Accepted Manuscript

Detecting discussion communities on vaccination in twitter

Gema Bello-Orgaz, Julio Hernandez-Castro, David Camacho

PII: S0167-739X(16)30217-5
DOI: http://dx.doi.org/10.1016/j.future.2016.06.032
Reference: FUTURE 3098

To appear in: *Future Generation Computer Systems*

Received date: 23 March 2016
Revised date: 17 June 2016
Accepted date: 23 June 2016

Please cite this article as: G. Bello-Orgaz, J. Hernandez-Castro, D. Camacho, Detecting discussion communities on vaccination in twitter, *Future Generation Computer Systems* (2016), http://dx.doi.org/10.1016/j.future.2016.06.032

Detecting Discussion Communities on Vaccination in Twitter

Gema Bello-Orgaz[a], Julio Hernandez-Castro[b], David Camacho[a,*]

[a]*Escuela Politecnica Superior, Universidad Autonoma de Madrid, Madrid, Spain.*
[b]*School of Computing, University of Kent, Canterbury CT2 7NF, UK.*

Abstract

Vaccines have contributed to dramatically decrease mortality from infectious diseases in the 20th century. However, several social discussion groups related to vaccines have emerged, influencing the opinion of the population about vaccination for the past 20 years. These communities discussing on vaccines have taken advantage of social media to effectively disseminate their theories. Nowadays, recent outbreaks of preventable diseases such as measles, polio, or influenza, have shown the effect of a decrease in vaccination rates. Social Networks are one of the most important sources of Big Data. Specifically, Twitter generates over 400 million tweets every day. Data mining provides the necessary algorithms and techniques to analyse massive data and to discover new knowledge. This work proposes the use of these techniques to detect and track discussion communities on vaccination arising from Social Networks. Firstly, a preliminary analysis using data from Twitter and official vaccination coverage rates is performed, showing how vaccine opinions of Twitter users can influence over vaccination decision-making. Then, algorithms for community detection are applied to discover user groups opining about vaccines. The experimental results show that these techniques can be used to discover social discussion communities providing useful information to improve immunization strategies. Public Healthcare Organizations may try to use the detection and tracking of these social communities to avoid or mitigate new outbreaks of eradicated diseases.

Keywords: Social Big Data, Community Detection, Vaccines

*Corresponding author
Email addresses: gema.bello@uam.es (Gema Bello-Orgaz), jch27@kent.ac.uk (Julio Hernandez-Castro), david.camacho@uam.es (David Camacho)

1. Introduction

The use of vaccines has contributed to dramatically decrease mortality rates from infectious diseases in the 20th century [1]. In 1920, 469,924 measles cases were reported in United States, and 7575 patients died. The number of cases decreased to fewer than 150 per year in the 50s, and in 2008 there were only 64 suspected cases of measles in the world. However, currently, social groups related to vaccines have emerged influencing on the opinion of population about vaccination. This fact could bring on disease outbreaks because they are more common when vaccination rates decrease [2, 3, 4].

The vaccination communities have taken advantage of social media technologies to effectively disseminate its message and to spread their theories [5]. In recent years, several studies on various social media services such as YouTube [6], MySpace blogs [7], and Social Networks (SN) [8], present this dissemination, and their effect. In addition, statistical analysis show how this vaccination information influences social media users in their treatment decisions [9].

Currently, one of the most popular social networks is Twitter [10], producing huge amounts of public information. Twitter users can generate new sources of collective intelligence through their comments and interactions, allowing the application of data mining techniques in several fields [11] such as marketing campaigns [12, 13], financial prediction [14] or public healthcare [15, 16, 17], amongst others.

In the related literature, there are several works investigating knowledge acquisition from social networks about vaccine sentiments using classification techniques [18, 19, 20]. These classification techniques usually obtain better results than Clustering techniques as a consequence of its supervised nature. However, clustering techniques are able to discover hidden information (or patterns) on a dataset, and they don not need a previous human-labelling process. Any human-labelling process can be really time-consuming, or even impossible, for huge datasets extracted from SN as Twitter.

The information extracted from a SN can be represented as a graph, where the vertices represent the users, and the edges represent the relationships among them (i.e. a re-tweet of a message or a favourited tweet). This graph representation can be clustered into user groups, or communities, based on the topology information of the graph. Each community should include

strongly interconnected vertices and few connections with the rest of graph vertices. Therefore, the problem of community detection within a SN can be handled using graph clustering algorithms [21]. These algorithms can automatically organize a set of users from a SN into similar communities to acquire collective knowledge about their behaviour, preferences, profiles, etc.

This work aims to detect communities in Twitter which are disseminating vaccine opinions in order to analyse how it could be influencing to the rest of users in a particular community, zone, or country. Many people looks for vaccination information on the internet, and the data found can impact on their vaccination decisions. Therefore, Public Healthcare strategies could be improved through the application of data mining and community detection techniques, increasing control and preventive measures in the identified risk zones. In this particular work, the use of these techniques are focused on discovering and tracking anti-vaccine movements arising in SN. For this purpose, firstly an analysis of the Twitter Social Influence on the vaccine coverage rates is carried out. The second part of the work is focused on the study of the re-tweet graph, representing the user interactions who talk about vaccination. Firstly, applying Community Detection Algorithms to this graph, the existing vaccine communities are found. Then, different network metrics are calculated over these communities to discover the most relevant users to analyse their social influences.

The rest of the paper has been structured as follows: Section 2 shows the state of the art concerning the health impact of social vaccine groups, web mining solutions to detect them, and community detection algorithms. Section 3 describes the methodology used in this work to analyse the social influence on healthcare of user groups talking about vaccination, and how to detect these social vaccine communities from Twitter. Section 4 describes the dataset used, and the experimental results obtained. Finally, in the last section, the analysis and conclusions of this work are presented.

2. Related Work

2.1. Studies on anti-vaccination health impact

Recent outbreaks of preventable diseases such as measles, polio, and influenza show the effect of the decrease in immunization rates. The MMR vaccine is an immunization vaccine against measles, mumps, and rubella, generally administered to children around the age of one year, with a second dose before starting school (4-5 years). The first 20 years of licensed measles

vaccination in the United States prevented an estimated 52 million cases of the disease. The reported cases decreased from hundreds of thousands to tens of thousands per year since the introduction of the vaccine in 1963 [22]. Fewer than 200 cases have been reported year on year since 1997, and the disease is no longer considered endemic.

In the UK, the MMR vaccine was the subject of avast controversy after the publication of a paper by Andrew Wakefield et al. [23]. This work reported the results of a study of the MMR vaccine on twelve children who had bowel symptoms along with autism or other disorders in 1998. The research was declared fraudulent in 2011 by the British Medical Journal [24]. However the MMR-autism controversy covered by popular media caused a decline in vaccination rates. Before this publication, the rate for MMR vaccination in the UK was 92%, decreasing after to below 80%. In 2003, a study by Jansen et al. [2] shown that if the low level of MMR vaccine persisted, the increasing number of unvaccinated individuals will make a measles epidemic again. In fact, the number of new cases has heavily increased over the last years [25]. As shown in Figure 1, while in 2000 there were 104 measles cases from UK, in 2013 there were 1919 cases, with 1 confirmed death.

Public Health Wales reported at the end of 2014, 44 cases in measles outbreak detected in that year. This outbreak has been linked to four schools in the Neath and Swansea area, and it follows the largest ever occurred in Wales with more than 1,200 cases in the same area between November 2012 and July 2013. In that outbreak, 88 people were hospitalised and one adult died. Although more than 70,000 catch up doses of MMR were given across Wales during the last outbreak, around 30,000 children and young people in the 10 to 18 age group remain unprotected.

In April 2014, Health officials of New York City reported that at least 25 persons, including 13 children, have contracted the measles virus. The outbreak emerged in northern Manhattan and the Bronx, and later spread downtown to the Lower East Side. Furthermore, a case of diphtheria was recently detected in Spain on 30 May 2015. A six year old child, who had not been immunized against the disease, was being treated with an antitoxin that proved ineffective, and the child died. There has not been a single case of diphtheria in Spain for the previous 28 years.

Another example of the potential effects on public health care due of distrust in vaccines is the influenza A(H1N1) vaccine. In June 2009, the World Health Organization (WHO) declared the influenza A(H1N1) pandemic. The influenza A(H1N1) virus was monitored around the world for

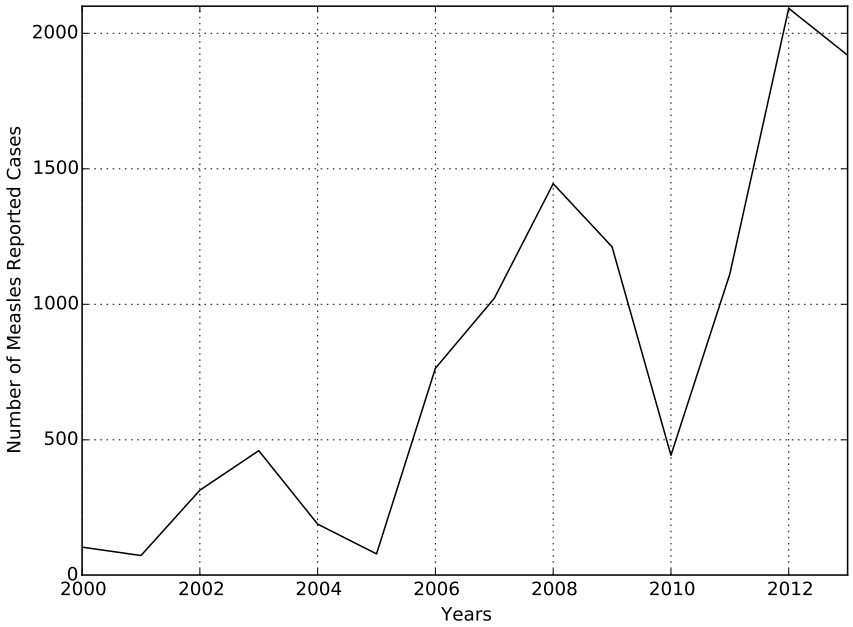


Figure 1: Number of Measles Reported Cases between 2000 and 2013 from United Kingdom.

changes in virulence or epidemiology, to have vaccines ready, but vaccine supply was insufficient in some areas [26]. The population wants to be assured that there will be enough vaccine when an outbreak occurs, but at the same time some were questioning the safety and effectiveness of the vaccine.

Finally, the controversy over polio vaccination happened in northern Nigeria between 2003 and 2004. It led to a resurgence of the disease and contributed to reinfection in 20 previously polio-free countries, reaching as far as Indonesia and still affecting Nigeria [27, 28].

The previous studies show that social groups related to vaccines can influence the opinion of population about vaccination, decreasing the immunization rates in some cases. Furthermore, this can bring on disease outbreaks because these outbreaks emerge when vaccination rates decrease. Therefore, Healthcare Organizations may try to use the detection and tracking of these groups or communities to avoid or mitigate new outbreaks of eradicated diseases.

2.2. Web Mining solutions for vaccination sentiments and attitudes

Nowadays, there are several works related to knowledge acquisition from web sources focused on vaccine sentiments. For this purpose, the VASSA (Vaccine Attitude Surveillance using Semantic Analysis) framework [18] combines Semantic Web and Natural Language Processing (NLP) techniques with online data for the assessment and analysis of vaccination attitudes and beliefs. Blog posts were sampled using the Google.ca search engine to search terms such as immunize, immunization, vaccine, and vax, among others. Then, using the Vaccine Sentiment Ontology (VASON) the framework identifies the concepts and relationships between them that can be used to infer vaccination attitudes and beliefs. The annotation scheme generated has been tested on a small sample of blog posts. The authors proposed as future improvements the application of their method for extraction and classification onto a larger sample to validate it.

In Botsis et al. [29], a multi-level text mining approach is presented for automated text classification of reports collected from the US Vaccine Adverse Event Reporting System (VAERS). A total of 6034 VAERS reports for the H1N1 vaccine were classified by medical officers as positive or negative, generating a corpus of text files. Firstly, text mining techniques were applied to extract three feature sets for important keywords. Then, several machine learning classifiers were trained and tested. The results of this work showed that Rule-based classifies, boosted trees, and weighted support vector machines performed well in terms of macro-recall, however at the expense of a higher misclassification error rate.

A novel modelling framework combining Social Impact Theory (SIT) characterization, with a game-theoretical analysis to study vaccination decision making is proposed by Shang et al. [30]. They used a social network representation of individuals to model the structure of their relationships. Moreover, they modelled using SIT characterization the strength of social influence on changing vaccination decisions by the influence of others, and the associated costs. The simulation results obtained suggest that individuals with high social influence increase the vaccination coverage, if the cost of vaccination is low. However, if individuals are social followers, the resulting vaccination rates depend on the vaccination sentiment rather than the associated costs. Another framework is presented in Shaw et al. [31] by modelling the spread of pathogens throughout a population to generate policies that minimize the impact of those pathogens. This framework combines agent-based simulation, mathematical analysis and an Evolutionary

Algorithm (EA) to determine the optimal distribution of vaccine supply.

In 2010, The Vaccine Confidence Project was launched to monitor and generate online reports about vaccines, vaccination programmes, and vaccine-preventable diseases, using data collected from the HealthMap system [19]. These reports were manually analysed, and categorised by concern, vaccine, disease, location, source of report, and overall positive or negative sentiment towards vaccines. Data from 10.380 reports (from 144 countries) was analysed between May 1, 2011, and April 30, 2012 showing that 69% of the collected corpus contained positive or neutral content, and 31% contained negative content. To further improve the system, extra efforts were focused on automating the data gathering and classification as much as possible.

Finally, in Salathe et al. [20], machine learning algorithms were applied to classify tweets. These tweets were labelled as negative, positive or neutral with respect to the user intent of getting vaccinated with the influenza H1N1 vaccine. The authors used an hybrid approach based on a naive Bayes and a maximum entropy classifier. Moreover, a study of the spread of health sentiment was performed. For this purpose, a statistical approach was used to measure the individual temporal effects of a large number of variables based on social network statistics. They found that negative sentiments are contagious while positive sentiments are generally not. These results suggest that the effects of behaviour spread on social networks are strongly content-dependent.

2.3. Community Detection Algorithms

The Community Detection Problem in Complex Networks has been the subject of many studies in the field of Data Mining and Social Network Analysis. There are several methodologies in the literature to find optimal groups of nodes into communities. The goal of the Community Detection Problem is similar to the idea of graph partitioning in graph theory [21, 32]. In computer science, the unsupervised process of identifying the underlying structure of the data in terms of grouping the most similar elements is called clustering [33]. Elements included in the same cluster should be similar, and elements included in different clusters should be dissimilar. The concept of similarity or dissimilarity will depend on some kind of metric. A cluster in a graph could be easily mapped into a community.

Graphs are structures formed by a set of vertices (also called nodes) and a set of edges which are connections between pairs of vertices. Graph clustering [34, 35] can be understood as the process of grouping the vertices into

clusters considering the edge structure of the graph. One of the most well-known algorithms for Community Detection was proposed by Girvan and Newman [36]. This method uses a similarity measure called "Edge Betweenness" based on the number of the shortest paths between all vertex pairs. This algorithm has however a high computational complexity. For this reason, Newman reformulated the modularity measure in terms of eigenvectors. The new characteristic matrix for the network is called modularity matrix [37]. The reformulated algorithm, based on modularity, has been employed by many authors to study community structures of complex networks, and it shows excellent performance when the size of the network is small. The main disadvantage of this algorithm is the high computational complexity on very large networks. Subsequently, the modularity measure was modified trying to reduce the computational demands significantly through several new approaches [38, 39, 40, 41].

3. Detecting and Analysing Social Vaccine Communities from Twitter

In this work, in order to extract collective knowledge from Twitter and to discover social vaccine movements or trends, two main phases have been performed. The first one is focused on measuring and analysing the potential healthcare influence of vaccine opinions from Twitter users. For this purpose, a comparative assessment of two factors is carried out: Topic Relevance (TR_f), and Kurtosis of Vaccination Coverage Rates (K_{VCR}). TR_f per country measures the importance of the countries which are talking about vaccination (see eq. 2). On the other hand, K_{VCR} per country measures the variation in the coverage rates of population vaccinated by antigen in a particular country (see eq. 3). Therefore, the comparative assessment between these two factors will allow us to perform an influence analysis of opinions from social networks on vaccination decision making.

The second phase is based on the analysis of Social Network structure, to retrieve information about how the different communities are constructed, and to find the most relevant users. In this case, users have been considered as network nodes, and their relationships (re-tweets) are represented as the edges. Once the graph is generated, several network algorithms and metrics are applied to the detected communities, discriminating the most relevant users within each community. Finally, an information diffusion analysis is

carried out to discover users that control the flow of information and therefore are most influential. The following subsections describe both phases in detail.

3.1. Analysing the potential influence of social vaccine communities on healthcare

To carry out a social influence analysis, firstly a dataset which contains vaccine-related tweets has been gathered. In addition, the vaccination coverage rates published by the immunization monitoring system of the World Health Organization (WHO) [42] have been retrieved and used. This official report shows, for each country, its official coverage estimation per year. Using both datasets, two factors (TR_f and K_{VCR}) have been calculated per country in order to measure the social influence on immunization rates. For this purpose, four sub-phases have been performed: Data Extraction, Data Preprocessing to Identify Tweet Locations, Social Data Analysis, and Visualization of Geo-Spatial Information. These are detailed in the following subsections.

3.1.1. Data Extraction

In this work, all gathered data has been extracted from two sources: Twitter and the Immunization Monitoring System of the World Health Organization (WHO).

- **Twitter** [10]: This is a Social Network where users share information about personal opinions in tweets. Tweets are posts, limited to 140 characters, containing information about opinions, photos, links, etc. A special kind of tweet is the re-tweet, which is created when one user reposts the tweet of another user. Users on Twitter generate over 400 million tweets every day, and they are available through public APIs which provide functionalities for searching by keywords, hashtags, phrases, geographic regions, or user-names. The information collected for this work were all the tweets containing the word *'vaccines'*.

- **WHO web site** [42]: The immunization monitoring system of WHO, collects reports including information such as the estimations of national Immunization Coverages, reported cases of Vaccine Preventable Diseases (VPDs), Immunization schedules, or indicators of immunization system performance, amongst others. This information is available by WHO Member State, as well as summarized by WHO Region.

3.1.2. Data Preprocessing to identify Tweet locations

Data preprocessing methods prepare the data to be analysed. Immunization coverage rates obtained from the WHO are reported per country, therefore the *location information* of the tweets is necessary to analyse social influence on vaccination coverage. Location information on Twitter is available from two different sources: geotagging (users can optionally choose to provide location information for the tweets using a system with GPS capabilities) or using the user profile information (user location can be extracted from the location field in the user profile).

Only 1% of all Tweets are geolocated, and it is often necessary to use the user profile information to determine location. In addition, the location string from the user profile must first be translated into geographical coordinates. There are several online services (Bing, Google, and MapQuest among others) which can take a location string as input, and return the coordinates of the location as output. The granularity of the location is generally thicker in the case of large regions, such as the center of town for a given city name. In this work the preprocessing has been divided into two further steps:

1. **Geocoding Process**: It is the process of converting addresses ("Mountain View, California") into geographic coordinates (37.42, -122.08). To determinate the location information for not geolocated tweets, the http geocoding service provided by the Google Maps API has been used.
2. **Finding country location**: This process translates the geographical coordinates into a particular country. The Geospatial Data Abstraction Library (GDAL) is a translator library for geospatial data formats, and it has been used for processing the geographic coordinates to find the origin country.

3.1.3. Social Data Analysis

Once the vaccine information is extracted and preprocessed, two factors per country are calculated to measure the healthcare impact of vaccine opinions: Topic Relevance Factor (TR_f) and Kurtosis of Vaccination Coverage Rates (K_{VCR}).

1. **Topic Relevance Factor (TR_f)**: The number of Twitter users who talk about vaccines in a given country can be used to quantify the relevance of this topic for each country. Countries with a higher number of tweets related to vaccination will be the most relevant ones. However,

there is a huge difference in Twitter usage per country, therefore a normalization will be made. Statista [43] provides information on the Twitter penetration per country. It is defined as the number of active twitter users relative to the total amount of internet users. On the other hand, the data on internet users by country can be extracted from Internet Live Stats [44] (see Figure 2). The TR_f factor for each particular country is calculated as follows:

$$TR_f(C) = \frac{\%NTwitterVaccineUsers}{\%TwitterPenetration \times \%InternetUsers} \quad (1)$$

Where C is a given country, $\%NTwitterVaccineUsers$ is the percentage of users retrieved who are talking about vaccines in this country, $\%TwitterPenetration$ is the percentage of Twitter Penetration for this country, and $\%InternetUsers$ is the percentage of Internet Users in the country. Finally, to scale the factor values into a range of $[0, 1]$, an unity-based normalization is applied:

$$\overline{TR_f(C)} = \frac{TR_f(C) - min(TR_f(C))}{max(TR_f(C)) - min(TR_f(C))} \quad (2)$$

2. **Kurtosis of Vaccination Coverage Rates (K_{VCR})**: The potential influence of social movements on vaccination coverages can be estimated by measuring the distribution of changes of the coverages rates. In probability theory and statistics, *kurtosis* is the measure of the "peakedness" of the probability distribution, also showing how heavy the tails are. A high kurtosis distribution has a sharper peak and fatter tails, whereas a low kurtosis distribution has a more rounded peak and thinner tails [45]. In this work, the kurtosis value is calculated as per Fisher's definition [46] where 3.0 is subtracted to the kurtosis values in order to obtain a result of 0.0 for normal-like distributions. Therefore, values equal to 0 correspond to a *normal* distribution, whereas values greater than 0 are indicative of a *leptokurtic* distribution. Finally, a *platykurtic* distribution correspond to values lower than 0. Regarding the variation of the coverage vaccination rates, high *kurtosis* values represent a sharp change on these rates. It can be assumed that these changes can be used to detect strong variations in the immunization rates. In this work, the kurtosis value has been calculated using the last 10 years of coverage vaccination rates for each country (see eq. 3).

$$K_{VCR}(C) = n\frac{\sum_{i=1}^{n}(X_i - X_{avg})^4}{(\sum_{i=1}^{n}(X_i - X_{avg})^2)^2} - 3 \quad (3)$$

Where C represents a given country, n is equal to 10 (the coverage rates in the 10 last years are used as sample to calculate the metric), and X_i is the immunization coverage rate of a specific year.

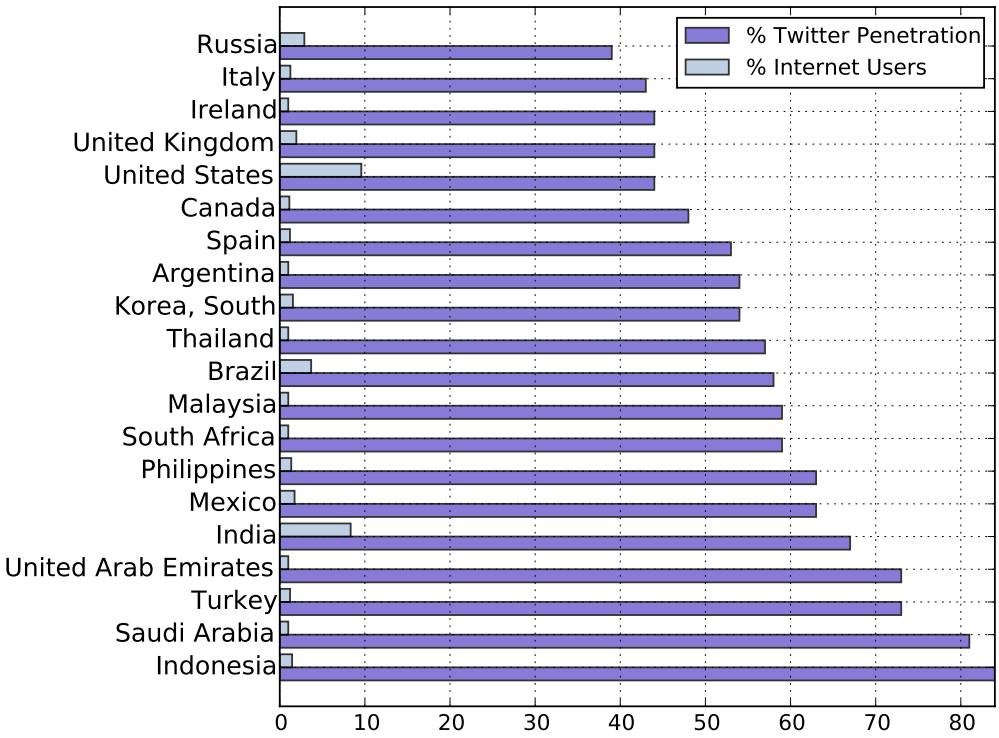


Figure 2: Penetration of Twitter Users and Internet Users per country (Top 20). Data taken from Statista and Internet Live Stats.

Finally, to validate the potential influence on immunization coverages of twitter opinions, the ***Spearman correlation coefficient*** [47] has been used. This coefficient is a nonparametric measure of the linear relationship between two datasets. If both previous factors (TR_f and K_{VCR}) are correlated, this could be due to a potential influence from social trends which

affect the vaccination decision-making. It means that the most relevant countries talking about vaccination (with higher TR_f), are the countries having strong variations in vaccination rates (with higher K_{VCR} values). The values of Spearman's correlation coefficient varies between -1 and +1, with 0 implying no correlation. Correlations of -1 or +1 imply an exact linear relationship. Positive correlations imply that as x increases, so does y. Oppositely, negative correlations imply that as x increases, y decreases. A table of critical values of the Spearman correlation coefficient for different significance levels is given in Zar [48]. In our work these critical values are used to validate whether there is a significance correlation between both factors.

3.1.4. Visualization of Geo-Spatial Information

Geo-spatial visualization can help to study the results obtained from the social data analysis. The location information can be used to show the most interesting locations or countries discussing on a specific topic. A Map is the best choice to visualize this kind of information. It can be used to effectively summarize location information, allowing an easy identification of interest regions on the topic (with high number of users opining about vaccines). In this case, the events measured have been the user locations grouping per countries. Therefore, the map is generated based on the TR_f factor per country, as previously calculated.

3.2. Vaccine Community Detection in Twitter

The second analysis phase is focused on the study of data network structure to provide information about two main aspects: on the one hand, the community detection of different user communities into the social network that are talking about a topic; on the other hand, the most relevant users of these communities to analyse the social information diffusion showing how each community is opining about vaccination. For this purpose, a network representation of the dataset has been generated. The users have been considered as the network nodes, and their relationships represent the edges. The relationships which have been considered in this work are the re-tweets. When any user re-tweets a message from other user, an edge between both users is generated. Therefore, for a social analysis of the re-tweet network generated, two phases are performed: Finding Social Communities and Analysing Social Network Data.

3.2.1. Finding Social Communities

Community detection algorithms have been studied extensively in computer science [21] but in particular, for social media mining [14, 49]. Individuals often form groups based on their common interests, and identifying groups of similar users can provide a global view of user interactions and their behaviours. In addition, some behaviours are only observable into a group and not on an individual level. This is because individual behaviour could easily change, but collective behaviour is more robust to changes.

There are several community detection algorithms, usually classified into two types: *member-based* algorithms that find groups based on the characteristics of their members; and *group-based* algorithms where the groups are formed based on the density of interactions among their members. In this work, a comparative assessment of group-based algorithms is carried out, to choose the most appropriate method for better identifying communities talking on a particular topic (vaccination in this case). The algorithms selected for this purpose were the following:

1. *Fast-Greedy* [38]: This algorithm merges individual nodes into communities in a way that greedily maximizes the modularity of the graph.
2. InfoMap [50]: This algorithm uses the probability flow of random walks on a network, as a proxy for information flows in the real system. Then the network is decomposed into modules by compressing a description of the probability flow. The result is a map that simplifies and highlights the regularities in the structure, and their relationships.
3. *Leading Eigenvector* [39]: In this algorithm the network is split into two components according to the leading eigenvector of the modularity matrix. Then recursively takes a given number of steps by splitting the communities as individual networks.
4. *Label Propagation* [51]: Initially, each vertex is assigned to a different label. Then, each vertex chooses the dominant label in its neighbourhood for each iteration. Ties are broken randomly and the order in which the vertices are updated is randomized in every iteration. The algorithm ends when vertices reach a consensus.
5. *Multi-Level* [52]: This is a bottom-up algorithm, where initially every vertex belongs to a separate community, and vertices are moved between communities iteratively in a way that maximizes the vertices local contribution to the overall modularity. The algorithm stops when it is not possible to increase this modularity.

6. *Walktrap* [53]: It based on random walks, according to the idea that short random walks tend to stay in the same community.

Community evaluation is a difficult task, partly because a list of community members is rarely known. A good community, based on their network structure, would be: modular, balanced, dense, and robust. Therefore, traditional metrics of network topology can be useful to evaluate the results obtained:

1. *Modularity* [54]: It is the most popular quality function to identify good partitions. This measure is based on the idea that a random graph is not expected to have a cluster structure. Therefore the possible existence of clusters is revealed by the comparison between the actual density of edges in a subgraph and the density that would be expected in a random subgraph.
2. *Density* [55]: The connectivity of a node is computed using the size of its neighbourhood. The average number of neighbours indicates the average connectivity of a node in the network. A normalized version of this parameter is the network density, which is a value between 0 and 1.
3. *Cohesion* [56]: This metric computes the vertex connectivity between some vertices of the graph. This value represents the number of vertices that need to be removed in order to disconnect two vertices into two separate components. The vertex connectivity will be the minimal vertex connectivity over all vertex pairs.
4. *Omega* [57]: It represents the clique number of the graph that is the size of the largest clique (a subset of its vertices such that every two vertices in the subset are connected by an edge).

3.2.2. Analysing Social Network Data

After the anti-vaccine communities have been detected, there are several networks metrics that can be used to discriminate the most relevant users within each community to allow the study of their social influence. Usually, the importance, or influence, in a social network is analysed through **centrality measures**. The most frequently used in social media analysis are [49]:

1. *Degree Centrality*: It used to analyse the interactions between users. Users with higher number of connections, or larger degree values, will

be considered the most representative. The degree centrality for a node in an undirected graph is calculated as the number of adjacent edges of this node.

2. *Eigenvector Centrality*: Using degree centrality, nodes with more connections are considered more important. However, in real-world cases, having more friends does not by itself guarantee relevance. Instead of this, having more important friends usually provides a higher relevance degree. This measure tries to generalize the degree centrality based on this idea by incorporating also the importance of the neighbours.
3. *Betweenness Centrality*: Other approach for measuring the centrality is to compute the number of shortest paths between a particular node and the ones that pass through it. This measure shows how central is a node connecting any other pair of nodes into the network.

Each of these measures provide a different view about who is important in the network. In the context of a re-tweet network, these measures allow to detect different aspects: who are the most re-tweeted users (Degree Centrality), who are the most influential users (Eigenvector Centrality), and who are the users controlling the information flow (Betweenness Centrality).

4. Experimental Results

4.1. Dataset Description

As described in Section 3.1, the data collected to perform the data analysis of vaccination influence was extracted from two sources: Twitter APIs, and WHO Web Site.

1. **Twitter APIs** [10]: The information collected from the Twitter APIs are comments mentioning the hashtags: *'vaccine, vaccines, #vaccine or #vaccines'*. These comments have been taken from all the countries between *'04-15-2014'* and *'11-08-2014'* (both dates inclusive). Table 1 shows the number of tweets gathered during this time period. As shown, less than the **1%** of all tweets are geo-located. However, after performing the preprocessing to identify tweet locations (described in Section 3.1), this value increases noticeably up to **51%**.
2. **WHO Web Site** [42]: The official country reported coverages are extracted at the last ***10*** *years* for ***five*** *different vaccines*:
   (a) *DPT1*: First dose of diphtheria toxoid, tetanus toxoid and pertussis vaccine.

Total Number of Tweets	1.448.010
Number of Geolocated Tweets	11.566 (**0,8%**)
Number of Geolocated Tweets after Preprocessing Data	761.924 (**51,2%**)

Table 1: Number of tweets on vaccine related topics during seven months.

(b) *DPT3*: Third dose of diphtheria toxoid, tetanus toxoid and pertussis vaccine.
(c) *HepB3*: Third dose of hepatitis B vaccine.
(d) *MCV*: Measles-containing vaccine.
(e) *POL3*: Third dose of polio vaccine.

The total number of countries with data on vaccination coverages is equal to 194.

4.2. Results of the potential influence of social vaccine communities on Healthcare

Using the aforementioned information the two factors (TR_f and K_{VCR}) have been calculated to measure the social influence on immunization coverage of social communities on Twitter. In this section, firstly a preliminary analysis of the results for TR_f is performed to identify the most relevant countries and their relevant regions. Finally, a comparative assessment between both factors has been carried out, analysing the potential influence that opinions from social networks have on vaccination decision making.

4.2.1. Preliminary analysis for Topic Relevance Factor (TR_f)

Figure 3 shows the results obtained for TR_f factor. To show the results, the TR_f values are ordered generating a ranking of countries by relevance. It can be noticed that there are only a few countries with large values while most countries have very low values. As shown in Figure 3, the top five countries have values higher than the mean plus one standard deviation. It means that most users talking about vaccines belong to these group of countries, therefore it can be considered that they are the most relevant for the topic.

To continue the analysis of results, a geo-spatial visualization has been performed to allow for a visual analysis a social data across all countries. A Map summarizing the location information of the users talking about vaccines is generated, allowing the identification regions of interest. The map is generated using the TR_f results obtained for each country.

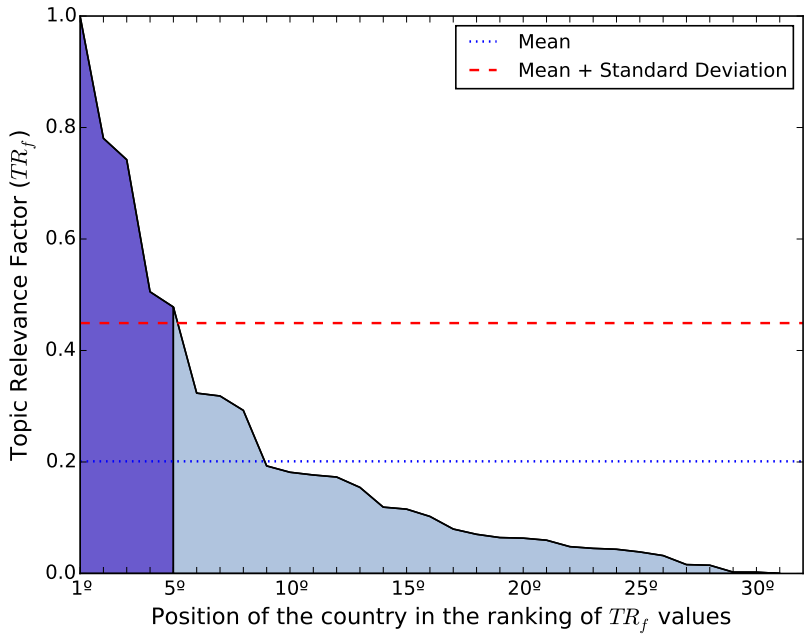


Figure 3: Values of TR_f for the countries talking on the vaccination topic. The results are shown in order as a ranking of countries by relevance.

Analysing the World Map shown in Figure 4, tree distinct blocks of interest can be identified. The first block is formed by Ireland, United States, United Kingdom, Canada and Australia which are the countries more active on vaccination in Twitter. For this five countries the TR_f factor takes values much higher than for the rest. The second block is build by several countries which show a moderate interest, such as France, Netherlands, Sweden, Malaysia, South Africa, Spain and the Philippines. Whereas, the third block is composed by most of the countries, which have very low values of relevance (less than 0.1). In the last block, there is not a clear influence effect over vaccination coverage.

4.2.2. Analysing the potential influence of Twitter vaccine communities on healthcare

Taking into account the previous results, countries belonging to the first block (top 5) are relevant, and have been selected in order to continue the

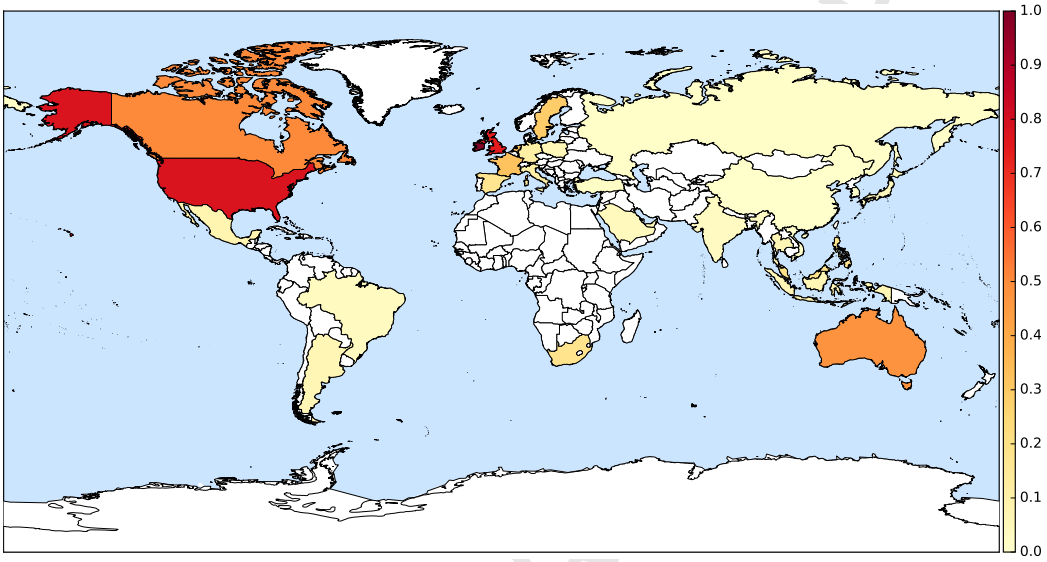


Figure 4: World Map based on TR_f, measuring the relevance of vaccination topics for each country.

analysis. Table 2 shows the results of both factors computed for these top five countries. If vaccine opinions affect user vaccination decision making, the immunization coverage rates of vaccination would show variations. This can be analysed studying the K_{VCR} values. In this work this measures are calculated using the Fisher definition. If all immunization coverage values are identical during the last 10 years, the value obtained will be -3, implying that there is not variation on the distribution. On the other hand, high kurtosis values would indicate a sharp change in the variation of vaccination. As can be seen in Table 2, almost all K_{VCR} values are higher than -3, being pretty high in cases such as Canada and Australia, which take values of 5.11 for some vaccines. Therefore, it is possible to notice a change on the vaccination pattern in these countries.

As mentioned in section 3.1, to discover a potential social influence of twitter opinions on immunization rates, it should appear a linear relationship between both vaccination factors (TR_f and K_{VCR}). For this, the *Spearman correlation coefficient* [47] has been calculated for each vaccine. This correlation coefficient measures the dependency between variables. It allows evaluating if countries who are talking more about vaccination correspond to

Country	***Number of Users***	TR_f	K_{VCR}				
			DPT1	***DPT3***	***HepB3***	***MCV***	***POL3***
Ireland	860	1	-1,14	-1,12	-0,67	-0,67	-1,12
United States	49278	0,78	-2	-0,63	**-0,58**	**1,14**	-0,5
United Kingdom	9560	0,74	-0,22	0,97	NaN	-0,83	-0,97
Canada	4117	0,5	**5,11**	-0,63	-1,65	-1,27	-1,23
Australia	1719	0,48	-1,24	**5,11**	-1,49	-3,0	**5,11**

Table 2: TR_f and K_{VCR} values for the top 5 most relevant countries on vaccination discussion. The K_{VCR} values are calculated for the five vaccines, considered in the last 10 years. There is no data available of HepB3 vaccine for United Kingdom.

Vaccine	***Spearman coefficient***	***p Value***
DPT1	-0,2	0,74
DPT3	**-0,82**	0,08
HepB3	0,59	0,4
MCV	**0,9**	0,03
POL3	-0,3	0,62

Table 3: Values of Spearman Coefficient Correlation applied to TR_f and K_{VCR} vaccination factors. This coefficient has been calculated using the top 5 relevant countries. For a ranking of 5 values, the minimal critical value is 0,5. In the HepB3 case, there is no data available for United Kingdom, being 0,6 the minimal critical value.

countries with higher variations in vaccination rates. The results obtained can be seen in Table 3.

In Zar [48], the critical values of the Spearman correlation coefficient for different significance levels were presented. Specifically, for a ranking of 5 values, the minimal critical value that shows a significant correlation between two variables, is **0,5**. As shown Table 2, there is no data available of HepB3 vaccination coverages for United Kingdom. Therefore, the minimal critical value is 0.6 for this specific case. Analysing the results shown in Table 3, two values are higher than this threshold. This means that vaccine opinions from social groups could influence the vaccination decision making for DPT3 and MCV vaccines. Positive correlations mean that both variables simultaneously increase (MCV). On the other hand, negative correlations mean that as one variable increases the other variable decreases (DPT3). In the results obtained, there is one vaccination coverage rate (MCV) that shows an increment directly related to the increase of TR_f in the countries.

On the other hand, there is one vaccine (DPT3) where the opposite effect occurs. This may be because not all social movements arising from Twitter are against vaccination. It can be that there are also supporting movements trying to increase immunization rates.

4.3. Results of Community Detection for groups on vaccination discussion

Once the social influence analysis from twitter opinions has been carried out, this section reports an additional analysis based on the data network structure. This analysis is focused on community detection for users talking about vaccination. Then, using these communities detected, a study of the most relevant users, user interactions, and their collective behaviour is performed. For this purpose, a network representation of the dataset based on the user re-tweets is generated. To select the most influential users, a minimum threshold number of re-tweets has been fixed. In this case this *threshold* is set to 10 re-tweets. There are *2865 users* exceeding this threshold in our dataset, which are used to generate the network for the social study.

In the literature there are several community detection algorithms which can be applied to solve this problem [38, 39, 51, 52, 53]. Therefore, firstly a comparative assessment of these algorithms is carried out to choose the most appropriate. Table 4 shows several topology network metrics (omega, cohesion, density and modularity) computed for each algorithm (see section 3.2 for a further description of these metrics). Algorithms showing the best results are Fast-Greedy and Walktrap, which generate the highest values for Omega, Cohesion and Density metrics. Regarding the Modularity metric, the Multi-Level algorithm obtains the best value, but Fast-Greedy algorithm obtains a value which is very close. Taking into a count all these network metrics, it can be concluded that the Fast-Greedy algorithm has obtained overall the best results. Therefore, it has been chosen to perform the detection later on.

Table 4.3 shows the communities found applying the Fast Greedy algorithm. To study the importance, or influence, of the different users into the re-tweet network generated, centrality network metrics have been computed. In addition, to identify the collective opinion for each community about the topic, a human-labelling process of the most frequent re-tweets has been performed. For each community, the top 10 of most frequent re-tweets are classified as positive or negative extracting the collective sentiment for the community. The last column in Table 4.3 shows the most frequently re-tweet for each community, and in the first column it can be seen the results of

Algorithm	*Communities*	*Omega*	*Cohesion*	*Density*	*Modularity*
Fast-Greedy	11	**0,36**	**0,27**	**0,06**	0,83
InfoMap	20	0,1	0,05	0,02	0,79
Loading Eigenvector	13	0,15	0,08	0,03	0,79
Label Propagation	12	0,24	0,18	0,04	0,80
Multi-Level	12	0,25	0,08	0,01	**0,84**
Walktrap	11	**0,36**	**0,27**	**0,06**	0,78

Table 4: Comparative assessment of community detection algorithms for different network topology metrics.

the human-labelling process, showing if the community has a positive (P) or negative (N) opinion.

As shown in Table 4.3, there are 7 communities (1,2,4,5,7,8 and 10) talking positively about vaccination against 4 which are talking negatively (3,6,9,11). Analysing the network structure, negative vaccine communities often include few users and have low values regarding centrality metrics. Specially, very low values are observed when Betweenness Centrality (representing the users that control the information flow) metric is analysed. Otherwise, the positive vaccine communities are generally bigger and have higher values of centrality metrics. This means that the most important and influential users, and those controlling the information flow, belong to positive communities. Therefore, as these results shown, it is possible to identify anti-vaccine movements from Twitter applying community detection algorithms. These algorithms are unsupervised data mining techniques, thereby human-labelling is not needed. This is a big advantage for huge datasets collected from social networks such as Twitter.

Regarding the social analysis of the communities found, each centrality metric shown in Table 4.3 provides a different aspect of its social influence. Firstly, analysing Degree Centrality, the most re-tweeted users can be identified as seen in Figure 5. Using this metric, users with more connections are considered as more relevant. In Figure 5 we see two main communities (1 and 4) including most of the important users, or institutions, which are discussion on vaccination. Several of these users correspond to relevant health organizations such as WHO, UNICEF or VaccinesToday, which belong together in the same community (1). In addition, Bill Gates also belongs to this community, and he is one of the most well-known and influential personalities who actively supports pro-vaccination campaigns. In the other most

Id.	*Top Users*	*N. Users*	*Degree*	*Eigen.*	*Betwe.*	*Most Frequently Re-Tweet*
1(P)	**VaccinesToday**,WHO, UNICEF, sanofipasteur, BillGates	42	14	1	4236,75	eu research commissioner: 'the best if people don't use it' vaccine in the world is worth nothing
2(P)	**CNN**, BeckOTR, cnni, TIME, CNNVideo	13	6	0,11	691,77	worried about childhood vaccines? don't worry, evidence strongly suggests they're safe see why it's important
3(N)	**megtirrell**, unicefusa, aetiology, StephenAtHome, sheridanmarfil	15	2	0,16	112,71	holy pharma deals: novartis buys gsk cancer biz for $14.5b, gsk buys novartis vaccines for $7.1b;...
4(P)	**CMichaelGibson**, washingtonpost, MiaFarrow, benoitbruneau, mikiebarb	26	7	0,22	3700,96	#cdc: vaccines prevent more than 700,000 child deaths in the u.s. reuters
5(P)	**timminchin**, mattliddy, LOLGOP, ChrisWarcraft, carnivillain	14	6	0,01	1633,99	when discussing vaccines, remember: stories work better than stats. help me spread this letter...
6(N)	**UnusualFactPage**, SteveStfler, FemaleTexts, BestProAdvice, LifeFacts	9	6	0	26	the scientist who developed the vaccine to fight leprosy is almost 100 years old, and he still working...
7(P)	**AmerAcadPeds**, Skepticscalpel, kevinmd, AAPNews, healthychildren	13	7	0,29	1106,50	immunize for a healthy future. know which vaccines you need.#ruuptodate with yours? if not...
8(P)	**CDCgov**, marstu67, MSF_uk, MotherJones, segmentis	21	5	0,18	1242,14	it's national infant immunization week! now is a great time to check what vaccines your baby needs...
9(N)	**VaccineXchange**, EVaccines, ELEN_A,the_refusers, cinderstella	12	8	0	739,11	drug giant glaxosmithkline has been using bribery and fudging its research in china. they make vaccines...
10(P)	**VacciNewsNet**, BBCWorld, GSK, OpposingThumb, RealDeanCool	16	6	0,26	2645,47	how many vaccine-preventable outbreaks have to happen before we realize this?
11(N)	**FeminineHygiene**, FollowIceland, TheEyesOfTexas_, ActOf1871, WeHateAmerica_	8	6	0	3,83	good morning patriots,enjoy your #nwo supplied #chemtrails #fakedebt #fluoride...

Table 5: Communities detected using Fast Greedy algorithm. The centrality metrics (Degree Centrality, Eigenvector Centrality and Betweenness Centrality) are related to the most influential user of the community, and are marked in bold. The Id column shows if the community has a positive (P in green) or negative **(N in red)** opinion on vaccination.

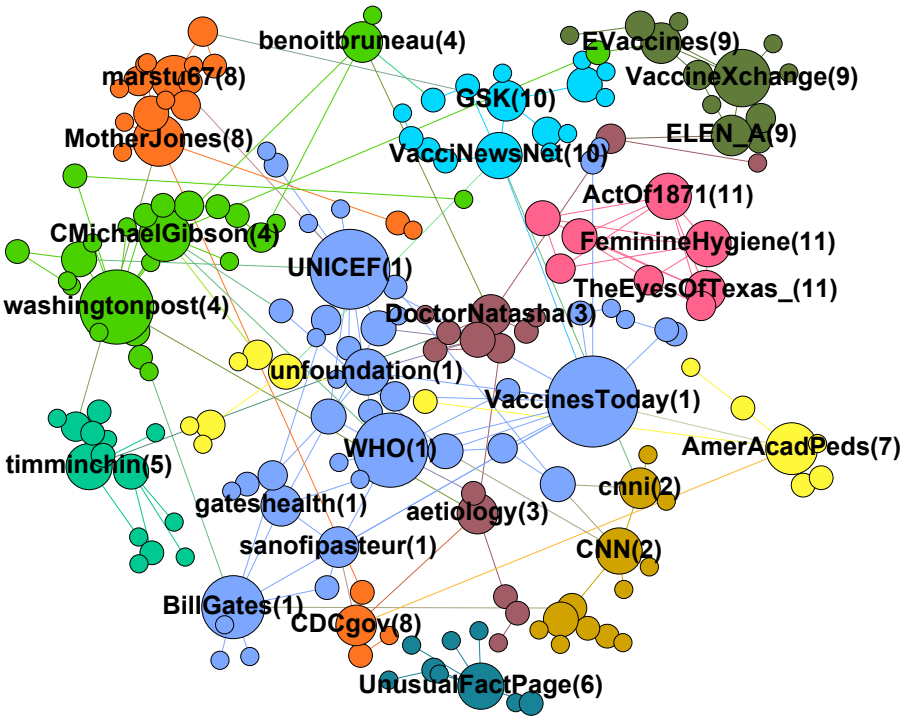


Figure 5: Vaccine Communities showing the **most re-tweeted** users based on their **Degree Centrality** metric (node size according to its value). Node labels are filtered by a degree value higher than 4. Top 5 users: VaccinesToday(1)(P), UNICEF(1)(P), washingtonpost(4)(P), WHO(1)(P) and BillGates(1)(P).

relevant community (4) based on Degree Centrality, an important international media as Washington Post appears. On the other hand, there is only a highly re-tweeted user (Vaccine eXchange) belonging to a negative vaccine community (9). This may be because negative communities tend to be small and poorly connected, as was previously discussed in the network structure analysis.

In real-world cases, users with more connections or number of re-twees do not have necessarily to be the more influential individuals. Betweenness Centrality is based on this idea, and it incorporates the importance of the neighbours to take into account the relevance of the friends. Using this metric, it can be identified the most influential person talking about vaccines from Twitter, as shown in Figure 6. The community that includes

the most important users based on Degree Centrality (1) still includes the largest number of influential users. But within this community, new influential individuals appear such as Shakira, who is a famous public personality. The most influential users of communities 4 and 7 (CMichaelGibson and AmerAcadPeds) remain as so in this new analysis based on social influence. Regarding the communities against vaccination, Figure 6 shows that only one negative community (3) includes influential personalities.

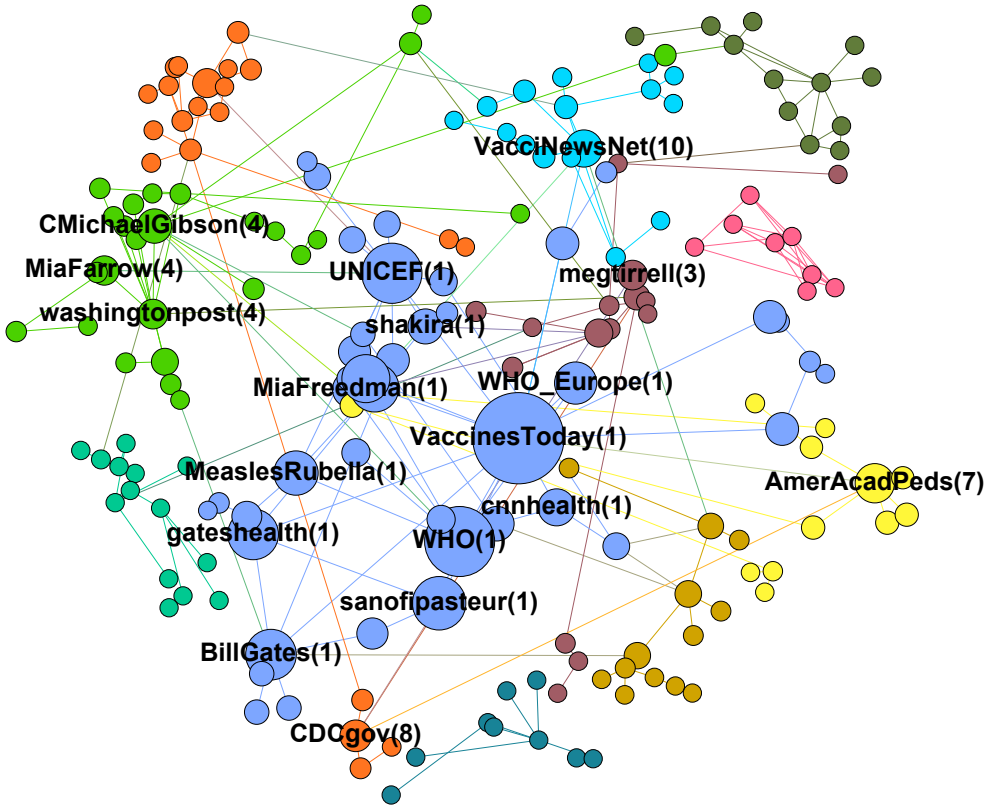


Figure 6: Vaccine Communities showing the **most influential** users and institutions based on the **Eigenvector Centrality** metric (node size according to its value). Node labels are filtered by a eigenvector value higher than 0,15. Top 5 users: VaccinesToday(1)(P), WHO(1), UNICEF(1)(P), BillGates(1)(P) and shakira(1)(P). Only one negative community (3) includes influential personalities.

To finalize the social influence study, the results of Betweenness Centrality measure have been used. This metric takes into account how important are

nodes connecting others (Figure 7 shows the results obtained). Users controlling the information flow can be identified using this information. Therefore, analysing the results shown in Figure 7, communities 1 and 4 include the largest number of users controlling the information flow. However, there are other communities that also include users with high values in this metric such as community 3 and 8. For example, a relevant health organization (CDCgov) belongs to community 8. In order to detect communities corresponding to negative discussion on vaccination, two negative communities (3 and 9) have been discovered. In addition community 3 has relevant users for this measure. As mentioned in different works on Social Networks [55, 20], this can be due to the effects of behaviour spread on social networks that are typically strongly content-dependent. Moreover, the negative sentiments are often contagious while positive sentiments are generally not.

A geographical visualization of the communities detected can help to analyse the results. Figure 8 shows a map summarizing the location information relating to communities. This map allows quickly identification of regions of interest (positive or negative) on vaccination. For this purpose, anti-vaccine communities are marked with red, while communities disseminating positive comments are shown in green. Analysing the results, it can be seen that four of the most relevant countries talking about vaccination (Ireland, United Kingdom, Canada, and Australia) mainly include positive communities. The European countries belonging to the block which show a moderate interest on vaccine topic such as France, Netherlands, or Sweden have only positive communities. On the other hand, most of the negative communities are located in EEUU, and as can be seen in the community detection results, these communities are relatively small and disconnected. Therefore, it can be concluded that strong communities supporting vaccination have emerged from the social networks.

Finally, considering all the results obtained for the different analysis performed, it can be concluded that the application of Communities Detection Algorithms is able to discover and track the discussion groups on vaccination arising by Twitter. In addition, the network structure analysis of the resulting communities allow to identify the most relevant users analysing their social influences, and their collective opinion, or sentiment, about the topic for each community.

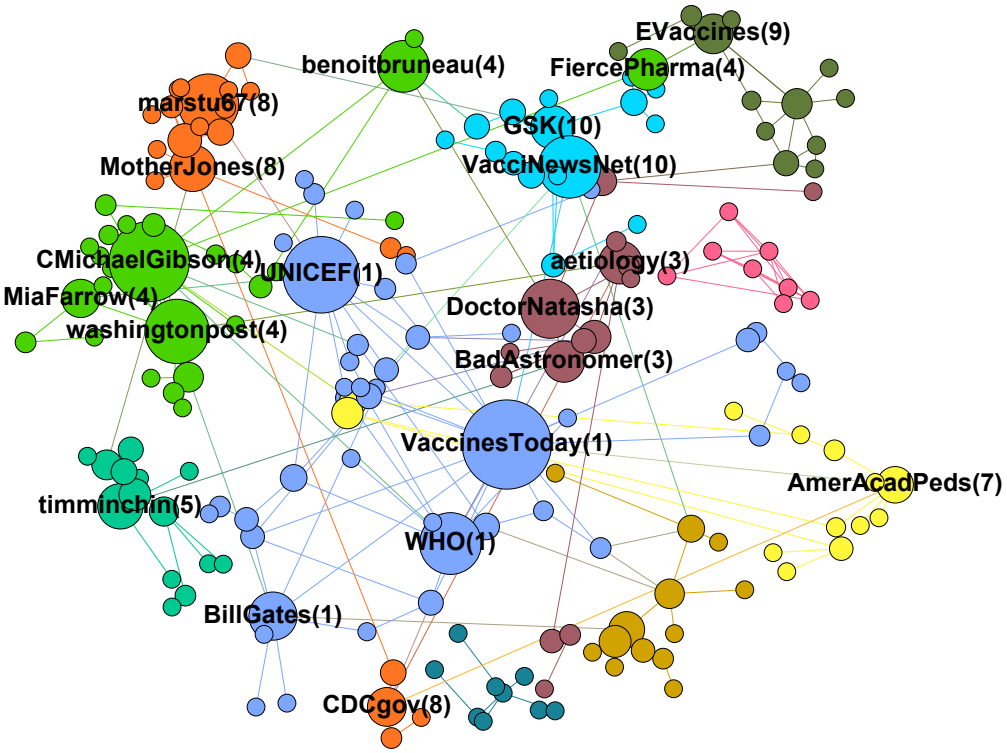


Figure 7: Vaccine Communities showing the users who **control the information flow** based on the **Betweenness Centrality** metric (node size according to its value). Node labels are filtered by a betweenness value higher than 1100. Top 5 users: VaccinesToday(1)(P), CMichaelGibson(4)(P), UNICEF(1)(P), washingtonpost(4)(P) and DoctorNatasha(3)(N).

5. Conclusions

This work shows a practical application of Data Mining Techniques to detect and analyse Twitter communities which are disseminating vaccination opinions. A dataset collected from Twitter, and the vaccination coverage rates retrieved from the immunization monitoring system of WHO, have been used to carry out several analysis. Using both datasets, an initial analysis is performed focused on measuring the potential influence of vaccine opinions based on the variation in the coverage rates. For this purpose two factors are used: Topic Relevance Factor (quantifying the relevance of vaccine topic in a given country) and Kurtosis of Vaccination Coverages (measuring the

Figure 8: Map summarizing the location information for the communities detected. Anti-vaccine communities are marked in red and positive communities in green. It can be seen that the most relevant anti-vaccines communities appear in EEUU.

distribution changes of vaccination coverages rates). Afterwards, generating a network representation of the Twitter dataset, Community Detection Algorithms have been applied to identify groups of similar users opining about vaccines. Finally, several centrality network metrics have been used to study these communities, discovering the most relevant users and analysing their social influence.

The results obtained in this preliminary analysis show that vaccine opinions from Twitter users could affect the vaccination decision-making process in some cases. However, it can be noticed that most of communities discussion on vaccination from Twitter are not against vaccines. In fact, currently most of the emerged movements are supporting vaccination and trying to increase the coverages rates.

The second part of the work is focused on the application of Community Detection Algorithms in order to discover communities opining about vaccines. The results obtained show that the most important and influential users belong to communities supporting vaccination movement, whereas negative vaccine communities often include few users that are not well connected. In addition, a geographical visualization of these communities shows that the

most relevant countries (Ireland, United Kingdom, Canada, and Australia) talking about vaccination are filled with positive communities. On the other hand, most of the communities disseminating negative opinions on vaccination are located in EEUU.

Taking into account all the experimental results presented, it can be concluded that the data mining techniques applied are useful for this kind of analysis. The methodology proposed can be used to find and track vaccine movements, discovering new knowledge in data that could be useful to improve Public Healthcare immunization strategies. Moreover, this new acquired knowledge could also be used to detect and locate communities against vaccination that could generate future disease outbreaks in different parts of the world.

6. Acknowledgements

This work has been supported by following research grants: Comunidad Autonoma de Madrid, under CIBERDINE S2013/ICE-3095 project, and EphemeCH (TIN2014-56494-C4-4-P) project, under Spanish Ministry of Economy and Competitivity, both supported by the European Regional Development Fund FEDER.

7. References

[1] C. for Disease Control, P. (CDC, et al., Impact of vaccines universally recommended for children–united states, 1990-1998., MMWR. Morbidity and mortality weekly report 48 (12) (1999) 243.

[2] V. A. Jansen, N. Stollenwerk, H. J. Jensen, M. Ramsay, W. Edmunds, C. Rhodes, Measles outbreaks in a population with declining vaccine uptake, Science 301 (5634) (2003) 804–804.

[3] D. J. Opel, S. B. Omer, Measles, mandates, and making vaccination the default option, JAMA pediatrics.

[4] K. S. Wagner, J. M. White, I. Lucenko, D. Mercer, N. S. Crowcroft, S. Neal, A. Efstratiou, D. S. Network, et al., Diphtheria in the postepidemic period, europe, 2000–2009, Emerging infectious diseases 18 (2) (2012) 217.

[5] A. Kata, A postmodern pandora's box: Anti-vaccination misinformation on the internet, Vaccine 28 (7) (2010) 1709–1716.

[6] J. Keelan, V. Pavri-Garcia, G. Tomlinson, K. Wilson, Youtube as a source of information on immunization: a content analysis, jama 298 (21) (2007) 2481–2484.

[7] J. Keelan, V. Pavri, R. Balakrishnan, K. Wilson, An analysis of the human papilloma virus vaccine debate on myspace blogs, Vaccine 28 (6) (2010) 1535–1540.

[8] N. Seeman, A. Ing, C. Rizo, Assessing and responding in real time to online anti-vaccine sentiment during a flu pandemic, Healthc Q 13 (Sp) (2010) 8–15.

[9] N. Sunday, The online health care revolution: How the web helps americans take better care of themselves, Pew Internet & American Life Project.

[10] Twitter web site, twitter.com (2013).

[11] G. Bello-Orgaz, J. J. Jung, D. Camacho, Social big data: Recent achievements and new challenges, Information Fusion 28 (2016) 45–59.

[12] W. Chen, C. Wang, Y. Wang, Scalable influence maximization for prevalent viral marketing in large-scale social networks, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2010, pp. 1029–1038.

[13] G. Bello-Orgaz, H. Menéndez, S. Okazaki, D. Camacho, Combining social-based data mining techniques to extract collective trends from twitter, Malaysian Journal of Computer Science 27 (2) (2014) 95–111.

[14] S. Asur, B. A. Huberman, Predicting the future with social media, in: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, Vol. 1, IEEE, 2010, pp. 492–499.

[15] N. Collier, Uncovering text mining: A survey of current work on web-based epidemic intelligence, Global public health 7 (7) (2012) 731–749.

[16] I. Batal, H. Valizadegan, G. F. Cooper, M. Hauskrecht, A temporal pattern mining approach for classifying electronic health record data, ACM Trans. Intell. Syst. Technol. 4 (4) (2013) 63:1–63:22. doi:10.1145/2508037.2508044.
URL http://doi.acm.org/10.1145/2508037.2508044

[17] G. Bello-Orgaz, J. Hernandez-Castro, D. Camacho, A survey of social web mining applications for disease outbreak detection, in: Intelligent Distributed Computing VIII, Springer International Publishing, 2015, pp. 345–356.

[18] S. Brien, N. Naderi, A. Shaban-Nejad, L. Mondor, D. Kroemker, D. L. Buckeridge, Vaccine attitude surveillance using semantic analysis: constructing a semantically annotated corpus, in: Proceedings of the 22nd international conference on World Wide Web companion, International World Wide Web Conferences Steering Committee, 2013, pp. 683–686.

[19] H. J. Larson, D. Smith, P. Paterson, M. Cumming, E. Eckersberger, C. C. Freifeld, I. Ghinai, C. Jarrett, L. Paushter, J. S. Brownstein, et al., Measuring vaccine confidence: analysis of data obtained by a media surveillance system used to analyse public concerns about vaccines, The Lancet infectious diseases 13 (7) (2013) 606–613.

[20] M. Salathé, D. Q. Vu, S. Khandelwal, D. R. Hunter, The dynamics of health behavior sentiments on a large online social network, EPJ Data Science 2 (1) (2013) 1–12.

[21] F. Santo, Community detection in graphs, Physics Reports 486 (3-5) (2010) 75 – 174. doi:DOI: 10.1016/j.physrep.2009.11.002.

[22] A. B. Bloch, W. A. Orenstein, H. C. Stetler, S. G. Wassilak, R. W. Amler, K. J. Bart, C. D. Kirby, A. R. Hinman, Health impact of measles vaccination in the united states, Pediatrics 76 (4) (1985) 524–532.

[23] A. J. Wakefield, S. H. Murch, A. Anthony, J. Linnell, D. Casson, M. Malik, M. Berelowitz, A. P. Dhillon, M. A. Thomson, P. Harvey, et al., Retracted: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children, The Lancet 351 (9103) (1998) 637–641.

[24] F. Godlee, J. Smith, H. Marcovitch, Wakefields article linking mmr vaccine and autism was fraudulent, BMJ 342.

[25] J. John Thomas, M. LLM, paranoia strikes deep*: Mmr vaccine and autism, Psychiatric Times 27 (3).

[26] H. J. Larson, D. L. Heymann, Public health response to influenza a (h1n1) as an opportunity to build public trust, Jama 303 (3) (2010) 271–272.

[27] J. R. Kaufmann, H. Feldbaum, Diplomacy and the polio immunization boycott in northern nigeria, Health Affairs 28 (4) (2009) 1091–1101.

[28] H. J. Larson, I. Ghinai, Lessons from polio eradication, Nature 473 (7348) (2011) 446–447.

[29] T. Botsis, M. D. Nguyen, E. J. Woo, M. Markatou, R. Ball, Text mining for the vaccine adverse event reporting system: medical text classification using informative feature selection, Journal of the American Medical Informatics Association 18 (5) (2011) 631–638. doi:10.1136/amiajnl-2010-000022.

[30] S. Xia, J. Liu, A computational approach to characterizing the impact of social influence on individuals vaccination decision making, PloS one 8 (4) (2013) e60373.

[31] L. Shaw, W. Spears, L. Billings, P. Maxim, Effective vaccination policies, Information Sciences 180 (19) (2010) 3728 – 3744. doi:http://dx.doi.org/10.1016/j.ins.2010.06.005.

[32] A. Clauset, Finding local community structure in networks, Phys. Rev. E 72 (2005) 026132. doi:10.1103/PhysRevE.72.026132.
URL `http://link.aps.org/doi/10.1103/PhysRevE.72.026132`

[33] A. K. Jain, R. C. Dubes, Algorithms for clustering data, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[34] S. S. Elisa, Survey: Graph clustering, Comput. Sci. Rev. 1 (1) (2007) 27–64. doi:10.1016/j.cosrev.2007.05.001.
URL `http://dx.doi.org/10.1016/j.cosrev.2007.05.001`

[35] H. D. Menéndez, D. F. Barrero, D. Camacho, A genetic graph-based approach for partitional clustering, International journal of neural systems 24 (03).

[36] M. Girvan, M. E. J. Newman, Community structure in social and biological networks, Proceedings of the National Academy of Sciences 99 (12) (2002) 7821–7826.

[37] C. G. Wang Xutao, L. Hongtao, A very fast algorithm for detecting community structures in complex networks, Physica A: Statistical Mechanics and its Applications 384 (2) (2007) 667–674.

[38] A. Clauset, M. E. Newman, C. Moore, Finding community structure in very large networks, Physical review E 70 (6) (2004) 066111.

[39] M. E. Newman, Finding community structure in networks using the eigenvectors of matrices, Physical review E 74 (3) (2006) 036104.

[40] M. E. J. Newman, Fast algorithm for detecting community structure in networks, Physical Review E 69 (6) (2004) 066133+. doi:10.1103/physreve.69.066133.
URL http://dx.doi.org/10.1103/physreve.69.066133

[41] P. G. Sun, L. Gao, Y. Yang, Maximizing modularity intensity for community partition and evolution, Information Sciences 236 (2013) 83 – 92. doi:http://dx.doi.org/10.1016/j.ins.2013.02.032.

[42] World health organization web site, http://www.who.int/en/ (2013).

[43] Statista inc. web site, http://www.statista.com/.

[44] Internet live stats, http://www.internetlivestats.com/internet-users-by-country/ (2013).

[45] L. T. DeCarlo, On the meaning and use of kurtosis., Psychological methods 2 (3) (1997) 292.

[46] R. A. Fisher, The moments of the distribution for normal samples of measures of departure from normality, in: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, Vol. 130, The Royal Society, 1930, pp. 16–28.

[47] M. Hollander, D. A. Wolfe, E. Chicken, Nonparametric statistical methods, John Wiley & Sons, 2013.

[48] J. H. Zar, Significance testing of the spearman rank correlation coefficient, Journal of the American Statistical Association 67 (339) (1972) 578–580.

[49] R. Zafarani, M. A. Abbasi, H. Liu, Social media mining: an introduction, Cambridge University Press, 2014.

[50] M. Rosvall, D. Axelsson, C. T. Bergstrom, The map equation, The European Physical Journal-Special Topics 178 (1) (2009) 13–23.

[51] U. N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Physical Review E 76 (3) (2007) 036106.

[52] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment 2008 (10) (2008) P10008.

[53] P. Pons, M. Latapy, Computing communities in large networks using random walks, in: Computer and Information Sciences-ISCIS 2005, Springer, 2005, pp. 284–293.

[54] M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) (2004) 026113. doi:10.1103/PhysRevE.69.026113.
URL http://link.aps.org/doi/10.1103/PhysRevE.69.026113

[55] S. Wasserman, J. Galaskiewicz, Advances in social network analysis: Research in the social and behavioral sciences, Vol. 171, Sage Publications, 1994.

[56] D. R. White, F. Harary, The cohesiveness of blocks in social networks: Node connectivity and conditional density, Sociological Methodology 31 (1) (2001) 305–359.

[57] R. D. Alba, A graph-theoretic definition of a sociometric clique, Journal of Mathematical Sociology 3 (1) (1973) 113–126.

*Biographies (Text)

Gema Bello-Orgaz is a teaching assistant in Universidad Autonoma de Madrid. She has a BSc in Computer Science from Universidad Carlos III de Madrid, and a MSc in Computer Science from Universidad Autonoma de Madrid (2012). Nowadays, she is a Computer Science PhD candidate at Escuela Politecnica Superior (UAM). She is involved with AIDA interest research group at EPS-UAM, her main research interests are related to Clustering, Graph-based algorithms, Social Data Analysis, and Evolutionary Computation.

Julio Hernandez-Castro is a Senior Lecturer in Computer Security at the University of Kent's School of Computing. His research interests are wide, covering from RFID Security to Lightweight Cryptography, and including Steganography and Steganalysis and the design and analysis of CAPTCHAs, to name only a few. He worked before for the University of Portsmouth and Carlos III University in Madrid, Spain. He has been a pre-doctoral Marie Curie fellow and a postdoctoral INRIA fellow. He is also affiliated with the Cybersecurity Center of Kent's University. He is currently the vice-chair of the EU COST project CRYPTACUS. He receives research funding from InnovateUK project aS, EPSRC Project 13375, and EU H2020 project RAMSES.

David Camacho is currently working as Associate Professor in the Computer Science Department at Universidad Autonoma de Madrid (Spain) and Head of the Applied Intelligence & Data Analysis group. He received a Ph.D. in Computer Science (2001) from Universidad Carlos III de Madrid, and a B.S. in Physics (1994) from Universidad Complutense de Madrid. He has published more than 200 journals, books, and conference papers. His research interests includes Data Mining (Clustering), Evolutionary Computation (GA \& GP), Multi- • Agent Systems and Swarm Intelligence (Ant colonies), Automated Planning and Machine Learning, or Video games among others.

jchernandezcastro.jpg

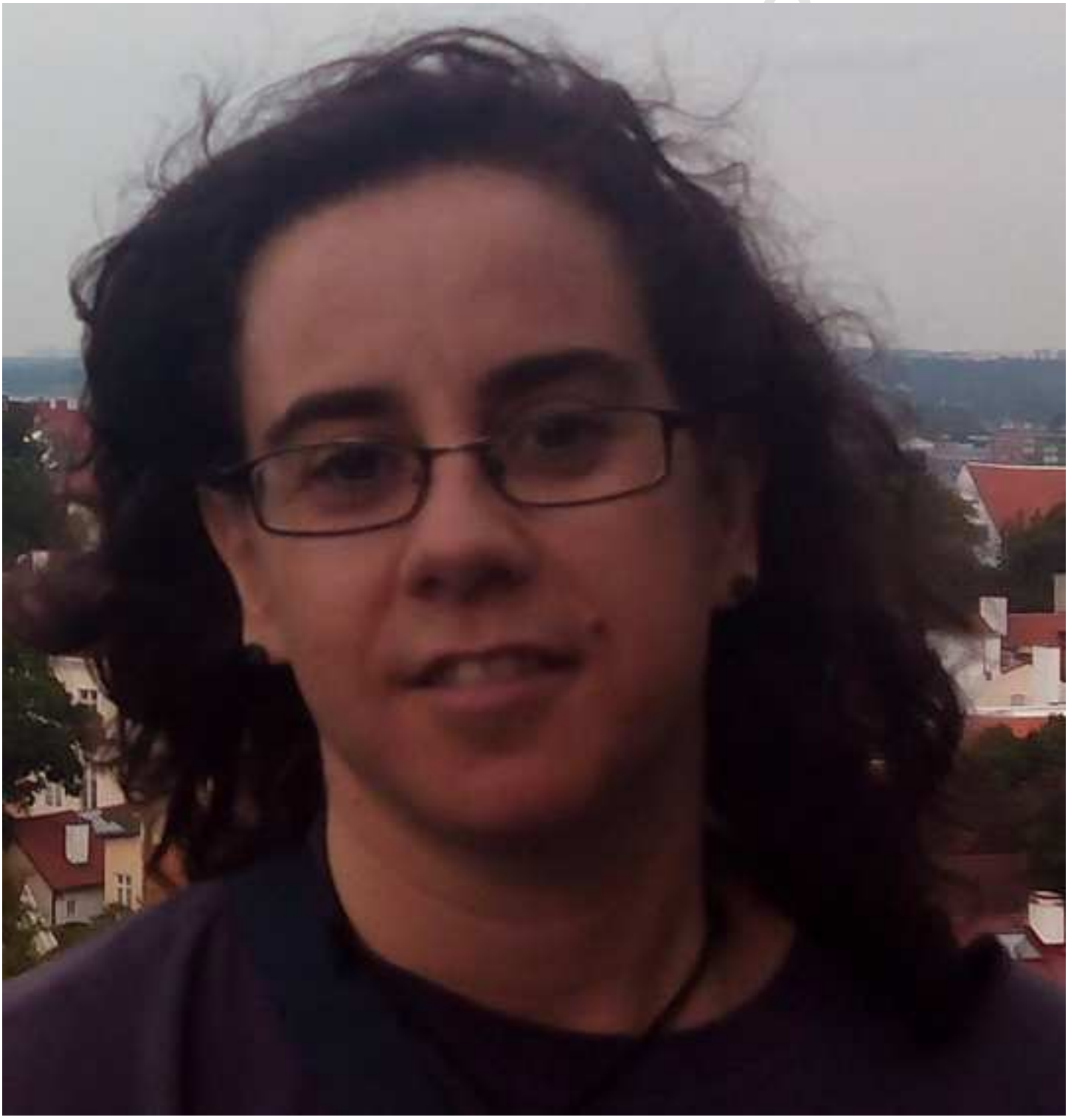
Click here to download high resolution image

davidbio.jpg

Click here to download high resolution image

gemabio.jpg

Click here to download high resolution image



Highlights

- A methodology to detect discussion communities on vaccination is proposed.
- Vaccine opinions in twitter can affect the decision-making about vaccination.
- The most relevant and influential users are identified analyzing the communities.
- The collective sentiment on vaccination has been studied for the detected groups.
- Results provide useful information to improve immunization strategies.