explanation for the use of *x* is as in *xk*.

In addition to spanish, *xk* is also used in the same meaning in Italian (for the Italian word "perche" meaning "why"), see (Rimay, 2010). Note that the first vowel is pronounced different than in Spanish.

When working with trendistic, we can use the ratio between the maximum and minimum of the activity level during a day as a measure of its locality. We shall say that a term is well localized if this ratio is larger than 2.

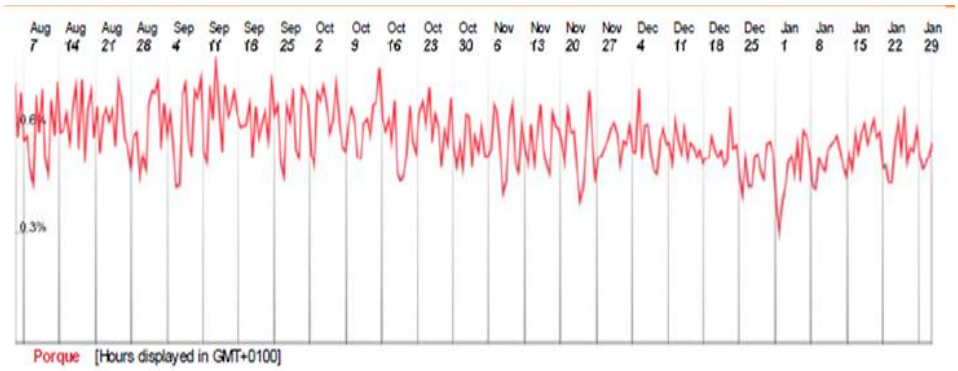


Figure 2.11: *The frequency of appearance of "porque"*

2.5 The Spanish word Porque

We discuss in some more details the spelling we find in Twitter for the words "porque" and "because".

We already mentionned the spelling "xk" for "porque". We found many other spellings. We list them along with the number of tweets in which they appear averaged over the six months period of beginning of Aug 2011 - end January 2012.

We tried also the following spellings: "porque" (0.5%), "xq" (0.07%), "porq" (0.04%), "xk" (0.012%). The frequency of their appearence in Twitter is depicted in Figures 2.11 and 2.12.

Other spelling had too few occurrences and trendistics gave the message "There is too little data for a full chart so we are showing only recent activity". These spellings are "podque", "podq", "podk". They are obtained by replacing the "r" by "d" in the word "porque" and then, for the two last spellings, "que" is abreviated. Such a replacement is a typical childish way of speaking spanish, as many children have difficulties to pronounce the *r* and replace it the by *d*.

We found no tweets with the spelling "xque". The spelling "pork" appears, but most tweets with this spelling correspond to the English word "pork".

2.6 The English word "Because"

This section uses the periodogram approach to study the word "because". The word "because" appears in Twitter with a large number of variations. In fact, the total fraction of tweets in which this word appears in a new form and/or spelling is larger than that corresponding to the original word. This is illustrated in Table 2.1 which provides the most frequent variations of "because" along with the number of tweets in which they appear (averaged over 300 samples) along with

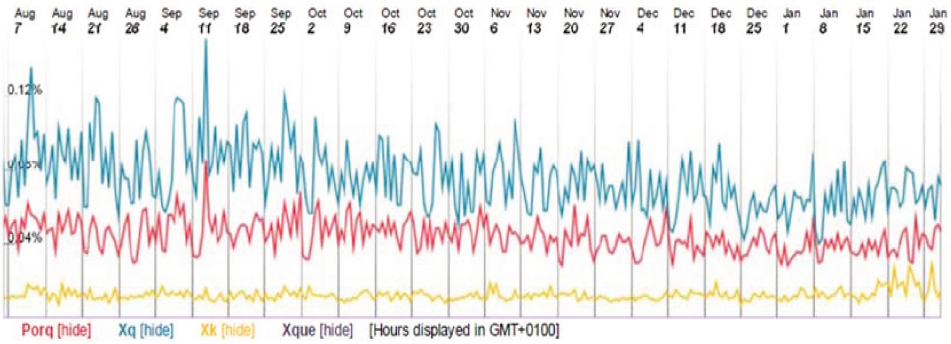


Figure 2.12: *The frequency of appearance of other spellings of "porque"*

Spelling:	because	bcuz	becuz
% of tweets in which it appears:	0.824	0.0156	0.0116
Standard deviation	0.107	0.00350	0.00355

Spelling:	cause	cuz	cos	cus	coz	cz
% of tweets in which it appears:	0.505	0.187	0.0490	0.0304	0.0135	0.0128
Standard deviation	0.0678	0.011	0.0141	0.064	0.011	0.00355

Table 2.1: *The most popullar spellings of "because" in Twitter*

the standard deviation. For reference, the average fraction of tweets in which the word "because" appears without changes is 0.824% of the tweets with a standard deviaition of 0.107%.

Two short forms appear frequently in Twitter: "cuz" (around 0.2% of tweets) and "coz" (around 0.02% of tweets). Such shortening of words are called "clipping" in linguistic research (mei Sun, 2010). Twinglish thus allows us not only to recognize the word but also to hear it, and hence distinguish between the American and British accents. (Note that the opposite happened with the word "xk" which means "why" both in Italian and in Spanish, but is pronounced differently in the two languages.))

In Figure 2.13 we observe the periodograms of both. We see that "cuz" and "coz" have exactly the opposit activity profile: the minimum activity of "coz" are during night hours in Europe where as those of "cuz" are in night time in USA and Canada. The maximum daily activity of "coz" is during day time in Europe where as "cuz" has its maximum activity at day time in America. The periodogram of both words show very well localization: the ratio between the peak and the minimum activity is around 5 for both "cuz" and "coz".

We also find the spelling "cus" and "cos" as seen in Figure 2.14-2.15. Again, the spelling is seen to correspond to the accent. The geo-linguistical finger print of "coz" is seen to be the same as

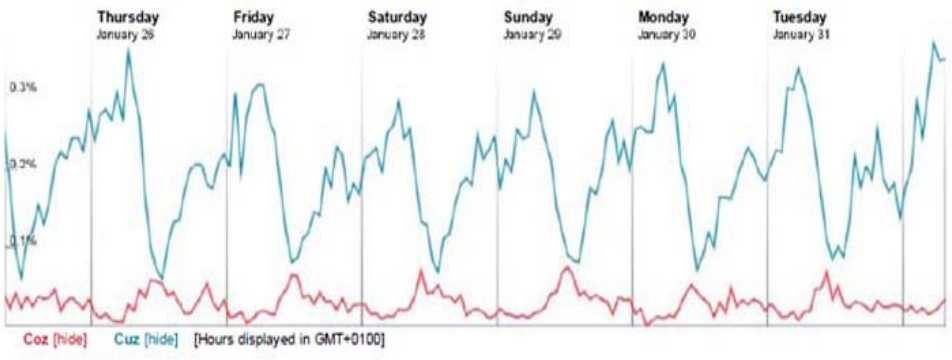


Figure 2.13: *The frequency of appearance of popular spellings of "because"*

"cos" (Fig 2.14). They both correspond to the UK where the second vowel of "because" sounds like "o", as opposed to the American pronounciation that sounds like "u" which we find in "cuz" and "cus".

When comparing the two "American" spellings "cuz" and "cus", we see that there is a very clear preference to "cuz" where as the British seem quite indifferent between the two British spellings "cos" and "coz". The preference of the version with "z" in the America is in line with the fact that there have been already much before Twitter differences between UK and USA with respect to the use of *s* versus *z*.

Further shortning of "coz" and "cuz" by elliminating the vowel is possible but it did not seem appealling to Twitternauts. We have not found "because" written as "cs". It appearead however as "cz", four times less frequently than "coz". From its periodogram in Figure 2.16, "cz" is seen to be very localized and it corresponds to the same activity period as that of "coz". We conclude that the use of "cz" is restricted to Twitternauts from UK.

Two other spelling, "bcuz" and "becuz", appear with lower activity Their periodogram in Figure 2.17 shows activity periods that correspond to in America. We again have high degree of localisation. We did not find tweets with the spelling "bcz" or "bcos".

Remark 1. *Many of the spellings that we presented have other meanaings than because. For example, cos is also the mathematical function "cosine", "cause" usually means "reason" and is also a verb, "cz" is used in other context for the Czech Republic (it has also other meanings). Thanks to the snapshots of the contents of the tweets that trendistics provides, we were able to confirm that the above spellings are indeed used in Twitter mainly in the sence of because. We shall later see that this is not the case in other electronic media.*

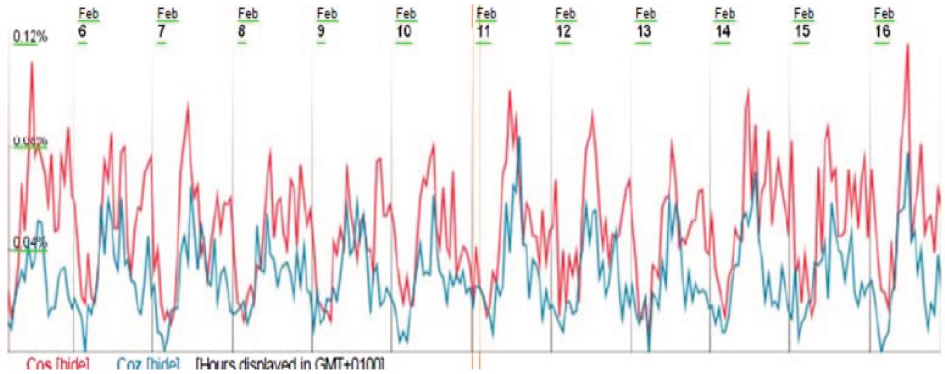


Figure 2.14: *The frequency of appearance of the spellings of "cos" and "coz" of "because"*

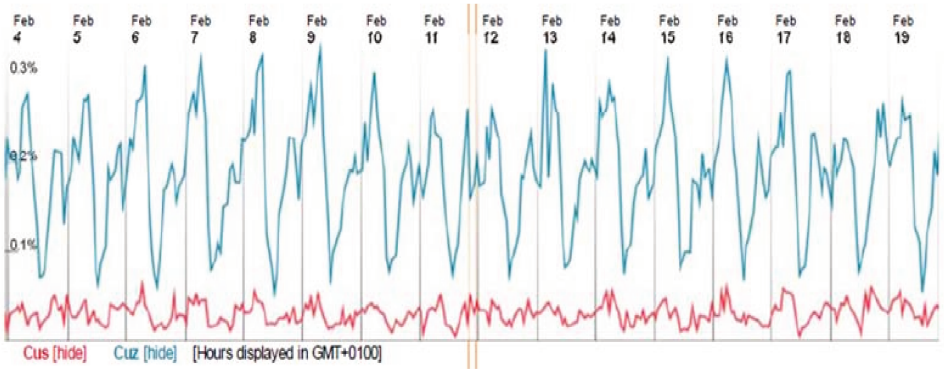


Figure 2.15: *The frequency of appearance of the spellings "cus" and "cuz" of "because"*

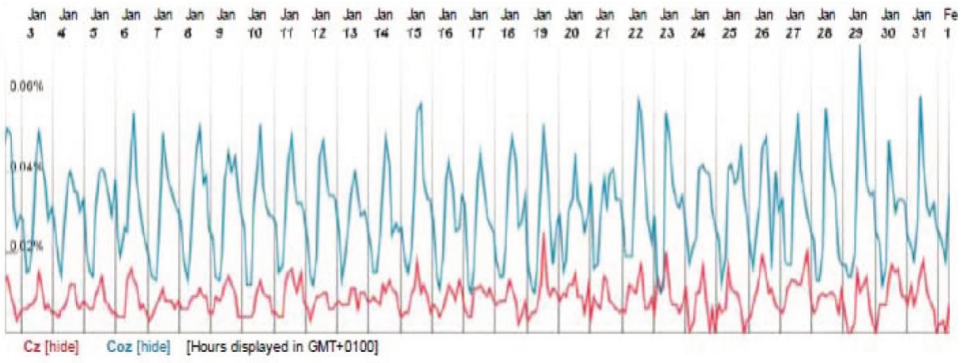


Figure 2.16: *The frequency of appearance of other spellings of "cz"*

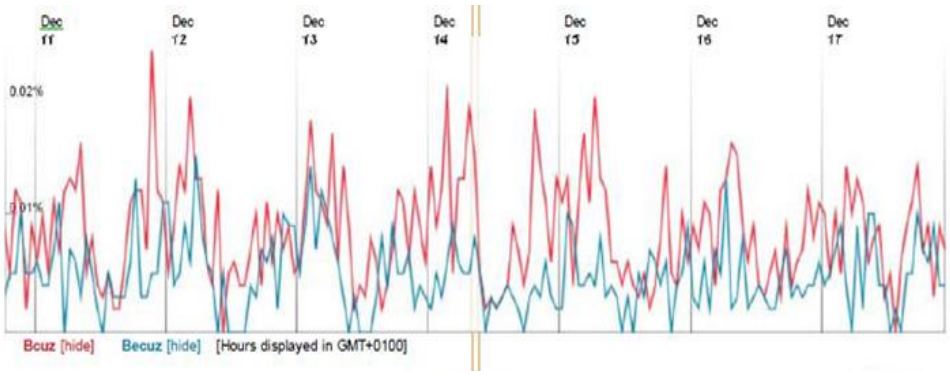


Figure 2.17: *The frequency of appearance of the spelling "bcuz" and "becuz"*

2.7 Combination of different spellings

In view of the large number of different spelling of words, the natural question that arises is of what information is conveyed by a different spelling of a word. We already saw that the spelling can convey to which geo-linguistic group one belongs, e.g. whether the message is written by a British or an American English speaking person. The differentiation between these English versions that already existed since long ago has increased in tweeter, not only with respect to other media but also with respect to other social networks.

In this Section we examine other information and semantic content that spelling conveys. This has been motivated by a large number of messages in Twitter in which we found various spellings of the same word in the same message. Is it intentional?

We focus on the word "because" in English and Spanish. Here are some examples of intentional use of different spellings or forms of the word.

We found different spellings appearing in the same sentence when comparing between two reasons for some action or some decision. Here are some examples that we found in tweets written in the end of March 2012 and the beginning of April.

i. 26 March, 2012: "laugh at u because ur different?? Laugh at them becuz they're the same!"
   The example did not specify in what sense the one or the others are different or not. This is left to the imagination of the reader. My own prefered interpretation is that the difference is in the way of spelling "because", and hence the message advises one not to be too hurt if they laugh at you because you use a different spelling for writing "becuz", but rather to laugh at them becuz they all write becuz using the same spelling.

ii. 26 March, 2012: if you haven't notice by now... i'm really bad at replying to ppl, not

becuz im ignoring you, just because im a lazy person lol.

iii. March 27th, 2012. BTW...that LOL was because of fun memories...not becuz of song that was tweeted by RT

iv. 27 march, 2012: i cannot wait til spring break 2013, nt because of the trip but becuz on my way back ill be saying, this is my last time drivin to muncie :)

v. April 2nd, 2012: In class a professor called US a pure socialist country becuz it bails out all its companies, and Russia a capitalistic one because it wudnt

vi. April 4th, 2012: U don't want to get to kno me becuz of something I did or somewhere I've been...u want to get to kno me just because of how I look..

The same structure appears also in Spanish, for example:

- 4th April, 2012: no llueve porque tu salgas del colegio, llueve xk son vacaciones y en vacaciones siempre llueve yo creo que es por joder!

which says - it is not raining because you got out of school, it rains because these are holidays ...

Another case in which we observe a tendency to alter the spellings of the word "because" in a sentence is when the sentence has a nested structure. For example

- April 4th, 2012: I isolate myself when I feel a certain type of way & I'm trying to get it off my mind because I don't like for people to be down becuz of me

In Spanish, the word "porque" means both "because" as well as "why". Observing all occurrences of both in the week of 26 March - 4 April 2012, 8 out of 9 used "xk" for why and "porque" for "because", and only one case was the opposit. This suggests that new spellings can be used to transfer more information on words that looked the same in previous spellings.

2.8 On the creation of new spellings in Twitter

The creation of new spelling and forms of words in Twitter is often explained by the advantages in writing shorter words: both the character limitation in Twitter as well as the fact that many tweets are sent from cellular phones whose small keyboard is not as confortable as that of a laptop.

In the creation process of new spellings, alpha-numerical symbols often replace cylables according to

(i) the phonetic sound that they are associated with. Examples are 3Q which is used in Chinese as for "thank you". 3 is "San" in chinese, Q sounds like "queue" which when combined gives

"SanQueue". This sounds like "Thank You". We call this an "audio association".
(ii) the graphic form that they have. That symbol "<3" is a "graphic association" of a heart or lips and is used for expressing affection. The number 7 has a form similar to that of the letter "cha" in Arabic and is thus used as such when an arabic keyboard is not available.
(iii) Composition of associations: we saw that "xk" means "because" The "x" is pronounced "por" through a two step association: first a graphical association is used to transform "x" to "multiply", and then the audio association of "multiply", which is "por" in Spanish, is used.

As is the case in "3Q", the audio associations are often innexact. Here are some more examples.

- The letter "k" is pronounced as "ka" in Spanish so that "xk" sounds as "porqua" where as it is used in the meaning of "because" in Spanish, which sounds like "porque".
- "k2" sounds as "KaDeu" in French and means a "present"; the pronounciation of "present" in French is, however, "KaDo".
- The word "your" is often shortened to "yo".

It is not a surprise that there is a big tolerancne to such imprecisions, as we know of natural languages in which the vowels, altogether, do not appear in the written version (e.g. Hebrew or Arabic). Yet, although we see vowels appear often in an imprecise way in Spanish, English and French, the Twitterenauts do not seem eager to drop them completely (as we saw in the shortning of "because").

We saw that Twitternauts often convay the accents they use. This was the case of the word "because" whose Twitter spelling spelling "cuz", "becuz" or "bcuz" suggest the USA accent whereas its spelling "cos" suggests the British one. We showed that this classification is confirmed with a high degree of localization obtained using the periodograms.

Further audible features of words appeared, e.g. in replacing the "r"s by "d"s in Spanish, as we saw in the word "porque". This feature also occurs in English, where the sound "th" in words such as "the", "this" and "that" is sometimes pronounceed as a "d". We illustrate this in Figure 2.18.

2.9 Numbers

We tried the following experiment. We compared the daily frequency of the appearance of the integers 1, 2, 3, ... in Twitter messages. We expected to obtain variations around some common average: we did not expect one number to appear more frequently than another in the long run. Figure 2.19 shows the measured frequencies.

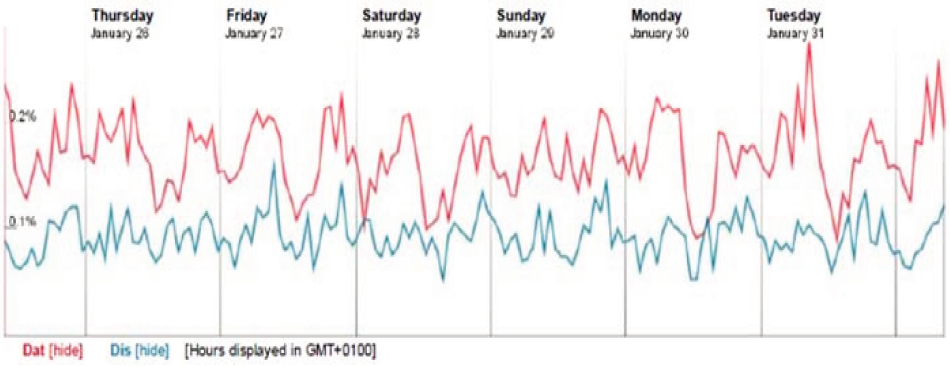


Figure 2.18: *The words "this" and "that" spelled as "dis" and "dat"*

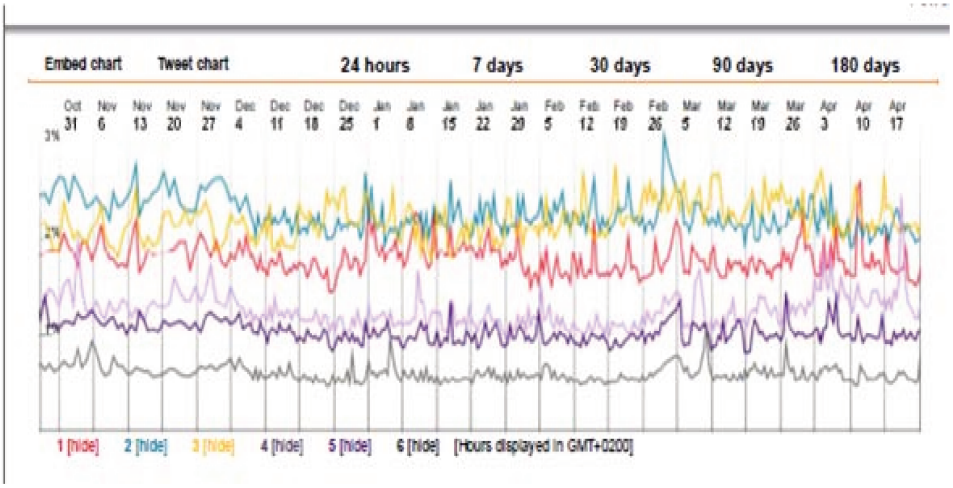


Figure 2.19: *The relative frequencies of the integers 1, 2, 3, 4, 5, 6*

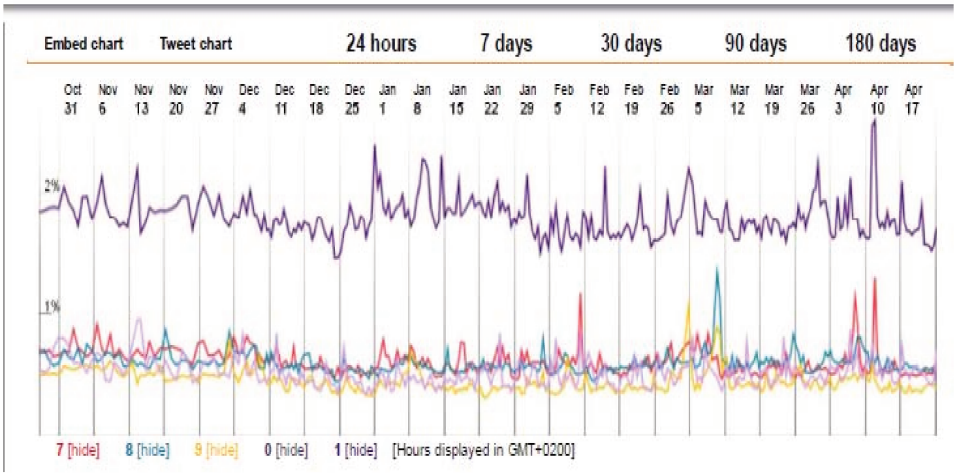


Figure 2.20: *The relative frequencies of the integers 7, 8, 9, 0, 1*

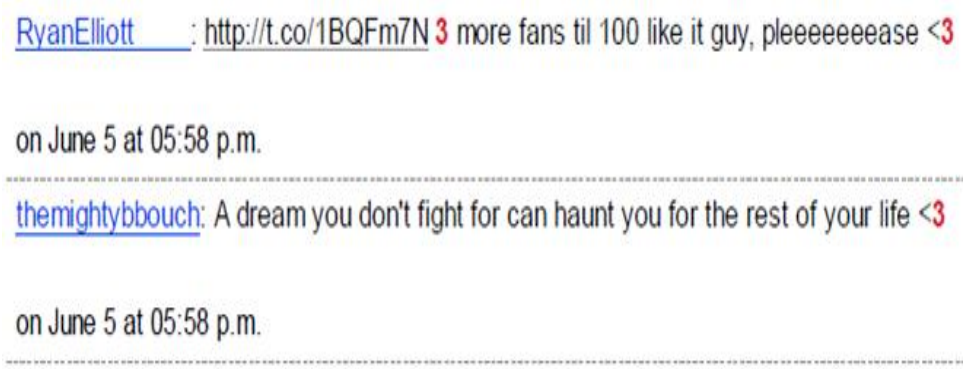


Figure 2.21: *Examples of tweets using "3"*

We observe significant differences between the frequencies of appearance of various integers. What is the reason for that? Looking into the messages themselves, one immediately observes that through the way of pronouncing them, integer numbers have other meanings as well. Integer numbers are then used so as to shorten the number of characters needed for transitting messages. This is extremely useful in Twitter since it

In particular

- 4 is pronounced the same as **for**. Therefore, 4u can be used to write the two words "for you" (which contain seven characters) in two characters only.
- 2 is pronounced like "to". It thus allows again to reduce the number of characters. E.g., writing 2u allows us to express "to you" in 2 characters.
- 3 is used frequently combined with $<$ which gives <3 This can be interpreted as a kiss or a heart, see Fig 2.21.

2.9.1 On 3 and its Evolution

The frequency of the use of 3 increases rapidly as is seen in Fig. 2.22. Since mid july 2011 till mid January 2012 its popularity increased from arounud 2% of the tweets to 2.7%. Its peak is achieved on Dec. 31 with a popularity of 3.42%. this period, the other numbers showed negligible variations in popularity.

The number 3 has several other meanings: 3Q means "thank you" in chinese: 3 is pronounced as "SUN" and "Q" is pronounced as "Queue" so that 3Q together immitates the sound of the English word "Thank you".

The digits 3, 5, 7, and 9 are used also as letters in Arabic when using a latin keyboard to write

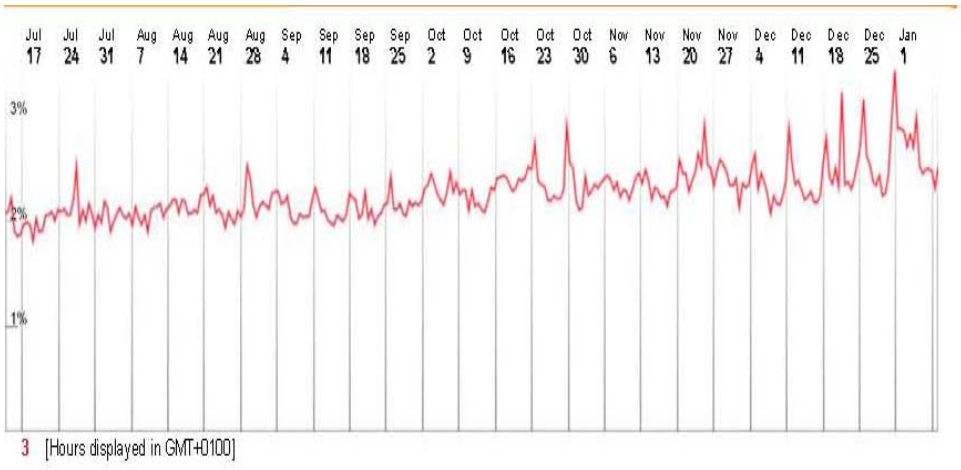


Figure 2.22: *The time evolution of "3"*

arabic. They correspond to the arabic letters "Ain", "Kha", "Cha", "Ka", respectively. The need for using digits is due to the fact that there are no other consonantns in most European languages (e.g. English, French Italian) that are pronounced like these arabic ones.

2.10 Comparisons with other types of media

In this Section we compare the freqency of occurrance of "because" in its different spellings obtained in Twitter to the frequency of its appearance in the World Wide Web. For the latter we used "google search". We repeated the same experiment restricting to those pages that fall into the category of "news" under google.

Our findings are summarized in the Table 2.2.

The table presents the normalized popularity: at each row, we divided the corresponding number by the first one in that row. In that way we can compare the relative "popularity" of each form of "because" on different media.

The fractions of the versions of "because" that are obtained by shortening, "becuz" and "bcuz", are seen to be much higher it tweeter than those obtained in google and in google news. The number of appearances of each one of these two versions, divided by the number of times that "because" appears, is more than 500 times larger in Twitter than over the whole Internet. It is more than 100 times larger in Twitter than it is over news documents found by Twitter.

This ratio in Twitter is also larger that in the whole internet for the versions obtained by further clapping the word "because" (where the first cylable disappears) with one exception: the word

Spelling:	because	bcuz	becuz
Normalized popularity			
trendistics	100	2.00	1.41
Google	100	0.00251538	0.00264231
Google News	100	0.0171975	0.0140127

Spelling:	cause	cuz	cos
Normalized popularity			
trendistics	61.3	22.7	5.94
Google	26.8461538	0.8076923	1.0384615
Google News	0.656051	7.1974522	48.4713376

Spelling:	cus	coz	cz
Normalized popularity			
trendistics	3.69	1.64	1.55
Google	0.0869231	0.2115385	0.9769231
Google News	0.656051	7.1974522	48.4713376

Table 2.2: *Normalized popularity of variants of because in %*

"cause" which is used more on the Internet since the other meanings of the word "cause" appear more frequently there. This is also true to the finding over google news.

Two other exceptions occur with respect to google News: "cz" is very frequent there as it is used with the meaning of the Czech Republic. The spelling "cos" also has many other uses that appear in google News.

2.11 Other applications of the periodograms

The fingerprints could be used to evaluate how popular a given product or event or person is among different communities. As an example, a politician may wish to know how "popular" he is and furthermore, within which community he is popular.

As a "popularity" criterion we shall take the presence in Tweeter. We wish to analyze the fraction of tweets sent by members of a given community which mention some keyword.

Example 1

Fig. 2.23 shows the frequency of the words "Obama", "these", "vamos" and "der". Observe that there are regions in which the frequency pattern of Obama is similar to that of "These" and "Vamos", and others where it is similar to "der".

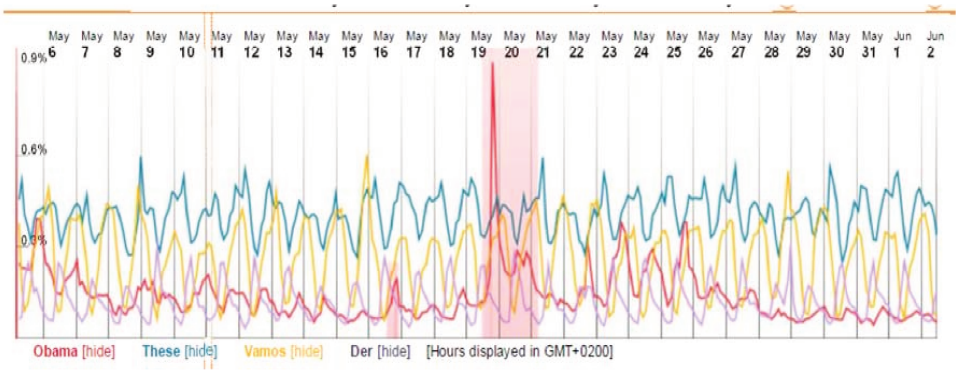


Figure 2.23: *The frequency of appearance of the words "Obama", "these", "vamos" and "der"*

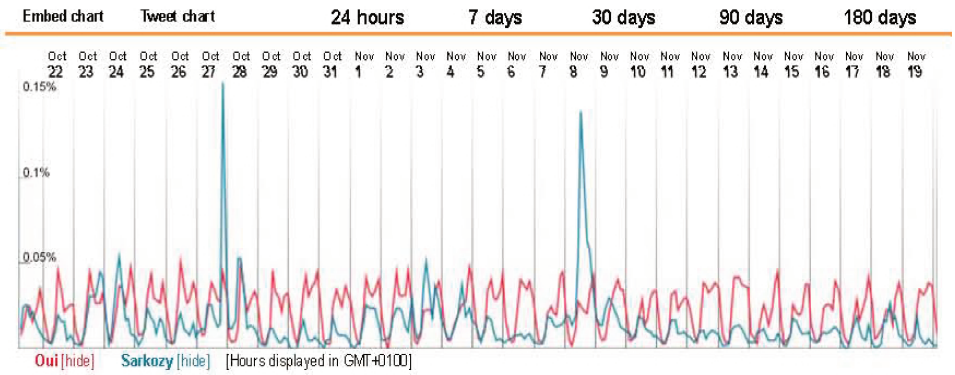


Figure 2.24: *Buzzes of Sarkozy*

For example, the minima of the curve corresponding to Obama coinsides with that of "der" in the week after May 20. i

His curve seems closer to that of "these" or of "vamos" in the week starting from May 8 during which he was in the USA. The geo-linguistic finger prints thus suggest a shift of the interest in Obama from the USA to Europe.

Examining the tweets, one can correlate the shift of interest in Obama to Europe with the fact that indeed Obama was visiting Great Britain in that week.

Example

In Figure 2.25 we observe 2 peeks in the popularity of Sarkozy within Tweets. We shall call such a peek a "buzz". We included in the figure the periodogram of the French word "oui". By comparing the periodogram to the two buzzes, we see that the first coincides with a peek of the use of "oui" where as the other one does not. This simple and quick comparison allows us to conclude that the first buzz concerned the French tweeters while the second one did not.

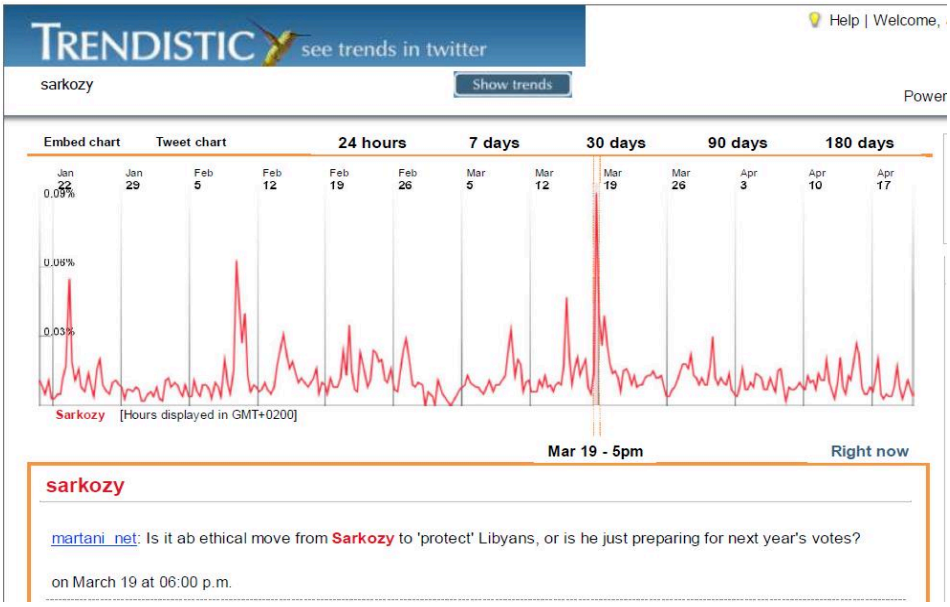


Figure 2.25: *The frequency at which the name of the French president appears in the tweets*

Looking into the tweets that correspond to the second buzz we see that they concern a report that Sarkozy called the Israeli prime minister, Netanyahu, a liar during a conversation with Obama. This is confirmed in the next Figure that shows that indeed Netanyahu has a peek at the same time. The first buzz, in contrast, concerns the birth of Bruni-Sarkozy daughter and it interested the French more than Netanyahu did.

The Figure also depicts theh word "liar" in several languages. It very much used in English tweets (even more than the word "Netanyahu"), it is used a little bit in Spanish, and almost not at all in French. This confirms the conclusion already obtainied from the French geo-linguistic finger print, that there was little interest in the event in France.

2.12 Understanding causality relationship

We show below an example that illustrates how the analysis of tweets periodograms can be used for political analysis, and in particular, that of election campaigns. We start by observing the periodogram of appearence of the French president during 2011, N. Sarkozy, in tweets, for a period of 6 months. This is given in Figure 2.25.

We observe that the highest popularity during this 6 months experiment was attained on March 19, 2011. This jump of popularity is due to the fact that France started bombing Lybia on that day. Many tweets mention the possible relation between this attack and the presidential

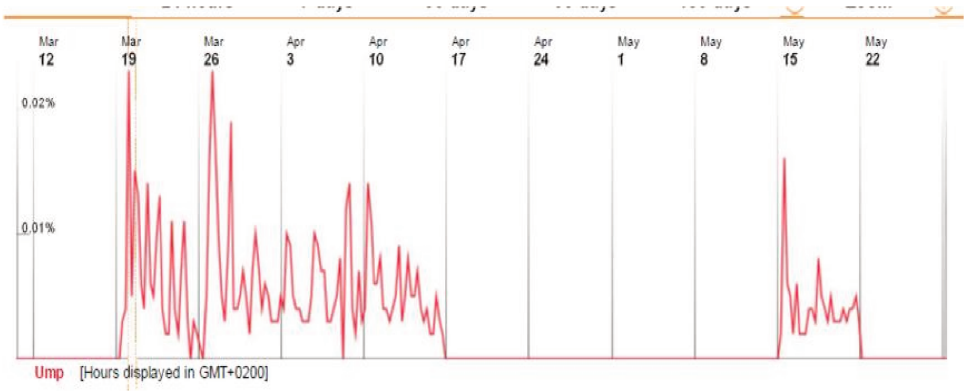


Figure 2.26: *The frequency at which the name of the French party UMP appears in the tweets*

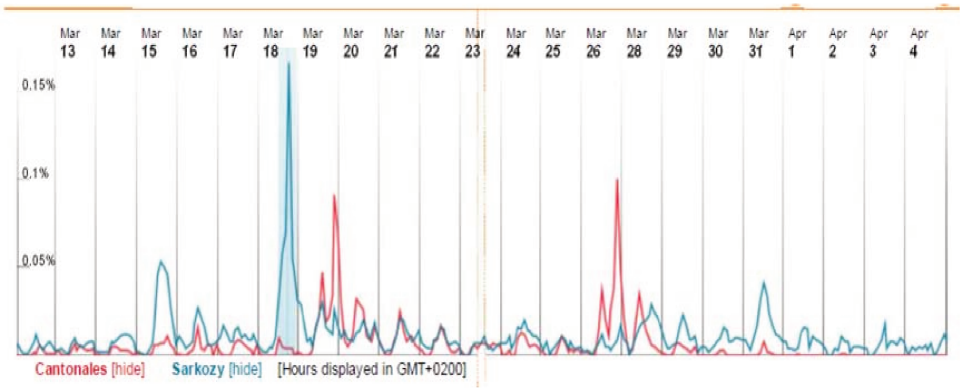


Figure 2.27: *The time relation between the French local elections and the revival of the UMP*

campaigns (the presidential election in France took place on 2012), see for example, the tweet that appears in Figure 2.25. But another causal relation that can be identified using tweeter would shed a light on this event.

The political party that supported the French president Sarkozy is called UMP (i.e. "Union for the Presidential Majority"). If we observe it in the tweets, we discover that it has most of the time no tweets at all. Indeed, we may observe in Figure 2.26 the popularity of the UMP as a function of time. We see that it comes to life on March 19 and remains alive till it becomes inactive again three weeks later.

The importance of the resurection of the UMP on that precise timing can be seen from Figure 2.27 that considers the words "UMP" as well as "Cantonales" (which stands for the local elections in France). the three weeks of life that the UMP is seen to get from the peak of popularity of Sarkozy's foreign policy is seen to cover the two peaks (which are one week apart from each other) that correspond to the timing of the two rounds of local French elections.

We can now summarize: the UMP party is faced with local elections. It is however dead from

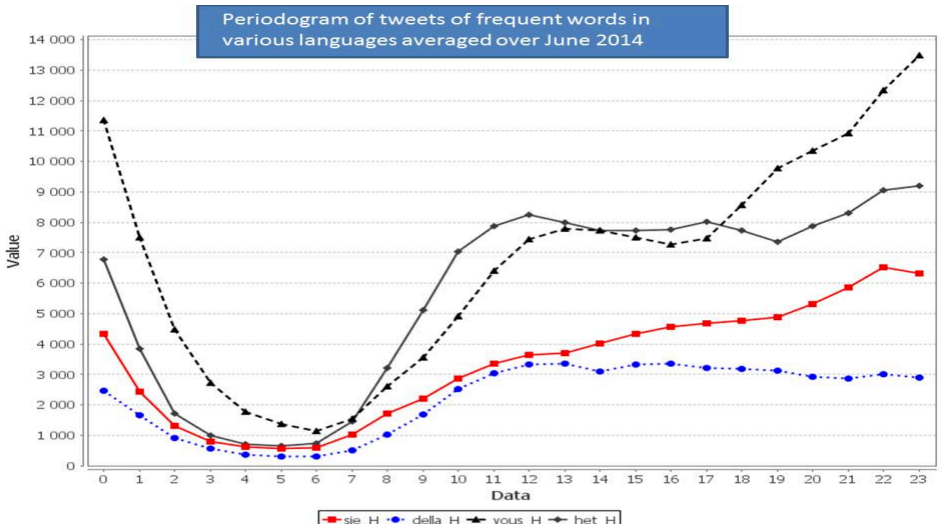


Figure 2.28: *The periodogram of appearance of words tweeted in different languages originating from areas that share a common time zone, averaged over all the days of June 2014.*

the point of view of coverage in some electronic media (Twitter). The attack on Lybia brings the UMP back to life (in Twitter) in order to be present (in the media) during both first and second rounds of the local elections.

2.13 Refining the geolocation

This chapter used a daily periodogram approach in order to locate the longitude of areas in which given words appear frequently in tweets. A future step will be to differentiate between words spoken in areas that share the same longitude. Will the periodogram enable to separate such areas? It turns out that the tweeting habits as a function of the time of the day do differ between various zones sharing a common longitude, but the periodograms are too noisy to detect these. In order to to reduce the noise we averaged the periodogram over the whole month of June in 2014. We chose words of four languages that have a common time zone:

i. German with the word "Sie" (you),

ii. Italian with the word "della" (of the), and

iii. French with the word "Vous" (you),

iv. Dutch with the work "het" (the)

The respective periodograms are given in Figure 2.28.

We first note that the quality of the geolocation (of the longitude) is quite satisfactory according

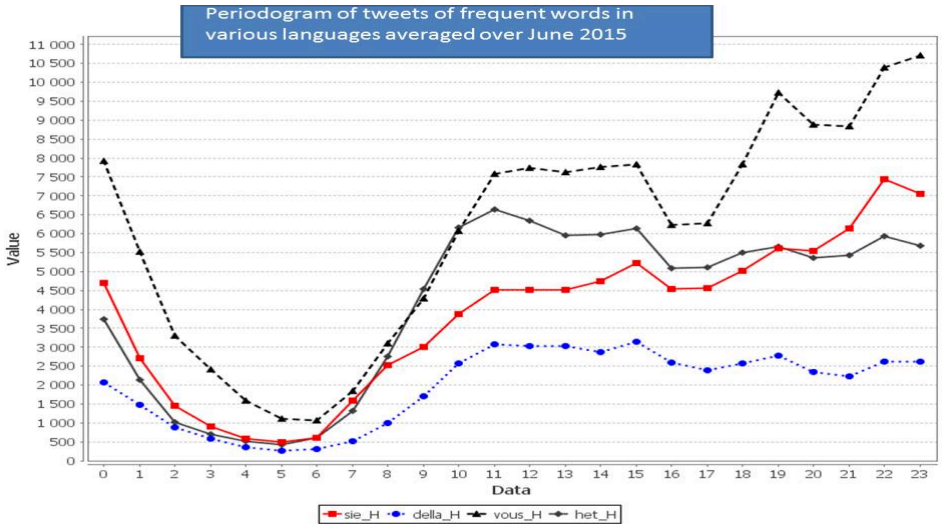


Figure 2.29: *The periodogram of appearance of words tweeted in different languages originating from areas that share a common time zone, averaged over all the days of June 2015.*

to our RTR criterion of Section 2.3 and in all words the ratio between the maximum and minimum tweeting rate in the periodogram is around 10. We see that the largest tweeting rate in Italy is from 11h00 till 23h00. During this period it is almost constant - it only has a slight decrease (by aboout 15%). While Dutch periodogram has a similar behavior, we see that the French and German curves have their peak activity at around 22h-23h, during which the tweeting intensity grows to more than 50% than than at 11h.

Note that French is also spoken in other time zones but the high quality of localisatioon of tweets with French worods suggests that the percent of French messages tweeted from countries other than France is quite small.

We repepat the same exdperimentation one year later, at June 2015. The periodograms are given in Figure 2.29 and are seen to provide similar trends as in 2014.

We conclude that it is possible using averaging to obtain more information from tweets so as to differentiate between features of communities that are separated geographically.

//

Acknowledgement

The authors wish to thank C. E. Contreras Oller (D partment of modern languages, Univ de Los Andes, Mérida, Venezuela), Prof Aharon G. Kleinberger (Department of Arabic Language and Literature, Haifa University, Israel) and Julien Gaillard (Univ of Avignon).

Chapter 3

Social Networks: a Cradle of Globalized Culture in the Mediterranean Region

3.1 Introduction

According to Wikipedia, the word meme is a shortening (modeled on gene) of mimeme (from Ancient Greek), and it was coined by the British evolutionary biologist Richard Dawkins in The Selfish Gene (daw, 1989)[p. 192] as a concept for discussion of evolutionary principles in explaining the spread of ideas and cultural phenomena. Examples of memes given in the book included melodies, catch-phrases, fashion, and the technology of building arches (daw, 1989)[p 352].

Wikipedia further defines Internet memes as a subset of this general meme concept specific to the culture and environment of the Internet. Wikipedia cites from (Solon, 2013) Dawkins who characterized an *Internet* meme as being a "meme deliberately altered by human creativity-distinguished from biological genes and Dawkins' pre-Internet concept of a meme which involved mutation by random change and spreading through accurate replication as in Darwinian selection".

We shall focus in this chapter on three memes and examine some of their impact on the mediterranean region: the Harlem Shake, the Gangnam Style and TiK ToK.

- The Harlem Shake is a cultural phenomenon that consists of recreation of a danse using always theh same song "Harlem Shake" by the electronic musician Baauer. It is not the ephemerary danse that is important but rather, the trace that is left on videos that propagate on the Internet which are often well editted versions of the original performance of

the shake. It lasts 31 seconds and consists of two parts. It starts with one person (often helmeted or masked) dancing for 15 seconds, often surrounded by other people who ignore him or who are unaware of him. Then the bass drops, and an entire crowd joins doing a strange convulsive dance for the next 15 seconds. The first video entitled "Do the Harlem Shake" was posted on February 2013 by the Japanese video artist Filthy Frank.
The impact of Harlem Shake on the Mediterranean area During March 2013, Tunisians participated in an unprecedented creativity wave of Harlem Shake. School pupils are the first to reproduce the danse. There was one performance of school children in Tunis that particularly shocked the Education Minister, who then reacting in blaming the headmaster of that school. This ignited massive protests of pupils and students in public areas, schools and universities in which the Harlem Shake was performed as symbol of the freedom of expression.

- The phrase "Gangnam Style" is a Korean neologism that refers to a lifestyle associated with the Gangnam District of Seoul (Wikipedia). According to Wikipedia, "Gangnam Style" is the 18th K-pop single by the South Korean musician Psy. The song was released in July 2012 as the lead single of his sixth studio album Psy 6 (Six Rules), Part 1, and debuted at number one on South Korea's Gaon Chart. On December 21, 2012, "Gangnam Style" became the first YouTube video to reach one billion views and since May 31, 2014, the music video has been viewed over two billion times on YouTube. Park Jaesang had been busted for marijuana and for avoiding the country's mandatory military service. His first album got him fined for "inappropriate content" and the second was banned (Fisher, 2012).

- "Tik Tok" (stylized as "TiK ToK") is the debut single by American recording artist Kesha. The song was produced by Benny Blanco and Dr. Luke and co-written by Blanco, Dr. Luke and Kesha. It was released on August 7, 2009 as the lead single from Kesha's debut studio album, Animal (see Wikipedia). With TiK-ToK, Kesha wan the best new artist MTV Video Music Award on 2010. Since it was uploaded on 2009 The number of views in YouTube of TiK ToK has been growing at quite a constant rate, exceeding 150 million views on October 2013 and reaching 192 hundred million views on Sept 2014.

In the next section we introduce the tools we used for the analysis of the above three Internet Memes. We then present a quantitative spatio-temporal analysis of the Memes followed by a section that describes the political impact of all three Memes. We end the chapter with a conclusion section. A 10mins video in French that summarises this chapter is available at (Altman and Portilla, 2013).

3.2 Analytical tools

Or first tool in this work has been Google Trends which is publically available. It allows to obtain the time evolution of the

- $N1$ - the normalized number of times that some sequences of words appear in the title of different videos in YouTube, and
- $N2$ - the normalized number of times of the number of different documents accessed by google that contain the sequences.

By normalized number we mean the following. The largest peak among the different sequences is normalized to 1 and so all other peaks are smaller than 1.

Google Trends provides not only a comparison in time but also a comparison in space. Indeed, it further provides the normalized numbers N1 and N2 in the seven countries and in the seven cities in which N1 and N2 are the largest.

The API Topsy of tweeter allows to obtain the number of tweets in which various sequence of words appear. We used our own tool to visualize these numbers. To use Topsy one should open an account and request a password. There is a limit on the number of a consultations allowed by Topsy per day.

3.3 Analysis of the three Internet mimes

In what follows, we compare the number of appearances of Harlem Shake on the WEB in different geographifcal areas.

Table 3.1 provides the normalized number of appearences of the Harlem Shake on the WEB ranked according to countries. In is normalized such that the country where the Harlem Shake is mostly viewed has 100 view units. Table 3.1 shows the same for Gangnam Style. We see that Tunisia appears in the 6th rank worldwide. Note that no European country appears in the sevem first positions. The ranking was obtained with the Google Trends publically available software.

Table 3.2 provides the normalized number of appearences of the Harlem Shake on the WEB ranked according to cities. These are obtaind again using Google Trends. Beside metropoles with long established cultural tradition as New-York or Paris, we see that the Mediterranean city, Istanbul, is ranked fourth worldwide in the appearence of Harlem Shake. The penetration of the Harlem Shake in the Mediterranean region is even more remarkable if we recall that the song is in Spanish.

The Gangnam Style is also seen to have penetrated into Turkey, as is seen by its fourth worldwide rank obtained with Google Trends in table III. Note that all other countries among the first seven ranked ones are from Asia, which is also the continent where the Gangnam Style come from.

We next use Google Trend to compare Gangnam Style with Harlelm shake. The results are given in Fig. 3.1. In the upper part of the figure we compare them over YouTube. This is not a comparison of the number of views of each of the original videos but the time evolution of number of new videos that have Gangnam Style (resp., Harlem Shake) in their title in YouTube. The second comparison in the lower part of the figure is on the number of appearences of these words on the WEB. Both the comparison over YouTube and over the Web provide very similar behavior and in both we see that the peek for Gangnam Style is slightly larger than that of Harlem Shake, whereas the total number (which corresponds to the area beneith each of the corresponding curves) is much larger for Gangnam Style.

In Figure 3.2 we compare the number of tweets in which Gangnam Style and Harlem Shake, respectively, are mentioned. This is done with the help of an application that we developed named TOPSY. The situation in Tweeter is reversed: The peek of the Harlem Shake is much larger than that of Gangnam Style where as the total number of appearences in tweets are similar.

Fig. 3.3 provides a comparison between TiK ToK by Kesha (corresponding to the first peek) and Harlem Shake (corresponding to the second peek) over YouTube using Google Trend. We note that the number of videos with Harlem Shake in their title is similar to that of TiK ToK. Yet the peek number is much higher for Harlem Shake and is attained during a period much shorter. And then the decay of Harlem Shake is much faster: the time to reach half its peek value is much shorter. The behavior of Gangnam Style is similar to that of Harlem Shake as we saw in the previous figures.

In the two following figures we present another way to compare the the geographic propagation of mimes, i.e. the penetration of a Mime to various countries. We focus on Harlem Shake and compare the number of co-occurrences of the words Harlem Shake together with each one of

Trinidad and Tobago	100
Indonesia	97
Puerto Rico	89
Paraguay	88
Jamaica	78
Tunisia	73
Honduras	69

Table 3.1: *The normalized number of appearence of "Harlem Shake" on the WEB at different countries.*

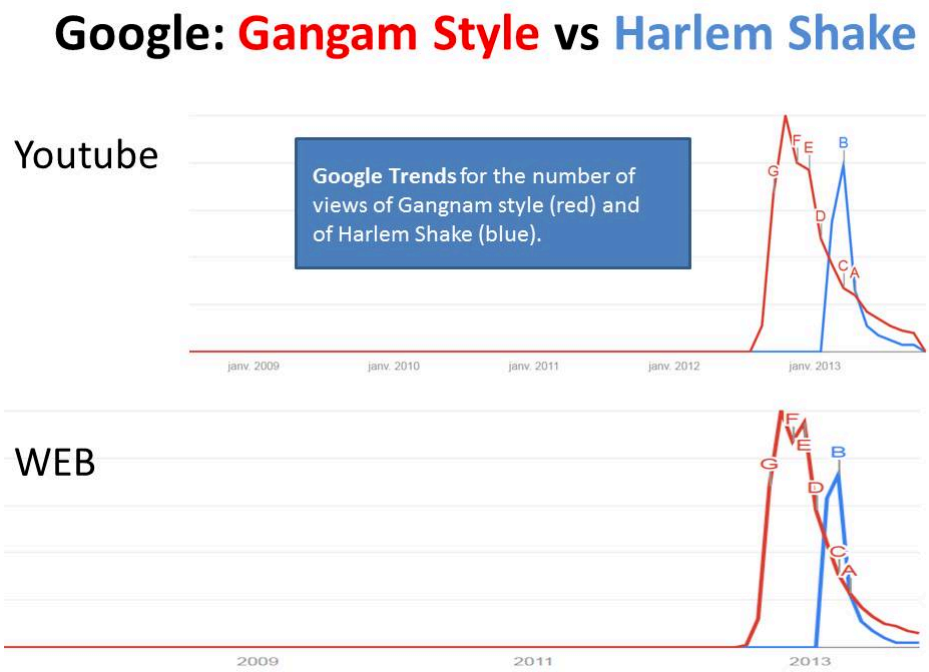


Figure 3.1: *Comparison between Gangnam Style (the left curve) and Harlem Shake (the right curve) both over YouTube as well as over the entire WEB.*

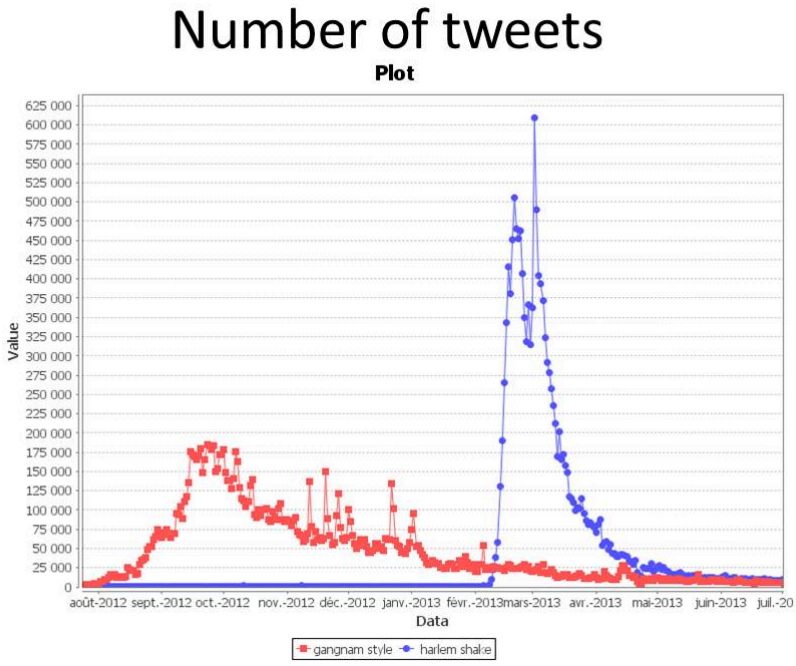


Figure 3.2: *Commparison of Gangnam Style and of Harlem Shake in Twitter*

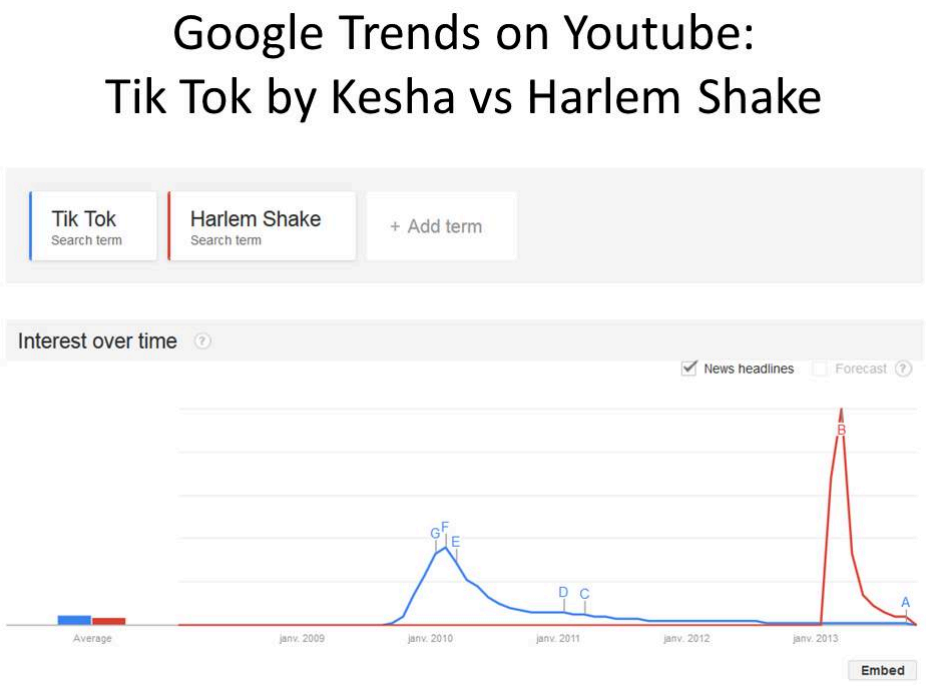


Figure 3.3: *TiK ToK vs Harlem Shake*

the words US, Egypt and Tunisia. We do so for both the number of co-occurrence over the WEB (Fig 3.4) as well as for the number of videos having these words in their title (Fig 3.5) with the help of Google Trend.

Both over the WEB and over YouTube we see that it took around a week from the moment that Harlem Shake started to propagate in the US till is started to propagate in Egypt and in Tunisia, where as its decay took much longer in the USA. Figure 3.4 shows that the size of the peeks at Tunisia and Egypt for the WEB is respecively one quater and one half the size in the US. This corresponds to the comparison of the number who speak of Harlem Shake. Figure 3.5 shows that the corresponding peek number of videos created in Tunisia is around one third the number in

Jakarta	100
Mexico City	79
New York	78
Istanbul	73
Chicago	68
Paris	67
San Paolo	62

Table 3.2: *The normalized number of appearence on the WEB of "Harlem Shake" on the WEB at different cities.*

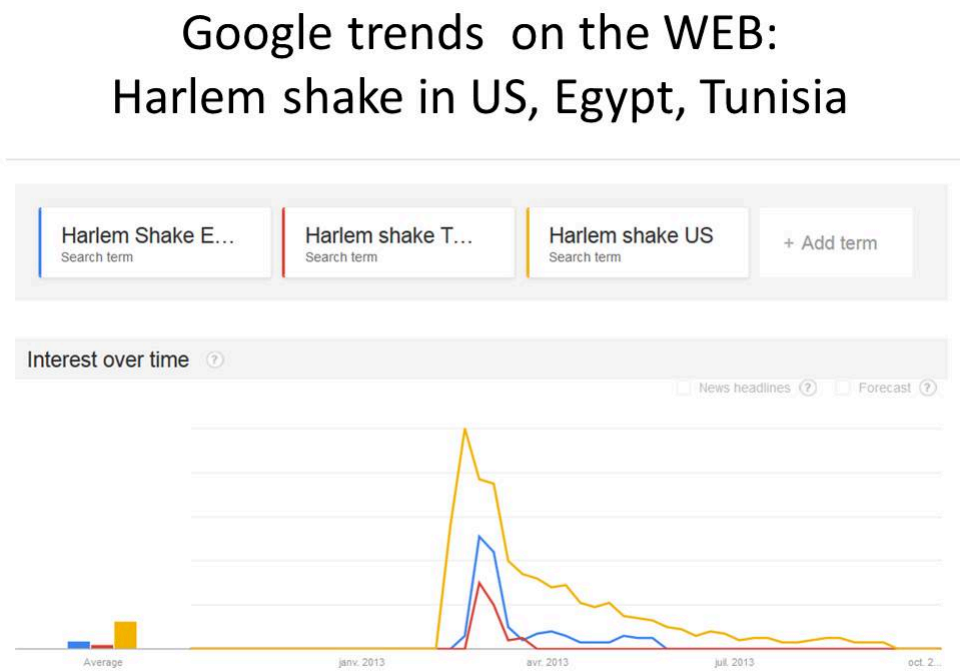


Figure 3.4: *Geo-localisation of events using co-occurrence of words on the WEB*

Egypt, which in turn is only around 10% lower than the peek in the US. This shows amaizingly high creativity in Tunisia and Egypt in terms of creation of Harlem-Shake videos.

All the three mimes were seen to be viral in that they reach a clear peak and then the daily number of views decrease. In the next Figures we provide the viewcounts of Gangnam Style and of Tik-Tok the videos. While the daily viewcount of Gangnam Style in Figure 3.6 has a similar form as the ones for other measures (N1 and N2 in 3.1 or the number of tweets in 3.2), the daily view count of TiK-ToK is flat (or equivalently, the cumulative viewcount that we see in Fig. 3.7 is increasing with a slope almost constant). Thus the viewcount does not have an epidemic behavior while the Internet mime that TiK ToK created does.

Vietnam	100
Mongolia	52
Cambidua	39
Turkey	39
Sri Lanka	36
Malaysia	35
Philippines	33

Table 3.3: *The normalized number of appearence of "Gangnam Style" on the WEB at different countries.*

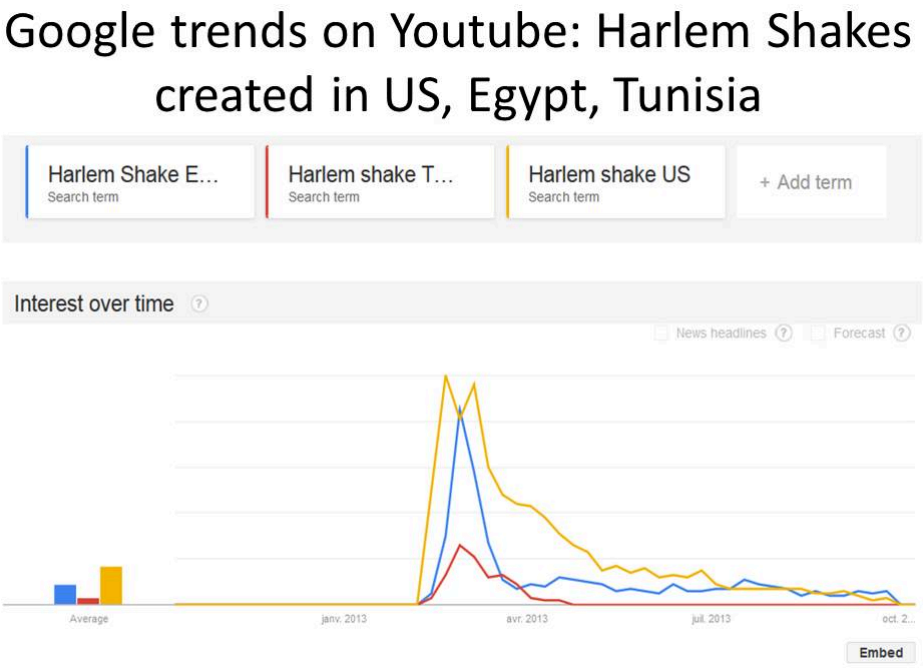


Figure 3.5: *Geo-localisation of events using co-occurrence of words on YouTube*

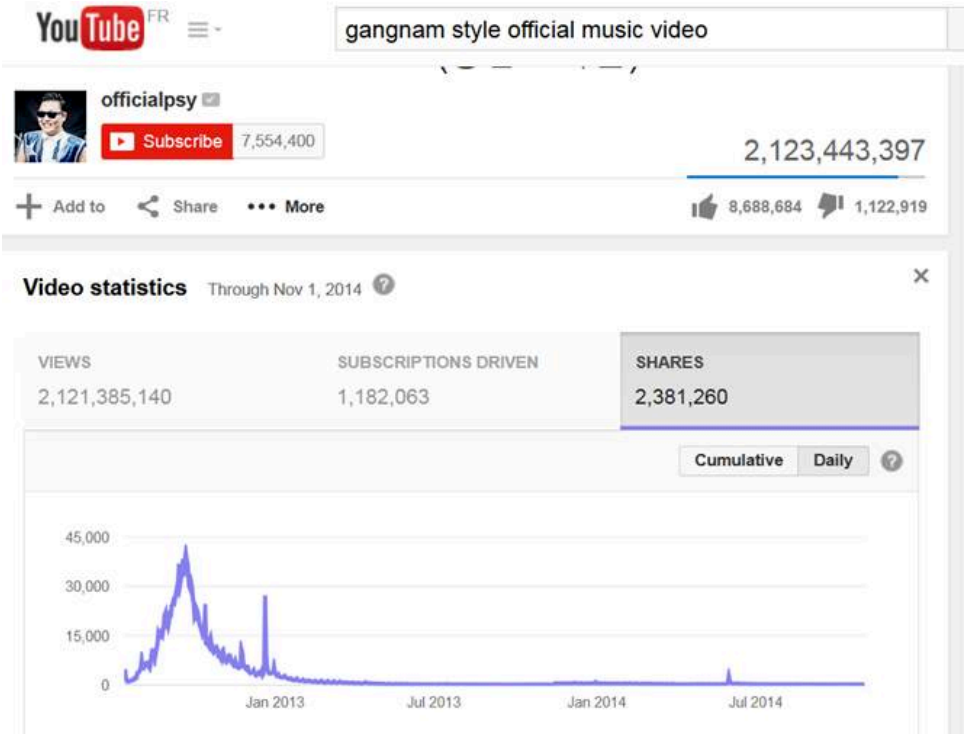


Figure 3.6: *Daily viewcount of the original video of Gangnam Style*

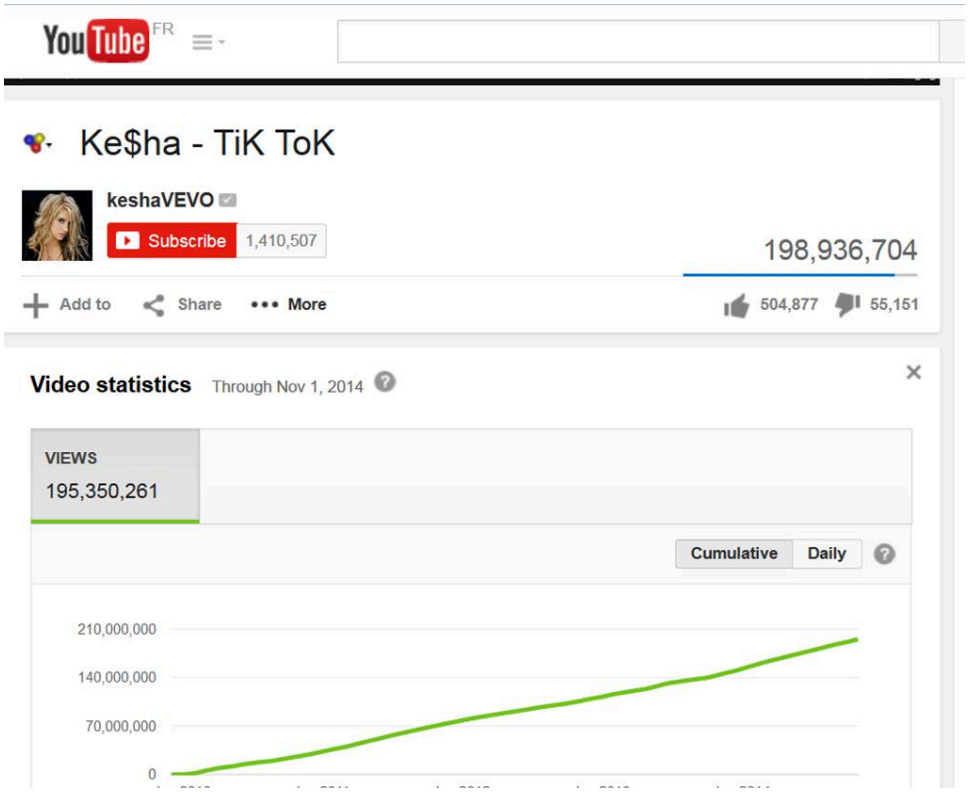


Figure 3.7: *Daily viewcount of the original video Tik-Tok*

The comparison with the viewcount of the original video of Harlem Shake is impossible since it is not available on YouTube. Indeed, not all videos have the viewcount curve available on YouTube. The viewcount may be blocked either by the creator of the video or it is YouTube that may block it if there are complaints concerning infringement of copyright by the video. It is the this second reason that holds in the case of Harlem Shake [Wikipedia].

Note that the whole cooperative process of creating new versions of Harlelm Shake involves copyright infringement since they are based on copying the same original song. From the authors experience as well as many other ones', this results in the fact that new versions of Harlem Shake cannot be uploaded to facebook. Yet, YouTube is more flexible and allows them to be uploaded without deleting them as long as there is no request to do so by someone who can proves that he is the copyright owner.

3.4 Impact of penetration of Memes into the Mediterranean region

We have shown that the new Internet Memes traverse borders separating countries. They create a globalised culture that bring nations closer to each other.

On August 29, 2013 Israeli soldiers were making the rounds in the city of Hebron when they heard the "Gangnam Style" from a wedding party. They accepted the invitation to join the party and joined dozens of Palestinian men dancing to the hit "Gangnam Style." In a rare video they are seen with their uniform and guns, carried on the sholders of young Palestiniens. Sommer writes in (Sommer, 2010): *Underneath the animosity and the history, the labels of "Israeli soldier" and "occupied Palestinian" are simply young people who just want to have fun. Watching them dance together instead of confronting each other offers a glimpse of a Middle East that while doesn't seem possible in our lifetime, is lovely to contemplate.*

It is hard for an army to accomplish occupation tasks against another nation that shares the same culture. The soldiers that participated in the Gangnam Style were thus suspended from the army. Ironically, Park Jaesang, the author of Gangnam Style, had been jailed busted for for avoiding Korea's mandatory military service. The new Internet Memes thus constitute a danger for armies as they constitute a cradle for fraternity.

This also explains the disciplinary actions against Israeli soldiers that were involved in organizing of Harlem Shake on February 2013 (Winer, 2013). A soldier produced the Harlem Shake video, was sentenced to 14 days in prison and his commanding officer was sentenced to 21 days in jail and was stripped of his command.

Another YouTube video posted in 2010, shows a six-man patrol of soldiers walking along a Palestinian street in the city of Hebron under Israeli occupation, in full battle gear. In the midst

of the patrol, the soldiers suddenly break into a one-minute dance on the TiK ToK music of Kesha (Pfeffer, 2010). The video was removed from the YouTube account where it first appeared after Israeli television reports brought it to the attention of the military, but not before other video bloggers managed to copy and republish it. Mackey writes in (Mackeyi, 2010) concerning this event: "It is a reminder that while the war of ideas in the Middle East might have ancient roots, it is often waged by young people more interested in contemporary youth culture than age-old texts."

/

Chapter 4

A Study of YouTube recommendation graph based on measurements and stochastic tools

4.1 Introduction

Online media constitute currently the largest share of Internet traffic. A large part of such traffic is generated by platforms that deliver user-generated content (UGC). This includes, YouTube and Vimeo for videos, Flickr and Instagram for images and all social networking platforms. Among these platforms, YouTube has become the most popular. Based on statistics available from the website Alexa.com, more than 30% of global Internet users visit youtube.com per day. Other statistics from *http://www.youtube.com/t/press_statistics* clearly illustrate the previous fact: "Over 800 million unique users visit YouTube each month" and "72 hours of video are uploaded to YouTube every minute". Of course, not all videos posted on YouTube are equal. The key aspect is their "popularity'", broadly defined as the number of views they score (also referred to as view count). This is relevant from a twofold perspective. On the one hand, more popular content generates more traffic, so understanding popularity has a direct impact on caching and replication strategy that the provider should adopt. On the other hand, popularity has a direct economic impact. Indeed, popularity or view counts are often directly related to click-through rates of linked advertisements, which constitute the basis of the YouTube's business model. Hence the revenue model of YouTube is based on a sophisticated advertising scheme. Indeed different types of advertising methods are used in YouTube, for example, "In-video graphical and text advertisements", "post-roll advertising", etc. Extracting incomes in such business models is

related to content visibility, which motivates YouTube to provide the recommendation list and to display features videos.

Models for predicting popularity of online content including YouTube videos and Digg stories, are proposed in (Cha et al., 2007; Crane and Sornette, 2008; Gill et al., 2007; Ratkiewicz et al., 2010; Chatzopoulou et al., 2010; Cha et al., 2009; Richier et al., 2014), with the aim of developing models for early-stage prediction of popularity features (Szabo and Huberman, 2010). Such studies have highlighted a number of phenomena that are typical of UGC delivery. This includes the fact that a significant share of content gets basically no views (Cha et al., 2009), as well as the fact that popularity may see some bursts, when content "goes viral" (Ratkiewicz et al., 2010). Visibility of content is not just of interest to YouTube. It is also of interest to the content consumers and to the content creators. There is a competition between the creators on visibility of their creations. Understanding of how view count of a video is driven by different sources of views is helpful for finding strategies for increasing the number of views of videos. For advertisers and for content providers, this is useful for strategic planning so as to increase the contents' popularity. This is often directly related to click-through rates of linked advertisement.

To achieve this, we will study properties of recommendation lists and their impact on content propagation. Several studies have showed that there exists a strong correlation of view count of a content and the average view count of the videos in its recommendation list (Zhou et al., 2010; Cheng et al., 2008). Indeed, from different measurement studies, the view count of videos in a recommendation list of a given video tends to match the view count of that video. Cheng et al. have provided different statistics of YouTube and showed that the graph of YouTube's video structure exhibits a small-world characteristic (Cheng et al., 2008).

Zhou et al. have showed the importance of the recommendation system on view count of a video (Zhou et al., 2010). They found that the recommendation system is the source for 40-60% views of a video. Performing several measurement, they have discovered that the position of a video on a related video plays an important role in the click through rate of the video. They have also identified that the recommendation system improves the diversity of video views which helps a user to discover unpopular videos.

Our work compliments these works by quantifying the relationship between a video's view and its related videos (i.e. the videos in its recommendation list). To that end we focus on users that browse through videos according to some random mobility model over the recommendation graph. In this directed graph, videos are nodes, and directed edges connect that node to the videos that appear in its recommendation list. Nodes have some attributes, or weights that may correspond to the view count, or to the total time that the video was viewed, or to the number of likes etc.

We first describe the model associated the recommendation system in section 4.2. Our first goal is to relate the attributes of a node to those of its recommendation graph. In section 4.3, we first describe our dataset. We then study empirical properties of the sequence of weights in a random trajectory of a user in the recommendation model as a function of their mobility model. In particular, we focus on two attributes. The first is the number view count of the video, and the second is its age. To each one of these attributes and every mobility model, the sequence of consecutive attributes on the random trajectory forms a stochastic process which we model as a Markov chain and we study its stability in section 4.4. We then derive properties related to the stability of the sequence of videos viewed from the stability properties of the Markov chain corresponding to the attribute process. Finally in section 4.5, we provide a theoretical result on the improvement by the recommendation system of the diversity of video views. Indeed, recommender system in YouTube is traditionally based on keyword between videos (tags, title and summary).

4.2 A model for YouTube recommendation system

We now provide some background on YouTube which has become a key international platform for socially enabled media diffusion. This platform allows not only to share videos, but also to create interaction between users (friends, creators, rating..). It becomes a most attractive and a popular media diffusion with a huge quantity of user-generated content. Our study on recommendation system in YouTube is based on the data sets crawled from YouTube. Here we describe how we collected the data sets.

4.2.1 Data on videos

A huge majority of YouTube videos are available to the general public and includes valuables data about the video (a video may not be available if its creator decided to attach to it the unlisted or the private privacy level or if there are copyright issues that do not allow to show the video at a given country). YouTube further proposes a list of recommendation from which the consumer can choose the next video to watch. We collected data sets of videos using our own software developed in JAVA. This tool allows us to collect some view statistic of videos in YouTube as views, titles, tags, ages and recommendation list. A Linear Least Squares Regression (LLSR) is used to adjust the model parameter in order to obtain the minimum error between the model and experimental data.

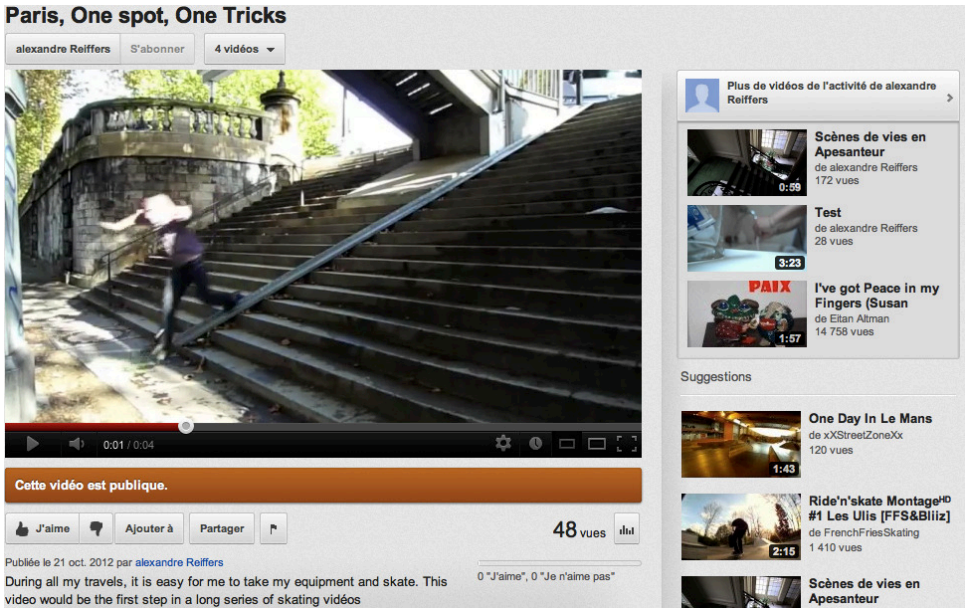


Figure 4.1: *Recommendation list in YouTube*

4.2.2 View graph

In this section we construct a graph based on YouTube recommendation. In particular, we will explore how the view of a video influences the view count of other videos through the recommendation list. Indeed, a user who views a video u may view a video v from its recommendation list. In that case, there is a directed edge between u and v (See figure 4.1).

Let us first introduce some terminology. We consider a connected network $G = (V, E)$, where V denotes the set of nodes with $|V| = n$, E denotes the set of edges and $w : V \to \mathbb{R}_+$ denotes the view count of the node v. We will now describe how we construct our connected graph. We imagine a random walker on YouTube starting from a video u. After viewing the video u, he selects randomly (with uniform distribution) a video v among top N videos in its recommendation list, and moves to its neighbour. We repeat the procedure again and again. A stochastic (transition) matrix $Q = [Q_{uv}]_{n\times n}$ is used to govern the transition of the random walk process where n is the number of videos in the graph. Q_{uv} is the probability that the transition from video u to video v occurs.

4.3 Statistical Study of the recommendation graphs

In this section, we explain how we obtain the influence of recommendation system using our data sets in YouTube. We will study two datasets. In the first one we randomly picked 1000

videos, in the month of july 2012 in YouTube by using our random extraction software. In the second one we crawl YouTube to get *NBVIDEOS* videos. We focus on two elements of the data. The first element is the view count of a video and the second element is the average view count of related videos recommended by YouTube recommendation system. We believe that the number of videos that we collected for each experiment is enough to capture all information. First, we investigate the relation between the view count of a video and the average view count of its recommendation list. We consider different values of the number N of videos in the list. In practice, the larger screen is, the larger is N. We shall show later that N plays a crucial role on the properties of the excursions over the recommendation graph. A very small N corresponds to list of recommendations viewed over cellular telephones.

In our statistical study, we test two types of models that relate a function of the number x_i of views of a video i and the function of views of videos in its recommendation list averaged over N:

- N-videos in which we use the linear regression between x_i and y_i, $i \in \{1,\ldots,I\}$, wher $I \in \{1000, NBVIDEOS\}$. Here,
$$y_i = \frac{1}{N}\sum_{j=1}^{N} y_i^j$$
where y_i^j is the view count of the j^{th} video in the recommendation list of video i. We thus identify parameters a and b so that y_i will be well approximated by $ax_i + b$. a and b are chosen in fact so as to minimize the mean square error of this approximation over all samples.

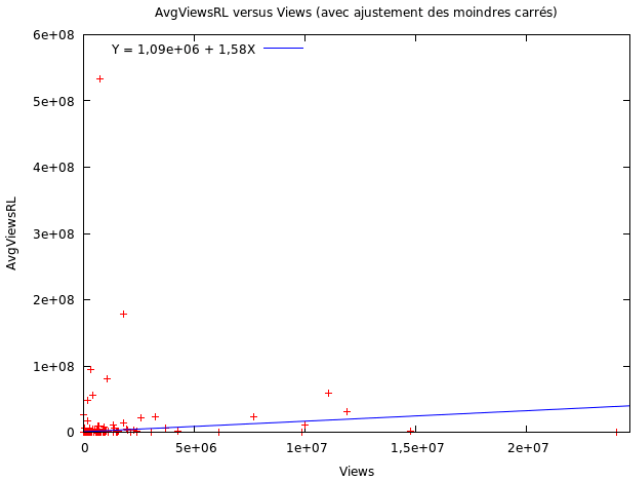
- N-videos wherein a linear regression is used between $log(x_i)$ and $log(y_i)$. We thus identify parameters a and b so that
$$\frac{1}{N}\sum_{j=1}^{N} \log(y_i^j)$$
will be well approximated by $a\log(x_i) + b$. a and b are chosen in fact so as to minimize the mean square error of the difference between the expressions over all samples.

4.3.1 Study of the data set

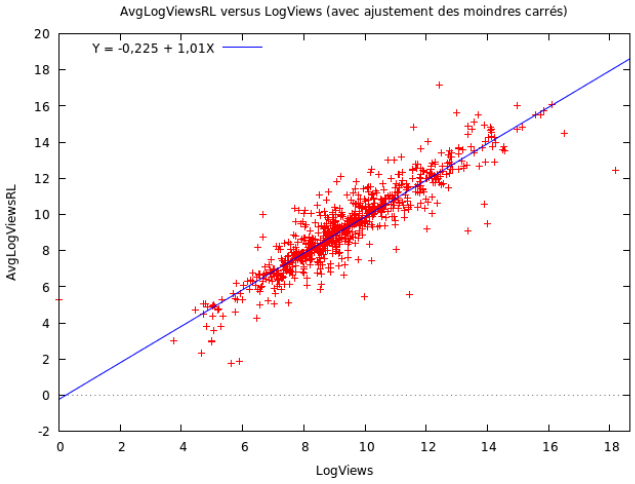
The distribution of the dataset according to age and to popularity

The measure of goodness of the linear fit is summarized by the statistical R^2 coefficient. It takes value in the unit interval $[0,1]$ and the better the fit is, the larger is its value.

(a) N=1



(b) N=3



(c) N=5

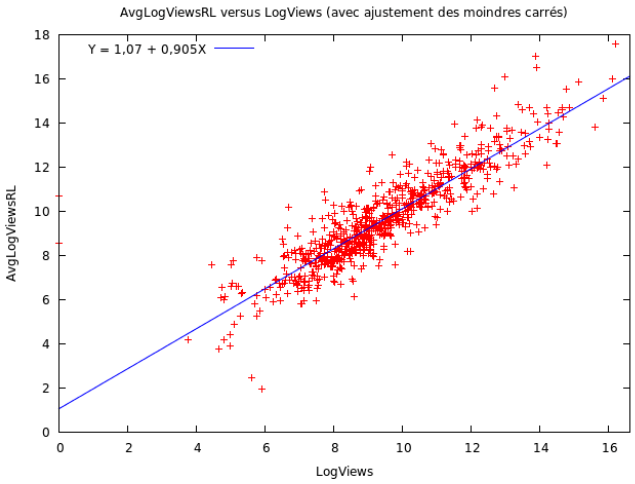


Figure 4.2: *The view count of one video on the X-axis, and the average of the logarithmic views of the top* $N = 1,2,3$ *of its recommendation list*

In figure 4.2a, we plot the view count of one video on the X-axis, and the average view count of videos in its recommendation list. We observe that the coefficient of determination R-square is small for all values of N (see fig. 4.4a) . This indicates that there are important deviations from any approximation of the y_i as a linear function of the x_i.

On the other hand, in figures 4.2b-4.2c, we use a logarithmic scale to plot the view count of one video on the X-axis, and the averaged of the view count of videos in its recommendation list. Both figures 4.2b-4.2c show the strong correlation between the view count of a video and the view count of top N videos in its recommendation system.

The first observation is on the monotonicity: the higher the the view count of the related videos in recommendation list of a given video is, the higher the view count of that video is. Hence, YouTube recommendation system will prefer to put videos in recommendation list based on the popularity of the current video, located in the same region of the popularity.

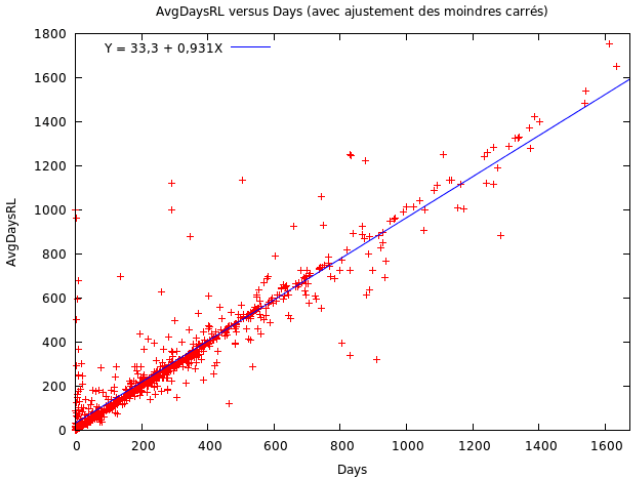
Table 1 shows the quality of the linear regression in the logarithm scaling. The table fits a set of data for several values of N. Our experimentation thus correspond to random walkers that after watching a video, it chooses the next one to see among its N recommended ones with equal probability. This can model space limitation on the screen or in the recommendation list which limit the number of recommendations seen by the random walker. It can also model user behaviour in which the top recommendations are preferred. Indeed, in was shown in (goo, 2009) the first position in the recommendation list attracts 39 times more clicks than the 10^{th} position.

Now we further investigate the correlation between the age of a content and the average of age of videos in its recommendation list. Figures 4.3a-4.3b show clearly the trend that the higher the average of age of the related videos, the higher the age of the video. This implies that YouTube recommendation will prefer to recommend the videos located in the same region of the age. This means that YouTube's recommendation will not significantly affect the overall videos based on the age of a video and will instead focus more on the same generation of that video. Moreover, Table 4.4b displays a high regression coefficient which is the additional evidence that the recommendation system has an important impact on the view count and a popular video can affect only the videos located in the same region of the popularity and the age.

4.4 Stability and video view diversity

We shall study in this section the stochastic process X_n, $n \geq 1$ of the attributes of the videos encountered by the random walker. For example, X_n can be taken to be the view count or the age of the nth visited video. For simplicity we shall assume that this constitutes a discrete time aperiodic communicating Markov chain (any state can be reached from any other state with

(a) N=3



(b) N=10

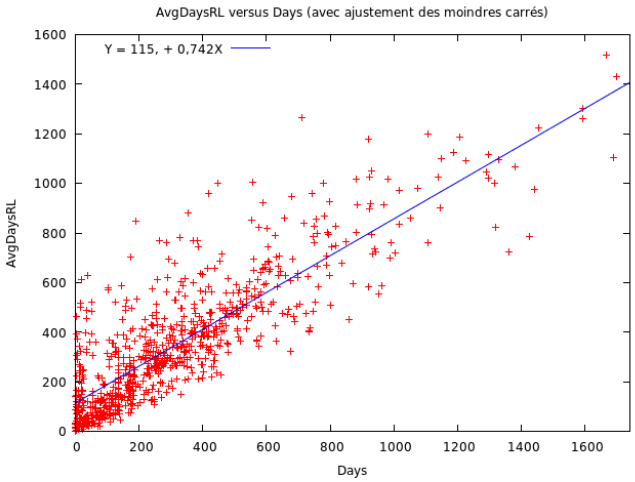


Figure 4.3: *The age of one video on the X-axis, and the average age of the top* $N = 3, 15$ *of its recommendation list*

Table 1	regression coefficients	R^2
$N = 1$ using logarithmic scale	1.03932	0.746919
$N = 1$ using linear scale	0.658874	0.025083
$N = 2$ using logarithmic scale	0.986524	0.987252
$N = 2$ using linear scale	2.73843	0.773921
$N = 3$ using logarithmic scale	1.00942	0.825262
$N = 3$ using linear scale	0.114771	0.005682
$N = 4$ using logarithmic scale	1.01938	0.816736
$N = 4$ using linear scale	0.858309	0.034767
$N = 5$ using logarithmic scale	0.905340	0.791218
$N = 5$ using linear scale	2.73843	0.221205
$N = 8$ using logarithmic scale	0.587379	0.335723
$N = 8$ using linear scale	0.798919	0.176054
$N = 10$ using logarithmic scale	0.529770	0.277402
$N = 10$ using linear scale	2.99831	0.195627
$N = 15$ using logarithmic scale	0.500745	0.258681
$N = 15$ using linear scale	1.21020	0.069401

Table 2	regression coefficients	R2
1 videos days	0.976898	0.918683
2 videos days	0.932103	0.882182
3 videos days	0.931216	0.879587
4 videos days	0.906025	0.870077
5 videos days	0.906025	0.870077
8 videos days	0.780917	0.776079
10 videos days	0.742248	0.713315
15 videos days	0.723398	0.685777

Figure 4.4: *Regression coefficients and coefficient of determination R-square for different values of N for the age study*

positive probability), although this assumption will be relaxed later. We investigate the stability of this Markov chain defined above in order to identify the impact of YouTube recommendation system on the visibility and popularity of videos.

We briefly recall some basic notions in Markov chains. The chain is said to be recurrent if it visits every state infinitely often. It is said to be positive recurrent if for each state x, the expected time between consecutive visits of x is finite. It is said to be transient if for some state x there is a strictly positive probability never to return back to that state. Thus in a transient Markov chain, a state x is only visited finitely often, and after some finite (random) time it never returns to x again. We shall say that X_n is stable if it is positive recurrent and that it is unstable if it is not.

We argue that a recommendation system tought to be designed with the objective of having X_n communicating and positive recurrent for various reasons. First, it would allow a random walker to return to a video he liked. Secondly, it does not get stuck in a subset of states due to the communicating assumption. The fact that one reaches a state infinitely often guarantees in particular that any video will be discovered by any random walker that stays long enough. Finally, this property has a positive impact on increasing the page rank of the video and thus search engines based on page rank will allow one to find the video quicker.

We next recall the Foster stability conditions for communicating Markov chains. Introduce a positive and increasing function f and a finite set B. Define the drift of the chain at state x as

$$\Delta(x) := E[f(X_{n+1}) - f(X_n) \mid X_n = x]$$

Note that due to the Markov property this does not depend on n. If is called a Lyapunov function.

Lemma 1. *(i) If* $\sup_x \Delta(x)$ *is finite and for some* $\varepsilon > 0$, $\Delta(x) < -\varepsilon$ *for all* $x \notin B$ *then the Markov chain is stable.*
(ii) If $\Delta(x) > \varepsilon$ *for all* $x \notin B$ *then the Markov chain is unstable. A sufficient condition for instability is that*

$$E[f(X_{n+1}) \mid X_n = x] \geq af(x) + b$$

for $b > 0$ *and* $a \geq 1$.

From different experiments (see for example Table 4.4a), we observe that for $N = \{3,4,5\}$ and for all $n \in \mathbb{N}$ we have

$$E[\log(X_{n+1}) \mid X_n = x] < \infty, \text{ and}$$

$$E[\log(X_{n+1}) - \log(X_n) \mid X_n = x] > \varepsilon > 0 \tag{4.1}$$

for some $\varepsilon > 0$ and for all x large enough (larger than some constant K). It follows from Theorem 3 in (Foss, 2008), that the Markov chain cannot be positive recurrent.

4.5 Discussion

The instability of the Markov chain X_n for $N = \{3,4,5\}$ (which is the case if the computer's screen is small) has several interpretations and consequences. Indeed, by the rapid adoption of smartphone, tablets and e-reader which are characterized by a small screen, our result may explain some results in (Zhou et al., 2010) which showed how YouTube recommendation can improve the view diversity and help users to discover videos.

The process X_n which we studied in this chapter can be interpreted as a random walk on the recommendation graph, where X_n corresponds to the number of views of the nth viewed video. The $n+1$ video viewed is chosen uniformly from the recommended videos of the previously viewed video. We identified in this chapter instability of the process X_n as a function of N (size of recommendation graph). The instability was obtained by establishing (through experiments) that the related drift is positive.

How sensitive are the results to the Markovian assumption on X_n? Note (by taking expectations in (4.1)) that as $n \to \infty$, $E[f(X_n)]$ tends to infinity. This type of instability result does not require X_n to be a Markov chain. Some of the stability results of X_n do hold even when X_n is not a Markov chain (Borovkov and Yurinsky, 1998).

Note that both for stability as well as for instability, the drift condition is required to hold for all large enough states, i.e. for all $x > K$ for some positive K. Instability of the process means that the number of views of the nth video tends to grow without bound. In practice however, there is some finite bound on the number of viewed videos (the bound is the number of views of the most viewed video). Thus, what instability means *in practice* is that the number of views grows until it is in some neighbourhood of that bound. For this to occur, we expect that it is sufficient to establish that the drift is bounded below by a positive number within some large integral $x \in [K_1, K_2]$.

//

Chapter 5

Development of tools to analyse social networks

5.1 Introduction

Internet is actually an increasing platform to communicate between machines and people. This is true especially since the introduction of new functionalities offered by the smartphones, tablets, etc. The understanding of their evolution and their influence over the social networks need leads us to analyse their dynamics, especially in the advertising and marketing sectors. Socio scientists are in general interested in the content generated by users; especially if these contents are available as public data (e.g. The Twitter stream API provides a 1% sample of the complete public tweets).

Online social network: there are online platforms in which people build social relations with other people who share similar personal topics, career interests, activities, backgrounds, real-life connections, etc. The online social networks allow all this sharing opinions, music, news, pictures, etc.

Because of the increasingly importance of data processing of this huge increasing data, it is necessary to develop tools able to collect and analyse the data. For the development of these aforementioned tools we first need to understand the structure of the platforms were the data is processed and how it is stored.

From the informatics point of interest, one way to better understand these complex networks is based in the building of software using API's (Application Programming Interface). The API's have the advantage that they can be set as a subroutine of syntaxes in the background as

communication protocol and a tool for further building a sustained software. API's uses a set of Library/functions for collecting data from the user, a useful feature even without knowing the API background.

Another way to collect data from the online social networks is using the WEBSCRAPING technique, which allows the extraction of information from websites through programs that simulate the human navigation using the http protocol. This technique is used in the process called ETL "Extract, Transform and Load" to extract data from the web. This technique allows us to analyse some online social networks.

There are many social networks and studying all of them is out of the scope of this dissertation, so we decided to focus on two important ones: YouTube and Twitter.

Origins and importance of Twitter and YouTube.

YouTube headquartered in California, US, is a video-sharing website were users can upload, share and view content. It was created by three former PayPal employees in February 2005 and bough by Google in November 2006 for US$1.65 billion; YouTube now operates as one of Google's subsidiaries.

Twitter based in California, US was created in March 2006 by programmers who worked at the podcasting company Odeo Inc. in San Francisco. Twitter is an online news and social networking service on which users post messages known as "tweets". Tweets were restricted to 140 characters, but on November 2007 this limits was doubled for all languages except Chinese, Japanese, and Korean.

According to the statistics shown in figure 5.1 in October 2018 on the site (www.statista.com, 2018) YouTube was the second social network in terms of persons with 1900 millions of registered users and Twitter was located in the 11th place with 885 million people registered. To have an idea of the important amount of information that is generated we can highlight the following data on (www.internetlivestats.com, 2018). Every second, on average, around 6,000 tweets are published on Twitter, which corresponds to over 350,000 tweets per minute, 500 million tweets per day, and around 200 billion tweets per year. In YouTube we have on average, around 70 thousand views per second, 4.2 million per minute, 6 billion per day and 1600 billion per year.

We chose the Figure 5.1 from Statista that shows the most popular social networks, but according to the concept we manage of social network we don't consider social networks tools like Skype, WhatsApp, Viber, etc. which appear in the figure.

Social networks such as YouTube and Twitter, represent an important broadcast media. They are used today to spread political messages, news, protests and even revolutions. Some of these uses will be analysed in the sections described in this chapter.

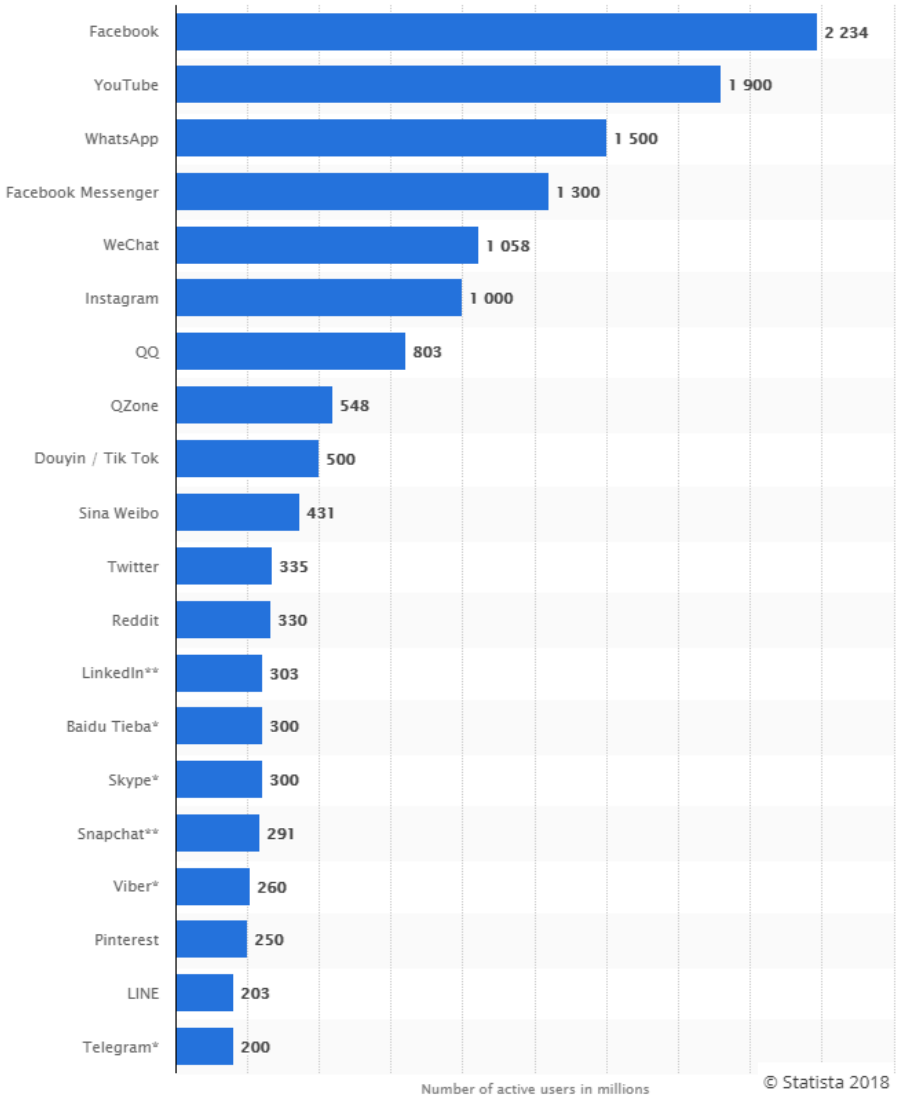


Figure 5.1: *Most popular social networks worldwide as of October 2018, ranked by number of active users (in millions)*

5.2 Objectives and topics

Discovering better methods to find, share and disseminate information on social networks is one of the objectives currently pursued by many people. And to achieve these goals we have developed tools to analyse large amount of data.

The general objective pursued here is the analysis and identification of the roles and dynamics of the social networks Twitter and Youtube. The goal is to create an or several computer tools

capable of analysing social networks, to make geographic analyses, suggest prediction models, and create popularity indicators.

First, to begin the analysis of a social network dynamics and roles we chosen Twitter because this tool allows to access data via API which provides methods for building software applications.

The work includes the geographic automatic extraction of information analysis of the tweets. We proposed a method to infer the geographical location of a user according their habits using the most commonly words in the language used in its surrounding. As a result of our analysis we presented and published a communication for the ASONAM conference in August 2012. ("Geo-linguistic fingerprint and the Evolution of languages in Twitter") (Altman and Portilla, 2012).

Secondly, we worked with various types of data on YouTube:

We started with the recommendation list of a video to see if there are specific behaviours that can be defined mathematically.

We propose six epidemic models that classify the behaviour of a large majority of videos. From a sample of more than 80.000.000 videos plus of a 90% of them fits one of the six propagation models with a mean square error (MSE) no more that 5% (Richier et al., 2014).

Additionally, we compared the evolution of memes on different social networks. We describe this work in section 5.6.

5.3 Creating Datasets of Twitter messages for geographic analysis.

Before talking about the methodology carried out we will talk about Twitter terminology,

A tweet is a message posted in Twitter.

Vocabulary and symbols associated with Twitter. The @, #, RT, and DM descriptors.

The "@" (at) attached to the nickname of a Twitter account it's used to inform that the recipient message will be sent to him. For example if you type " Hello @USER", the message will appear on the home page of this USER.

The "#" (hashtag) followed by a word (avoid accents, spaces and other special characters) operates as a keyword allowing define the main topic of a message written making the tweets be more visible.

A message with the descriptor "RT" (Retweet) is a message published by a person making use of the message published by another person. The message is written as follows: "RT

@user_who_wrote_the_message + original message".

A Direct Message (DM) is a message sent directly to a person. A DM is not public. It can be interpreted as an internal email in Twitter. We can send DM to another person only if that person follows you and vice versa.

After having a broad general view of the principal concepts of Twitter, we proceed to describe the work performed with it:

First, we analyse graphically the platform with a representation of the frequency of words used as function of the time of the day. We started this analysis using the tools Trendistic and Topsy (mentioned below in section 5.4.1 and 5.4.2) but later we stopped using them because the web pages were put offline, after this inconvenient we realized that we had to develop a method that allowed us to see the data in our particular way. By using the Twitter streaming API, we developed a tool called TwitterStreamDownloader (TSD) that automatically downloads 1% of the messages that are written on Twitter's platform. The data downloaded with TSD occupies approximately 4 GB on disk daily. Approximately 10% of the data automatically collected has the geographical position defined in its structure. Automatically downloaded data may contain duplicates, and by consequence they require post filtering to eliminate them. We develop a tool called CannesInsideScrapping (CIS) that uses web-scraping techniques to analyse messages at the pages called CannesInside during the Cannes festival promoted by Vodkaster (`https://www.vodkaster.com`), a social network used to film micro critics. Other work and analysis have been done on platforms such Google Trends and others that are mentioned below in sections 5.4.1, 5.4.2 and 5.6. We can summarize our tool development as follows:

i. Data collection either using Web-Scraping or API

ii. Transformation of the data either into CSV or JSON files or databases

iii. Statistic analysis

iv. Graphic representation

5.4 Analysis of Twitter Data

Before worked with TSD we used a cURL tool for access to messages posted on Twitter. A cURL is defined by Wikipedia as " a computer software project providing a library and command-line tool for transferring data using various protocols".

The use of the Twitter Streaming API with cURL was not stable, and data extraction was stopped after one day or two.

The quantity of data that can be downloaded from Twitter was further increased when we decided to develop our own automatic download tool rather than using the cURL approach. The TSD was written in Java and allows to automatically downloading the message flow. With TSD was obtaining an increase of more than a week of data downloaded. Due to the large volume of tweets downloaded, statistics analysis of them was takes a long time. It became therefore advantageous to organise the data in smaller files containing the tweets in JSON (JavaScript Object Notation) format. Unfortunately the Twitter API analysis of tweets relies knowledge of the geographic position, which was available only in 10% of them.

We started organizing the messages downloaded by date in a single file, but in the latest version of TSW they were grouped by hours in two different blocks, a file containing messages with geographical information (corresponding to 10% according to the API description) and another without the geographical information (90%). As a result of our analysis we presented and published "An Author-Topic based Approach to Cluster Tweets and Mine their Location" in Spatial Statistics: Emerging Patterns 2015 and "A Topic Modeling Based Representation to Detect Tweet Locations. Example of the Event "JE SUIS CHARLIE"" in ISPRS Archives, ISPRS Geospatial Week 2015.

5.4.1 Trendistic analyses

Trendistic (www.trendistic.com) was a server online until 2012. This tool helps us to analyse, and find the source of trending topics on Twitter. With Trendistic a simple search could be graphed presenting the popularity of any keyword on Twitter.

This tool uses trends lines to display periods ranging from 24 hours to 180 days, comparing several words in the same query. Trendistic had predefined values of time periods for the analysis and also offers options to select the tweets that contained a defined term from the query (ex: Figure 5.2).

Figure 5.2: *Topsy screen*

We extend the Trendistic functionality allowing periods of time greater than the 180 days predefined, by the software.

We found the way to customize this query by setting the display time from 180 days up to 365 days. We also define the display by day or by hour.

5.4.2 Analyst of the Topsy website with API.

By the time the work was developed the Trendistic website was offline in 2012, because that, we had to find another tool to replace it. After investigating and comparing different options, we found Topsy a webtool that have the functionalities loses in Trendistic.

Topsy is a search engine that stores a comprehensive index of Tweets — as its site says, "all Tweets since 2006." A Twitter certified partner, Topsy behaves a lot like Google Search for Twitter

Topsy allows us to search around a large amount of shared content on the network: web pages, photos, videos and also simple messages on the Twitter network and Google+.

Topsy have advanced search options than comparing Trendistic. These options allowed us to analyse in a better way our data, because we have access at values of the graphic representation.

With this new tool we find the real values associated with the trends and not only the graphical information associated at it, as in Trendistic.

The graphic representation was visualized on Topsy as shown in the Figure 5.3:

Figure 5.3: *Topsy Histogram*

Fig 5.3 shows the graphical representation that obtained with the API offered by Topsy. The API includes several functionalities, but for this work we use the function described by the searchhistogram: "https://code.google.com/p/otterapi/wiki/Resources#/searchhistogram". The result of our analysis is presented in chapter 2 and published in "Geo-linguistic fingerprint

and the evolution of languages in Twitter" in Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference.

5.4.3 Creating Twitter datasets with the Vodkaster platform for the Cannes Film Festival 2012 data

The analysis of the web structure with Vodkaster and the way as this website organize the data was utilized to collect the necessary parameters to retrieve information associated with 32 users profiles associated to Vodkaster called as Insiders and Messages. For each Cannes Festival role one or more profiles are associated in order to post news, images, and opinions for their respective meeting spaces. After the above each profile was analysed in order to infer and associate the Vodkaster messages with the dataset of tweets tagged in association with the Cannes Festival. . As a result of our analysis we presented and published "Les usages de Twitter au Festival de Cannes. Pratiques numériques et construction des opinions esthétiques" . in Culture et musées Nº24 (décembre 2014)

5.4.4 Creation of a tool to analyse the structure of messages on Twitter. The TweetAnaliser tool

There are two types of databases used along the thesis: relational (A database organized in terms of ordered list of elements called tuples, grouped into relations) and object oriented (a database in which information is represented in the form of objects as used in object-oriented programming. Object databases are different from relational databases which are table-oriented. Object-relational databases are a hybrid of both approaches.).

To develop this tool previously an in deep analysis of Twitter message's structure was necessary to be made. The reason of this previous analysis is because hidden behind a publish message of 140 characters there is a meta-data that describes it and all their elements.

TweetAnaliser was developed in a Java environment with a relational database.

TweetAnaliser was able to make the analysis of the popularity of a word assembling the results according with its geographical apparition; this classification is displayed in a map generated by GoogleCharts according the options menu provided by the software. TweetAnaliser managed to accomplish several objectives defined at the beginning of the chapter, but the amount of processed data was insufficient. The cause of this issue is the lengthy time to process a query containing more than 20 million messages.

5.5 Analysis of YouTube Data

One of cores of this work is the data collection using the YouTube APIs (version 2 and 3), which allow collecting data associated with a video. We used YouTube API version 3 in Java with the available data from API. These APIs are developed in different programming languages as: Apps Script, Go, Java, JavaScript, .NET, PHP, Python, and Ruby.

Since the integration of Google and YouTube is actually possible by log in YouTube using an existing Google account we proceeded to use the YouTube's API. To use YouTube's API, we must acquire first a "Developer ID". This is an additional YouTube property available for the developer.

Below is presented the sequence of steps followed to obtain an API Key.

i. Go to https://developers.google.com/ and log in (or create an account, if necessary).

ii. After logging in go to the link https://console.developers.google.com/project and click on the blue button CREATE PROJECT. Wait a moment as Google prepares your project.

iii. YouTube will ask you for a project name.

iv. Then click GoogleAPIs link in the top left corner and after click the link option "YouTube Data API". Which is under YouTube API's.

v. Now click on the "ENABLE" button.

vi. Next click on the 'Go to Credentials' blue button on the right.

vii. Choose the select option YouTube Data API v3 for the first select option and Web server for the second selection. Then choose Public data. Now click the blue button, "What credentials do I need?."

viii. Almost done, wait for Google to create your new project and you should see the screen where you can copy your API Key.

ix. Paste the API Key in our YouTube project.

We started with the API version which allow to find the data of a video following advanced searching criteria's. With this API library uploaded we search for videos without a keyword (this option is not available directly on YouTube website).

Others important information's, such as the description of the video, category, views, date of publication, list of recommendations of a video, etc., allows us to discover the "secrets" behind

the platform studied used for the dissemination of videos.

The data obtained in YouTubeData API 2 allowed us analyse the dynamics of number of videos viewers but did not provide enough features beyond that. Because that we decided to explore YouTubeData API 3 an posterior version that allowed to obtain the videos according to the date of publication, thus facilitating a broader view of the time period we were interested.

Unfortunately some data cannot be collected using the API since it is presented only in graphical form showing the history of the video's views, their statistics, time of visualization, etc. The most interesting is that this data screen visualization provides useful information of the evolution of a video in cumulative options. The screen was accessible by using some tools like cURL (this is a command-line tool for getting or sending files using URL syntax) making a call to the address `http://www.YouTube.com/insight_ajax?action_get_statistics_and_data=1&v=VideoID`.

All the methods described above allowed us to collect the data needed for later analysis, which were manipulated mainly in two ways: 1) CSV files, used to save basic data structures and 2) JSON files used for the complex structures.

5.5.1 Analysing recommendations lists of a YouTube video through the API

In our work we used some functions offered in the YouTube API implemented with a self-created Java application that generates CSV files (comma separated values files).

We did some preliminary analysis at the beginning of our search, but for each query we only had access to 1,000 elements, and this amount of data was not statistically significant compared to the number of videos published online by the website.

Updates on the website and new API versions beyond version 2 have reduced by half the output results (500 elements per query) and forced us to look for new solutions, because less that 1000 elements per day was not sufficient for our needs.

We have created a database to analyse the recommendations lists of YouTube videos. This tool was developed with Java based on the two previous versions of the API. The YouTube API 3 allowed us to go from 500 items to more than 1 million elements resulting from the use of a new method of search on the API. The result of our analysis is presented in chapter 4 and published in "A Study of YouTube recommendation graph based on measurements and stochastic tools" in IEEE/ACM International Conference On Utility and Cloud Computing (UCC-2015).

5.5.2 Analysis of YouTube website statistics

When we used the API for the YouTube statistics we found limits, one of these is that only some elements associated with the video were available, for this reason we did an analysis of the HTML content generated by the YouTube website.

Analysis using Google Charts: In this approach, we analysed the structure of the generated images corresponding to a Line Chart into Google Charts, these imagens followed an specific structure; with several transformations it was possible to retrieve the screen values associated with the graphics of each video (these values were restricted to 100 values of the evolution per video).

A new method different to the implemented using Google Charts was developed through the updating of the web interface. Because in API all elements associated with video are not available as crude data, then we also made an analysis of the HTML content generated by YouTube web site.

Analysis using a web service: This method is also obtained by analysing how the YouTube website works after the updating of the web interface; we have been able to retrieve the statistics necessary presented in graphical form with the video. The good thing with this method is that these screen data show us the evolution of the video.

At the beginning we only have access at the views of a video normalized, later we access at video views, subscribers, shares and watch-time showed from his uploading date until the day before the consultation. We develop the tool to collect the data in Java but due to changes in access mode later we develop our tool with Python.

5.6 Analysing statistics obtained in Topsy, YouTube and Google Trends

We compare the views evolution from YouTube videos and we search the keywords associated at with this video with Topsy and GoogleTrends. This is reported in chap 3 of the thesis. The data from YouTube was extracted with the method described in the section 5.5.2. The data was extracted with keywords in the Topsy search engine described in section 5.4.2 and the data was extracted associated with queries made with GoogleTrends. We performed a manual analysis for processing the results and compared the trends associated with a word to discover possible similarities with the statistics obtained in the different platforms. As a result of our analysis we presented and published “Social Networks: a Cradle of Globalized Culture in the Mediterranean Region” in the 2nd World Congress on Computer Applications and Information Systems (WCCAIS’2015).

Chapter 6

Conclusions and Future Work

Designing and improving tools for social networks analysis becomes crucial, today social networks have a tremendous impact on our culture, in business, etc. Social media websites are some of the most popular places for socializing on the internet. Understanding how the social networks evolve is a core of this work. In this thesis we proposed three approaches for analyzing social networks and their impact on society. The first is the analysis of most used words in English, French, Spanish, Portuguese, Italian and German and proposed a way to get the geolocation of messages based in their use. In the second approach we study the impact of memes in Mediterranean Region and in third we analyse the recommendation list of videos in YouTube based on measurements and stochastic tools.

In Chapter 2 we introduced a method that allows performing some rough geo-localisation using the periodogram of Twitter. This is applied to identifying and understanding some of the transformations in spelling and in the use of words over Twitter. This includes a geo-linguistic analysis that allows one to track different types of transformations in different communities that have the same language in common. Among the many examples presented here, we have studied in more detailed the transformations of the word "because", both in English as well as in Spanish. We saw that some common form of using two different spellings in the same sentence has emerged both in Twinglish as well as in Twitter-Spanish. We managed to differentiate between the locations of various versions of "because" in English. We then showed how the geo-localisation can be used to get insight on political events.

We note that there are other ways of obtaining geo-localisation of messages as well as the identification of language in which they are written, based on information that are available in some of tweets. Since only a small fraction of tweets are geo-localized and since we do not have figures on the reliability of the language used, we decided to focus our work on a methodology

that can be widely used relying on the trending APIs. We used the "trendistic" API which was publicly available on the Internet at http://trendistic.indextank.com/ when we started this work. Later on this API was removed. This chapter reports on results obtained using programs based on that API. Since the API is no more available, our results may not be reproducible. Yet results of experimenting with these tools may be valuable precisely because they are not available anymore.

We have shown that displaying one's geographic community has an important role the evolution of languages. We plan in the future to introduce this insight into mathematical models of evolution of languages and in particular in refining the definition of fitness function in existing models based on evolutionary game theory (Trapa and Nowak, 2000; Christina Pawlowitsch, 2011). In Chapter 3 we described some recent Internet Memes that have traversed the borders of countries and/or languages and have had a social/political impact on the Mediterranean Region. They contribute to the creation of a globalized culture that has no borders and which has a potential in bringing closer together different nations. In Chapter 4 We have shown in this chapter through measurements the relation between the number of views of a video and the number of views of the videos in its recommendation list averaged over the first N slots in this recommendation list. We considered not only relations between the numbers of videos viewed but also between functions of these numbers. More precisely, we established the relation between a function of the number of views of a video and the function averaged over the number of views of videos in the recommendation list of the video.

We show good fits of a linear relation between the number of views of a video and the number of views averaged over its recommended videos when considering the log-log scale. The fit was not good when considering a linear scale.

Based on this relationship we explored the evolution on number of views that a random walker sees when travelling through recommendation graph. We showed in particular that if the number of videos in the list is small, the number of views tends to increase along the trajectory of random walker. We conclude that the random trajectory is not sable which means that the trajectory doesn't not contain cycles.

For showing instability, we used Foster Criteria where we approximated conditional expectations by conditional averages. Note that Foster Criteria is valid independently of the value of the R^2 parameter.

This study was restricted to YouTube and based on a dataset that reflects YouTube's protocol at a given time (july 2012). However the methodology is valid for other recommendation systems as well or for newer versions of YouTube recommendation.

6.1 Future works

Currently most of tools developed throughout the thesis are outdated and new API's versions are available, for which new tools are needed. In order to study the evolution of social networks these tools need to be developed. An important point in the analysis of geo-location in Twitter will be to improve the analysis of opinion in micro-regions (changing from the analysis of countries to the analysis of smaller regions). Another important point is to compare and analyse tweets of 140 characters with those of 280 and see if the result of the analysis depends on the size of the tweet. It will also be interesting to analyse the tweets by taking others attributes in tweets metadata other than time and geo-location.

List of Publications

- Modelling View-count Dynamics in YouTube. Cedric Richier, Eitan Altman, Rachid El Azouzi, Tania Altman, Georges Linares, Yonathan Portilla. ArXiv 2014
- An Author-Topic based Approach to Cluster Tweets and Mine their Location. M. Morchid, Y. Portilla, D. Josselin, R. Dufour, E. Altman, M. El-Beze, J-V. Cossu, G. Linares, A. Reiffers-Masson. Spatial Statistics: Emerging Patterns 2015
- A Study of YouTube recommendation graph based on measurements and stochastic tools. Yonathan Portilla, Alexandre Reiffers, Eitan Altman, Rachid El-Azouzi, IEEE/ACM International Conference On Utility and Cloud Computing (UCC-2015)
- A Topic Modeling Based Representation to Detect Tweet Locations. Example of the Event "JE SUIS CHARLIE". M. Morchid, D. Josselin, Y. Portilla,R. Dufour, E. Altman, and G. Linares. ISPRS Archives, ISPRS Geospatial Week 2015.
- Social Networks: A Cradle of Globalized Culture in the Mediterranean Region. Eitan Altman, Yonathan Portilla. 2nd World Congress on Computer Applications and Information Systems (WCCAIS'2015).
- Bio-Inspired Models for Characterizing YouTube Viewcount, Cedric Richier, Eitan Altman, Rachid El-Azouzi, Tania Jimenez, Georges Linares, Yonathan Portilla, Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference
- Geo-linguistic fingerprint and the evolution of languages in Twitter. Eitan Altman and Yonathan Portilla. Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference
- Dynamiques des popularités dans YouTube.Cedric Richier, Georges Linares, Rachid El Azouzi, Tania Jiménez, Eitan Altman, Yonathan Portilla. CORIA2015

- Les usages de Twitter au Festival de Cannes. Pratiques numériques et construction des opinions esthétiques. Myriam Dougados, Jean-Louis Fabiani, Julien Gaillard, Yonathan Portilla. Culture et musées Nº24 (décembre 2014) : Démocratisation culturelle et numérique

List of Figures

List of Tables

Bibliography

(1989). *The Selfish Gene*. 33

(2009). Google adwords click through rates per position, http://www.accuracast.com/seo-weekly/adwords-clickthrough.php. 50

Altman, E. and Portilla, Y. (2012). Geo-linguistic fingerprint and the evolution of languages in twitter. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Istanbul. 10, 13, 15

Altman, E. and Portilla, Y. (2013). Les nouveaux réseaux: Berceaux de culture mondialisée en méditerranée. 34

Altman, E. and Portilla, Y. (2015). Social networks: A cradle of globalized culture in the mediterranean region. *International Conference on Advances in Social Networks Analysis and Mining*. 6

Borovkov, A. A. and Yurinsky, V. (1998). *Ergodicity and stability of stochastic processes*. J. Wiley. 54

Breuer, A. and Farquhar, D. (2012). Social media and protest mobilization: Evidence from the tunisian revolution. *European Commu. Conference*. 3

Bulut, E. and Szymanski, B. K. (2015). Understanding user behavior via mobile data analysis. In *IEEE ICC Workshops, Dynamic Social Networks, DYSON*. 9

C. Danescu-Niculescu-Mizil, J. L. e. a. (2013). No country for old members: User lifecycle and linguistic change in online communities. *WWW*. 9

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., and Moon, S. (2007). I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proc. of ACM IMC*, pages 1–14, San Diego, California, USA. 45

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., and Moon, S. (2009). Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Transactions on Networking*, 17(5):1357 – 1370. 45

Chatzopoulou, G., Sheng, C., and Faloutsos, M. (2010). A first step towards understanding popularity in YouTube. In *Proc. of IEEE INFOCOM*, pages 1 –6, San Diego. 45

Cheng, X., Dale, C., and Liu, J. (2008). Statistics and social network of youtube videos. In *International Workshop on Quality of Service*. 45

Christina Pawlowitsch, Panayotis Mertikopoulos, N. R. (2011). Neutral stability, drift, and the diversification of languages. *Journal of Theoretic Biology*. 67

Christina Pawlowitsch, P. M. and Ritt, N. (2011). Neutral stability, drift, and the diversification of languages. In *Journal of Theoretical Biology*. 4

Crane, R. and Sornette, D. (2008). Viral, quality, and junk videos on youtube: Separating content from noise in an information-rich environment. In *Proc. of AAAI symposium on Social Information Processing*, Menlo Park, California, CA. 45

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., and Potts, C. (2013). No country for old members: User lifecycle and linguistic change in online communities. In *ACM International Conference on World Wide Web (WWW)*. 10

Fernández, I. G. M. E. and Seemann, P. A. A. (2009). A study on language changes of written spanish in internet chats. *Trab. linguist. apl*. 9, 15

Fisher, M. (2012). Gangnam style, dissected: The subversive message within south korea's music video sensatio. *The Atlantic*. 34

Foss, S. (2008). Coupling again: the renovation theory. *Lectures on Stochastic Stability, Lecture 6, Heriot-Watt University*. 53

Gill, P., Arlitt, M., Li, Z., and Mahanti, A. (2007). YouTube traffic characterization: A view from the edge. In *Proc. of ACM IMC*. 45

Mackeyi, R. (2010). Israeli soldiers dance into trouble on patrol. *The Lede, The New York Times News Blog*. 43

mei Sun, H. (2010). A study of the features of internet english from the linguistic perspective. *Studies in Literature and Language*. 9, 15, 18

Monde.fr, L. and AFP (2017). Twitter va tester les messages de 280 caractères. 4

Morchid, M., Portilla, Y., Josselin, D., Dufour, R., Altman, E., El-Beze, M., Cossu, J. V., Linares, G., and Reiffers-Masson, A. (2015). An author-topic based approach to cluster tweets and mine their location. In *Procedia Environmental Sciences*. 10

Nowak M. A., Komarova N. L., N. P. (2002). Computational and evolutionary aspects of language. In *Nature*. 4

O'Connor, B., Balasubramanyan, R., Routledge, B. R., , and Smith (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *4th Intl. AAAI Conf. on Weblogs and Social Media (ICWSM)*. 3

Pfeffer, A. (2010). Soldiers to face punishment for youtube video of hebron boogie. *Haaretz*. 43

Ratkiewicz, J., Menczer, F., Fortunato, S., Flammini, A., and Vespignani, A. (2010). Traffic in social media ii: Modeling bursty popularity. In *Proc. of IEEE SocialCom*, Minneapolis. 45

Richier, C., Altman, E., El-Azouzi, R., Jimenez, T., Linares, G., and Portilla, Y. (2014). Bio-inspired models for characterizing youtube viewcount. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 20)14)*, pages 297–305, Beijing, China. 6, 45, 58

Rimay, Z. (2010). Cybercultural communication. *Budapest Univ. of Technology and Economics, Faculty of Economic and Social Sciences*. 16

Salmon, J. M. (2016). 29 jours de révolution. histoire du soulèvement tunisien, 17-12 2010-14 janvier 2011. *Les Petits Matins*. 3

Solon, O. (2013). Richard dawkins on the internet's hijacking of the word 'meme'. *Wired UK*. 33

Sommer, A. K. (2010). Israeli soldiers gangnam style with palestinians - and the world goes wild. *Haaretz*. 42

Stephan Gouws, Donald Metzler, C. C. and Hovy, E. (2011). Contextual bearing on linguistic variation in social media. In *Workshop on Language in Social Media*. 9, 15

Szabo, G. and Huberman, B. A. (2010). Predicting the popularity of online content. *Comm. of the ACM*, 53(8):80–88. 45

Trapa, P. E. and Nowak, M. A. (2000). Nash equilibria for an evolutionary language game. *J. Math. Biol.* 67

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *4th Intl. AAAI Conf. on Weblogs and Social Media (ICWSM)*. 4

Winer, S. (2013). 'harlem shake' troops sent to prison. *The Times of Israel*. 42

www.internetlivestats.com (2018). Internet live stats. 56

www.statista.com (2018). Most popular social networks worldwide as of october 2018, ranked by number of active users (in millions). 56

Zhou, R., Khemmarat, S., and Gao, L. (2010). The impact of youtube recommendation system on video views. In *Proc. of IMC 2010*, Melbourne. 45, 54

Index