Detection and Reinforcement of Celiac Communities on Twitter Argentina

Andrés Giordano, Santiago Banchero, Natacha Cerny,
Mauricio De Marzi, and Gabriel Tolosa

Universidad Nacional de Luján, Departamento de Ciencias Básicas, Argentina
{agiordano, sbanchero, ncerny, mdemarzi, tolosoft}@unlu.edu.ar

Abstract. Social Networks have shown great growth relating the number of their users and generated content. For example, Twitter is used as a means to gather support, express ideas and opinions on various topics or interact with users with similar interests. In the latter case, the idea of community formation appears, that is, groups of users that are more closely related to each other than the rest of the nodes in the network.
In this work we propose the detection of the community of users of Argentina interested in the celiac disease. We apply a series of techniques to detect and characterize them. In addition, we propose and use a methodology for the detection of more influential and active nodes (users), showing how the community can be reinforced by the recommendation of some *particular* links. The results show that with only a low percentage of accepted recommendation the network becomes denser and average distance between two users decreases quickly, thus improving the spread of information.

1 Introduction

Social networks have shown great growth in terms of the number of users and content generated, mainly in the last few years. A clear example is Twitter, in which not only users publish their activities but, in some cases, it is used as a means to gather support, express ideas and opinions on various topics or connect and interact with users with similar interests.

Starting with this dynamic, the forms of communication have expanded, generating patterns of union and behaviour among users who have emerging properties that are of interest to know to understand their scope and effectiveness. These relationships, which occur both in nature and in social phenomena, can be represented and analyzed in terms of a network, or formally, a graph. In general, at a macroscopic scale, these networks offer some degree of organization [28].

One of these phenomena is the formation of communities in social networks. People tend to group instinctively in the digital world as in the real world, with others with whom they share ideas, tastes, hobbies, etc., which facilitates communication. Although there is no global and unique definition of what a community is, it can be defined as a set of people that interact in time with an

objective, interest or need [33]. Regarding the analysis of the underlying network, these are groups of nodes that are more closely related to each other than to the rest of the nodes.

There are implicit and explicit communities [30]. In the first case, a community is formed by the daily interactions of a group of users which are not always seen by everyone (for example, user posts on a topic on Twitter, with its group of followers). In the second case, explicit communities are those in which users make a conscious decision to participate in a group, they can know the group of members of the same and the scope of their publications (for example, a closed group of Facebook on a subject particular).

In the latter case, the community is clearly delimited and the analysis of interactions is relatively simple. However, the identification of implicit communities in social networks is a slightly more complex task, whose result is not exact and that can provide useful information about the dynamics and behaviour of groups of users with common interests with different objectives, for example, provide services related to your interest.

Communication patterns tend to be more intense among members of the same group, with respect to others. These follow the sociological principle known as *homofilia* which proposes that people have to relate to a greater extent with similar pairs by some characteristic (age, education, religion, among others). Another important relation that appears is the *influence*, in which some members of a group develop similar ideas or visions about some concept following the opinion of one or several of its members [7].

Community detection is a relevant problem in the world of social network analysis or, more broadly, in Network Science[1]. On the one hand, it allows the identification of non-trivial relationships between members of the network and their self-organization and, on the other, helps to understand the processes that take place for their formation and dynamics [30, 33]

In addition, not all users who connect with each other share the same interests [22], so considering only the structure of links in the network may be an incomplete criterion or one that does not apply to all cases. Therefore, the similarity between users through the content of their publications and other data such as location, sex or age can also help determine membership of the same group or community.

The ability to detect communities in a social network has practical implications in multiple domains. In this work we propose the detection of a particular community in the graph formed by argentinian users of Twitter that are interested in celiac disease, as a complement of epidemiological studies[2] [6].

Celiac disease is the most frequent chronic intestinal disease in Argentina, characterized by a permanent intolerance to gluten (protein found in wheat,

[1] Network Science is a relatively new field of research that studies complex systems and their representation as networks of both natural and social phenomena, trying to obtain predictive models of the behaviour of their actors.

[2] This work is related to an interdisciplinary project whose main objective is to characterize the incidence of celiac disease and its relationship with related pathologies.

oats, barley and rye) that occurs in genetically predisposed people. This disease interferes with the absorption of nutrients by damaging part of the small intestine and is linked to other pathologies such as thyroid disorders, osteoporosis, infertility, diabetes, among others. Although the exact cause of celiac disease is unknown, it is known that environmental, genetic and immunological factors intervene in its pathogenesis; which makes accurate diagnosis difficult. Although there are still no epidemiological records, preliminary studies in our country indicate a prevalence of approximately 1:200. However, it is currently estimated that 1 in 100 people is celiac[3].

In recent years, case finding strategies have been developed in order to identify those people that belong to the "risk groups", such as: relatives of those affected by celiac disease, people with autoimmune diseases or with symptoms that could indicate a celiac disease such as limited growth, persistent bowel problems, anemia, etc. It is necessary to be more aware of this "chameleonic" pathology, as well as to conceive possible strategies for carrying out mass tests in order to extract the iceberg of celiac disease as much as possible, that is, the multitude of cases not diagnosed In this sense, consultation in social networks is a good approximation of the interests in the society under study.

This pathology has an impact on the everyday life of patients, including their social life, mainly in the formation of informal social capital, that is, contact with friends, family members, colleagues [36, 39]. Social networks collaborate with the maintenance of certain digital social capital. On the one hand, facilitating and expanding communication with other people and, on the other, allowing the exchange of information. In this case, they become powerful tools to obtain, generate and propagate sensitive information related to the celiac disease that can be of help to others, from indications to obtain gluten-free foods, recipes to discussions about signs, symptoms and diagnoses in each case. Many times, sharing experiences opens new perspectives to those who suffer from diseases.

1.1 Motivation and Main Goals

The relationship between social networks and pathological behaviours in groups of people is a topic of interest [40]. However, no prior work has been found related to celiac disease and its impact on a digital life of groups of people. Taking into account the growth of social networks, the intensity of participation of its users and the possibilities offered by being able to massively study groups of users almost in real time, it is of particular interest to generate methodologies and specific studies that support other disciplines in to characterize from a different perspective a human phenomenon, as a particular pathology.

The main objective of this work is to detect potential Argentine users with an interest in celiac disease (patient/family/friend) and suggest links that reinforce communities according to this interest to facilitate the exchange of valuable information in the context. In particular, the main contributions of this work are:

[3] https://www.argentina.gob.ar/salud/glosario/enfermedadceliaca

- The detection of communities of users interested in celiac disease combining both the relationships (*links*) and the content of their publications. Both approaches are combined showing an improvement in the final accuracy.
- The use of a clustering technique combined with the search for *certain* users of interest to determine the cluster that represents the community. It shows how the accuracy in the identified community varies.
- The identification of the most influential and active users in the community and this metric is used to recommend links to subgroups of users. It shows how the community becomes denser as links are accepted, which reinforces the propagation of internal information.
- The recommendation of users based on a combination of two metrics. We show that the seleccion of the most influential users together the most active ones (interested on the target topic) becomes a useful metric to recommend links to subgroups of users. We show how the community becomes denser as new links are accepted, which reinforces the propagation of internal information.

The rest of the work is organized as follows: Section 2 introduces some works highly related to ours, while Section 3 introduces the necessary basic concepts of the context of the work. Then, a methodology is proposed in Section 4 that determines the experiments (and their results) in Section 5. Finally, conclusions and future work are presented (Section 6).

2 Related Work

There are several works that address the problem of community detection on digital social networks [2, 14, 31]. In general, these make use of the existing connections within the network.

On the contrary, the vertical search of communities is a more complex task. Using only the existing connections among users is not enough to achieve a good performance, so it is necessary to explore the contents of the publications. This information is necessary to determine the topics of interest of a set of users.

The detection of topic-oriented communities using a combination of grouping techniques and link analysis has been addressed in the past [44, 43]. In a first stage, these techniques group objects into thematic groups using the *Entropy Weighting K-Means* [20] algorithm. Then, link analysis is carried out within each thematic group, using the modularity metric in order to detect potential communities that already exist for each topic.

The semantic search of communities is another approach to face the problem. There are techniques that use the contents within the network such as Latent Dirichlet Allocation (LDA) [3]. The model *Link-Block-Topic* uses LDA and performs the detection of thematic communities without the need to indicate the number of communities to look for or the size of them [42].

Using a matrix-based approach, Guo et al. [17] build a dissimilarity distance matrix of the network to identify community centers. They first estimate the

distance between all pairs of nodes and then use an affinity propagation algorithm to extract a candidate center set of community (they call the method as CDMIC). The evaluation on three real-world networks and some synthetic ones shows that CDMIC has higher performance in terms of classification accuracy and normalized mutual information.

Another way to solve this problem is mining user interactions to discover such communities. Correa et. al. [38] presented an algorithm called iTop to discover interaction based topic centric communities by mining user interaction signals (@-messages and retweets) which indicate cohesion. iTop takes any topic as an input keyword and exploits local information to infer global topic-centric communities.

Other works, based on local approaches, link themes based on identifying users with a large number of followers, considering that the selected users are representative of a category of interest in which they make the most publications [13]. These techniques then use an overlap calculation, between the followers of referents and the communities of the network through *Clique Percolation Method* (CPM) [10].

In addition, Yang et al. introduce CESNA [41] (*Communities from Edge Structure and Node Attributes*), a method that uses a probability model based on Bernoulli distributions. Here, the membership of a community is combined with the structure of the network and the attributes of the nodes from the model. This solution is based on the assumption that vertices are more likely to be neighbours as more communities share them. Although the CESSNA algorithm has a linear runtime with the size of the network, the interpretation of the results is not good enough [5].

3 Preliminaries

The underlying model of a social network corresponds to a graph $G = (V, E)$, where V is the set of nodes or vertices that represent the users of the social network and E is the set of edges that represent the relationships between the users. Considering Facebook as an example, if user $u \in V$ is a friend of user $v \in V$ then there is an edge $(u, v) \in E$ that represents the simetric relationship between them. Otherwise, in the case of Twitter the edge $(u, v) \in E$ represents the relationship u *follows* v but not the reverse way (*followers vs followings*). Thus, if v also follows u, then there exists an edge $(v, u) \in E$ (this denotes the directed nature of this graph). At the same, the intensity of the relationship is given by the weight of the edge ($w_{u,v}$), calculated according to some metric related to the user (for example, the number of retweets the follower user makes).

3.1 Communities

As previously mentioned, there is no single definition for the concept of community, but there is a common feature among all of them: a community is composed of users who have a subject or topic of common interest.

There are communities composed of users who periodically publish news, anecdotes or generate talks or discussions on some specific topic according to their interests. That is, the users of a given community have a high degree of interaction with each other.

However, there are communities clearly distinguished by their "explicit relationships" within the network. That is, there are "*followers*" and "*followings*" in the case of Twitter or "friends" on Facebook. In this case, it may (or may not) exists a high interaction between users but there must exist a high density of intra-community links.

Even more, there are communities whose only link is a common interest, without the existence of an explicit relationship or interaction within the social network. An analysis of the contents of the publications is necessary to find the features that bring them closer to a specific topic and, thus, detect the community.

3.2 Community Detection

There are different methods to detect communities, which are more or less appropriate according to the type of community, the patterns of interaction among their users or the portion of the social network explored. For example, in the case of interactive communities like Twitter, it is necessary to collect all types of interaction among users (posts, retweets, mentions and comments). Lim [25] proposes to generate a graph where the relations are the mentions between the users and then to apply a community detection algorithm on the generated structure. Consequentely, community detection algorithms are classified as:

- **Topology-based:** these methods are based only on the analysis of the graph underlying the network. That is, the algorithms analyses the structure of relationships among users [4, 24, 34]. While algorithms that apply this approach are effective, grouping users who are interested in different topics, although densely connected, lead to a lack of high precision.
- **Content-based:** this approach explores the contents of the publications of the users and does not consider the structural information of the network (i.e., the density of connections that may exist in a set of users). On Twitter, for example, this refers to the contents of the tweets after separating free-text from hashtags, URLs and mentions [25].
- **Hybrid:** this approach merges together the two previous ones [21, 35, 38, 43]. Basically, the graph of structural relationships among users if built first and then, some features based on content similarity are added. For example, contents similarity may be used as a weight (or importance) of the relationship between two users. Once this structure is generated, some detection algorithm that considers the weight of the edges is applied.

3.3 Algorithms

Below we describe the two specific algorithms for community detection used in this work. The first one, known as the Louvain method [4], is based on the opti-

mization of the modularity of the partitions obtained as the algorithm progresses in its execution, in a greedy way. On the other hand, the Infomap method [34] is based on the information theory for representing the communities.

Louvain Method: This approach tries to maximize the modularity of the graph as the nodes are grouped into communities. It is robust and efficient since it has been used and revised in several works [23, 24] and new community detection algorithms are based on this [8, 9, 15, 32]. Its complexity is $O(n \log n)$.

Modularity is established in order to assess the quality of the partitions [28] and thus, it has been widely used for this purpose [2, 37] as a measure of the quality of the resulting communities. This metric is defined as:

$$Q(G) = \frac{1}{2m} \sum_{l=1}^{K} \left(\sum_{i \in C_l, j \in C_l} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \right) \tag{1}$$

where:

K is the number of communities,
A_{ij}, the weight of the edge between i y j,
k_i is the sum of the weigths of the edges incident to i,
C_l is the community where i and j are asigned,
$m = \frac{1}{2} \sum_{i,j \in V} A_{ij}$
$\frac{1}{2m}$ is used as a normalization factor (between -1 and 1).

Then, the algorithm groups the nodes of G in two steps that are repeated at each iteration:

1. **Modularity optimization**
   (a) Assign each node to a different community.
   (b) For each node i, process all its neighbours j by calculating the modularity gain of moving node i to the community of j. Then, i is moved to the community whose profit it is the maximum if and only if the gain is positive.
   (c) Repeat step (b) until reaching a maximum of local modularity, that is, when there are not any more movements between communities.
2. **Community Aggregation:**
   In this phase the algorithm builds a new network ($G_a = (V_a, E_a)$) whose nodes are now the communities found in the previous step. That is, for each community $c_i \in C$ of the previous step is a new node n_{c_i} is added to the new network. The weights of the links between the new nodes are given by the sum of the weight of the links between nodes in the two communities. In case of links between nodes of the same community, self-loops are generated. Specifically,
   (a) For every $n_{c_i} \in V_a$ a self-loop is added with weight equal to the size of the subgraph formed by nodes in c_i.

(b) For every $n_{c_i}, n_{c_j} \in V_a$ where $n_{c_i} \neq n_{c_j}$, a new link (n_{c_i}, n_{c_j}) is added with weight equal to the number of links between nodes in c_i and c_j.

In case of a weighted graph, the sum of weights is used instead of the number of links. This step allows us to set up a cut parameter in order to find bigger or smaller communities.

Infomap Method: This approach relies on information theory to represent communities, using a code that must be as shortest as possible. Basically, it proposes to represent a random walk on a graph in an effective and compact way. To this aim, it uses two levels of description based on Huffman codes [19]:

1. The first level establishes a unique code for each intra-community node whose length is inversely proportional to the number of times that a node was visited during the walk.
2. The second level defines codes in the same way. In this case, to identify the different communities.

Then, the problem of finding the best partition of the graph in groups of users (or communities) is expressed as the problem of finding the minimum amount of information required to represent the random walk using the levels of the above description.

Huffman code is designed to assign short codes to more frequent symbols in a given language (and vice versa). It is expected that the walker stays for a long time within each community visiting several times the same nodes since the number of intra-community links is greater than those that link nodes in different communities.

Thus, within each community, it is possible to generate an optimal code to represent each node that only needs a couple of extra codes to indicate that the walker entered or left a particular community. This representation allows us to express the entire journey in the minimum number of the possible code. This method has been widely used [11, 24], even contributing in other areas such as biology [12].

Clustering: Another technique for community detection is to apply classic clustering algorithms based on user features, such as the well-know *Kmeans* algorithm. This method is widely used by the scientific community in different areas of computing such as image processing, data mining, among others [18, 27]. The approach used by this algorithm to identify k clusters relies on assigning each instance (user or, generally speaking, item) to the cluster whose centroid (centre of mass) is the closest. To this aim, it is required to represent each individual using a vector of characteristics (or feature-vector). For example, processing publications and constructing the frequency distribution of the terms a given person uses. Optionally, a topic detection algorithm such as LDA (Latent Dirichlet Allocation) could be applied in which the frequency of a published topic is accumulated.

3.4 Community Reinforcement

Strategies for community reinforcement are aimed to establish a greater number of intra-community links, resulting in a denser structure. The main idea is based on link suggestions (or recommendations)[4] to users. As soon as they accept new links, the graph becomes denser, which enhance its ability to spread information.

In general, these methods try to estimate the probability that a link between users a and b is established in the near future. To offer a recommendation, some links that maximize a given metric are selected.

In this work, we propose to combine two features of the nodes (users): their influence and their activity (Section 4.5). That is, given the objective of the community, we prefer users who can disseminate information quickly (influencers) but do it periodically (actives).

3.5 Metrics

In this section, we describe the metrics used for the analysis that basically correspond to measurements on the graph $G = \langle V, E \rangle$ or its nodes.

Diameter $(D(G))$**:** The distance between two vertices $(u, v \in G)$ is defined as the shortest path length between them. Then, the diameter of G is the maximum distance between all pairs of nodes.

Closeness $(C(u))$**:** The metric *Closeness* of any node $u \in G$, tries to quantify how close u is to the other nodes of G. It is defined as the inverse of the sum of the distances of u to all other vertices v, $C(u) = \frac{1}{\sum_{v \in V} d(v,u)}$.

Clustering Coefficient (CC)**:** The CC of a vertex $u \in G$ quantifies how much it is grouped or interconnected with its neighbors. It corresponds to the ratio between the links connected to their neighbors (e_{ij}) and the number of existing links in a click (maximum connectivity). It is defined as $C_i = \frac{|e_{ij}|}{k_i(k_i-1)}$. Then the Average CC (CCP) of G is $\frac{1}{n}\sum_i^n C_i$.

4 Methodology

In order to identify the target community, we start with a network topology-based approach. We sample publications (*tweets*) of users using the public Twitter API. Then, we build the corresponding directed graph based on the existing links between users. Finally, for each user in the graph, we analyze its membership (or not) to the community of interest.

4.1 Data Collection

Tweets were collected between April 20 and July 2, 2017 (74 days). For the positive identification of the publications, we use keywords related to the subject of our interest *celiaco, celiac, celiac, coeliac, celiaquia, celiaquía, sintacc, tacc,*

[4] For example, on Facebook a list of " People you may know "

gluten, 'sin gluten', gluten-free[5]. We collect 131,550 publications with a total of 76,233 unique users.

Filtering by Location: Given that our main objective is the detection of a community of celiacs on Twitter in Argentina, tweets were filtered to obtain only those published by Argentine users. This task was done in two different ways:

1. If the tweet is geolocated, the coordinates in the `'coordinates'` field of the tweet are taken and a reverse resolution was made using a map service.
2. Otherwise, the user's `'location'` field is analyzed and compared to a list of localities and provinces in Argentina.

4.2 Graph Generation

To build the graph, we only use Argentine users as nodes (U_{arg}), according to the location of their publications . For each user $u \in U_{arg}$ we get the set of users u follows (*friends*) and add the corresponding edge only for friends that also belongs to U_{arg} . The resulting structure is a directed graph $G = \langle V, E \rangle$ where each edge $(u_1, u_2) \in E$ represents the relation u_1 follows u_2.

Users that have no connections with others in the network (isolated nodes) are eliminated. Thus, the resulting graph (base graph or G_{base}) is comprised by 2,068 nodes and 20,675 edges.

4.3 Celiac Community Identification

We use a similar technique as reported by Lim [25], where the interest of the users in a specific topic is detected through the concept of *celebrities*.

Celebrities are users with more than n followers (where n is always a high number with respect to the remaining users) and it is known a priori that they have an interest in the specific topic (although it might not be the only interest of that celebrity).

In that work, the authors get the set of users that follow all celebrities. Then, they apply a community detection algorithm, thus verifying that users tend to follow "reputed" ones in the topic of interest.

However, the identification of celebrities related to "celiac disease" in Argentina exhibits very low numbers. Given that there are not enough Argentine users interested in this topic with a high number of followers, this requirement was eliminated. Five of the six Twitter accounts that were selected (Table 1) have been specifically created with the aim of sharing news or information about the subject. The remaining user corresponds to a person who describes himself as a celiac in his profile.

[5] The use of some English words responds to being detected as being used in some *hashtags*.

User	Description
@asoc_celiaca_ar	*The first one in Latin America. Offer help to people that need to follow a gluten-free diet*
@CeliacoCom	*Are you celiac? All you need is here!. Recipies, videos, interesting information about where to buy, eat and much more!!!*
@cocinaceliaca	*I am a high-cooking chef, specialist in celiac suitable food.*
@SoyCeliacoNoET	*Recipies, experiences, tips and information about celiac disease and gluten-free diet. #SinTACC #SinGluten#GlutenFree #Food*
@$anonUser_1$	*Daughter of a celiac person, celiac person and mother of a celiac person.*
@rojasglutenfree	*Supermermarket exclusive for celiac people.*

Table 1. Identified celebrities about "celiac disease". We replace the names of individual users ($anonUser_x$) to preserve anonymity. We only maintain usernames that identify institutions, associations or companies and do not individualize people.

Validation After executing each method, we perform a validation of the community based on expert assessment. We request a group of volunteer experts who judge whether the user is interested in celiac disease (or not) based on observing its public Twitter profile.

4.4 User Similarity

Content-based methods for community detection require some technique to compare users (instead of using links). A commonly used possibility is to calculate a similarity measure among users by taking their publications as representative of their interests.

To this aim, we concatenate the last n Tweets[6] of each user to build a single document [43]. We apply standard tokenization and normalization procedures: stopwords, URLs, numbers, punctuation, emoticons, arrows and tokens exceeding 30 characters are eliminated.

Finally, similarity among users is calculated based on the vector space model, a well-known and established approach used in Information Retrieval [26]. In this case, we use the cosine-similarity metric, defined as:

$$sim(d_u, d_s) = \frac{\overrightarrow{V}(d_u) \cdot \overrightarrow{V}(d_s)}{\left|\overrightarrow{V}(d_u)\right| \left|\overrightarrow{V}(d_s)\right|} \tag{2}$$

where $\overrightarrow{V}(d_n)$ is the vector of weights that corresponds to each document that represent users u and s, respectively. Denominator corresponds to the product of the norm of both vectors and it is used to normalize document lengths. To weight

[6] In this case, we are able to get the last 3200 tweets due to Twitter API limitations.

terms in each $\overrightarrow{V}(d_n)$ we use TF/IDF [1] approach. TF represents the normalized frequency of the term i in the user's j document, $TF = \frac{frec(i,j)}{max_f rec(j)}$. The IDF value corresponds to the inverse of the frequency in documents of the term in the collection, $IDF(t) = log(\frac{N}{df})$, where N is the total number of documents in the collection (in our case, the number of users) and *df* is the sum of the frequencies of the term in each document. After calculating the similarity between each pair of users, we use it as a weight or importance of the relationship between both.

4.5 Influencing and Active Users

Given the topic of interest (Celiac disease), our aim is to find a set of users that are both influential and active simultaneously. Influential users are those who redistribute contents generated by other users or have effects on the activities of their followers. In an analogous way, active users "talk" frequently about the target topic.

First, we start by generating separated rankings of influential and active users. To classify users according to their influence we use the method suggested by Cha [7]. A directed graph is generated (G_{infl}) where each node represents a user but the edges represent one of two possible relationships: *u retweets to v* and/or *u mentions v*. The weight of the relationship is given by the number of times each action happens. Then, we run PageRank [29] on G_{infl} as a metric of the importance of the nodes, obtaining the final list of the most influential users (ℓ_{infl}).

Then, to get the ranking of active users we start by analyzing the last n tweets of each user and compute the proportion p of terms that belong to the domain of the topic (Section 4.1). The resulting list of most active users regarding celiac disease (ℓ_{act}) is obtained sorting by the proportion of p.

Finally, we select a set of users after intersecting both lists ($\ell_{infl} \cap \ell_{act}$), at percentage p from the top. The objective of this procedure is to obtain a set of users to recommend to the remaining nodes as a means of community reinforcement.

5 Experiments and Results

We start using the base graph G_{base} to run community detection experiments. In a complementary way, we build two weighted versions of it.

- G_{base_w}: In this graph we weigh the edges according to the user similarity criterion (Section 4.4)
- $G_{base\_un\_w}$: In this case, we asume the edges as not directed, reflecting with more weight the symmetry of the similarity between users.

From now on, we execute all the experiments on the three graphs evaluating the results according to the proposed structural changes.

5.1 Community Detection

To identify the underlying communities, we start using the Louvain method on the three graphs. We vary the threshold parameter in the $[0.1; 1]$ range, increasing in 0.1 at each step. This process can be seen as the *height* at which a dendrogram is cut. When the cut-off value approaches 1, we obtain larger communities (lower resolution). Otherwise, when it approaches 0, the communities formed are smaller (higher resolution). This effect is related to the Clustering Coefficient (CC) of the target community and follows the idea that networks with underlying communities tend to have an average CC value (CCP) much higher than random networks with the same number of edges and nodes [37].

Finally, the selected threshold value is determined by the highest CC obtained in the celiac community. Then, that community is validated as specified in the section 4.3. Table 2 shows the results for each threshold value and graph. The values that lead to the best CCP are 0.3, 0.2 and 0.1 for the graphs G_{base}, G_{base_w} and $G_{base\_un\_w}$, respectively.

In the three cases, detected communities (U_{com}) have a high percentage of interested users (exceeding 65%) over the total number of individuals that form the group. Particularly, on the graph $G_{base\_un\_w}$ a greater accuracy (74.6%) is achieved at the cost of a decrease of 7.79% in the number of identified users (Table 3).

In a similar way, we run the Infomap method. To evaluate the variability of the resulting celiac community, we run 10 trials and compute the intersection over the union of the set of users in different executions. The results show that the community varies in only 1% verifying the consistency of this algorithm with regard to a random technique related to the degree of the nodes.

Table 4 shows the number and percentage of users interested in the topic within the celiac community found (U_{com}). The accuracy achieved on G_{base_w} and $G_{base\_un\_w}$ reaches 77% in both cases. Here, we observe a decreasing number of recovered users (8.73%) in the worst case, regarding G_{base}.

Threshold	Average CC		
	G_{base}	G_{base_w}	$G_{base\_un\_w}$
0.1	0.276	0.298	**0.523**
0.2	0.334	**0.299**	0.521
0.3	**0.358**	0.254	0.484
0.4	0.355	0.251	0.469
0.5	0.343	0.254	0.430
0.6	0.346	0.217	0.424
0.7	0.281	0.185	0.470
0.8	0.324	0.276	0.455
0.9	0.325	0.204	0.319
1.0	0.166	0.233	0.360

Table 2. Threshold value vs CCP for each graph (highest value in bold) using the Louvain method.

Graph	$\|U_{com}\|$	Interested	% Interested
G_{base}	104	68	65.4%
G_{base_w}	82	57	69.5%
$G_{base\_un\_w}$	71	53	74.6%

Table 3. Number of celiac-interested users found within the detected community U_{com} (Louvain method) for the three graphs.

Graph	$\|U_{com}\|$	Interested	% Interested
G_{base}	91	63	69.2%
G_{base_w}	71	55	77.4%
$G_{base\_un\_w}$	74	57	77.0%

Table 4. Number of celiac-interested users found within the detected community U_{com} (Infomap method) for the three graphs.

5.2 Clustering

As mentioned above, one possibility is to treat the formation of communities as a problem of document clustering. In this work, we use the widely used K-means method that requires to represent each item (user) as a feature vector. Before obtaining the vector for each user, the collection is preprocessed to reduce the data dimensionality (we can express this process as a pipeline).

1. **Preprocessing:** We remove symbols, mentions, numbers and stopwords. In addition, we also remove frequent terms (those that appear in more than 50% of the documents) but appear at least in five documents. Finally, we remove very short terms (less than 4 characters).
2. **Frequency-based Vectorization:** We vectorize all documents according to the term frequency of their terms. Then, we only keep the 10,000 most frequent terms.
3. **Topic Detection:** In order to reduce data dimensionality, we select the most important terms using a topic-based approach. To this aim, we use LDA (Latent Dirichlet Allocation) and set the number of topics to be detected in 10^7. The remaining parameters (α and β) are established according to Griffiths and Steyvers [16]. In this step, the 50 most important terms for each topic are obtained.
4. **TF/IDF-based Document Vectorization:** Then, we vectorize all documents taking into account the terms obtained in the previous step, using the well-known TF/IDF metric.

[7] We experimentally set this parameter after testing up to 50 topics.

5. **Factor analysis:** Matrix factorization is done using TF/IDF weights using *Singular Value Decomposition* (SVD). We obtain 115 components that explain a 70% of variance level.

Finally, we run the K-means algorithm to cluster all documents. For each K value we identify the celiac community (as explained in section 4.3). Then, we compute Recall, Precision, cluster size and the number of celiac users metrics based on the ground truth obtained from the experiments.

Figure 1 shows the values of Recall and Precision metrics for each cluster size. It is possible to apretiate that the total number of target users (belonging to the celiac community) are all identified when $K \geq 23$. This fact may be seen with the behavior of the Recall measure that remains with a minimum variation until the maximum number of clusters is reached. Following this trend, Precision increases when $k \geq 23$ up to its top value, close to 0.9.

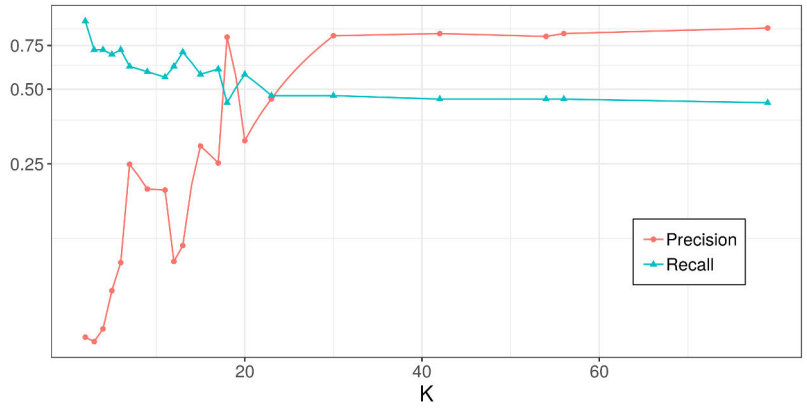

Fig. 1. Relationship between cluster size and Recall/Precision curves.

In an analogous way, Figure 2 shows the number of target users with respect to the total size of the cluster, for each value of K. The target users line represents users interested in the subject (Table 1), thus it is possible to see that the proportion of target users remains invariant and approximate to the total when $K \geq 30$.

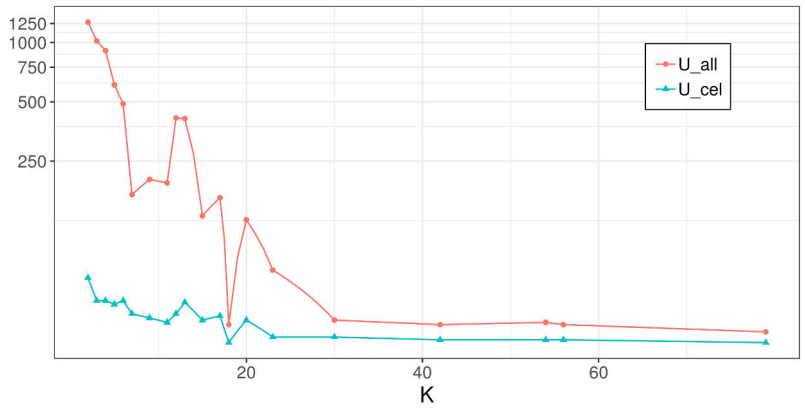

Fig. 2. Number of target users (U_{cel}) and total users (U_{all}) of the identified community, for each value of K.

5.3 Recommendation of Users

The main goal of this experiment is to evaluate the structural change in the community when some referring users in the social network regarding celiac disease are recommended to others. In this instance, we only simulate the process without intervention on the real network. The underlying idea is to select recommendations based on two attributes of the users: their influence and their activity in the social network, described in Section 4.5.

User	Score
$anonUser_1$	12.350
AlimentoSinTacc	9.333
$anonUser_2$	8.916
goutcafe1	8.472
GlutenFreeArg	7.859
TaccAway	7.744
sansglutenmdp	7.633
$anonUser_3$	7.078
Cocelia1	6.023
rojasglutenfree	5.901

Table 5. Most active users on celiac disease community (*top-10*). In a similar way as in Table 1, we replace the names of individual users ($anonUser_x$) to preserve anonymity.

Recommendation process: We start with the set of users of the celiac community found by the Louvain method on the main graph, G_{base} ($U_{com\_l\_base}$). Then, we use the two aforementioned attributes (influence and activity) to se-

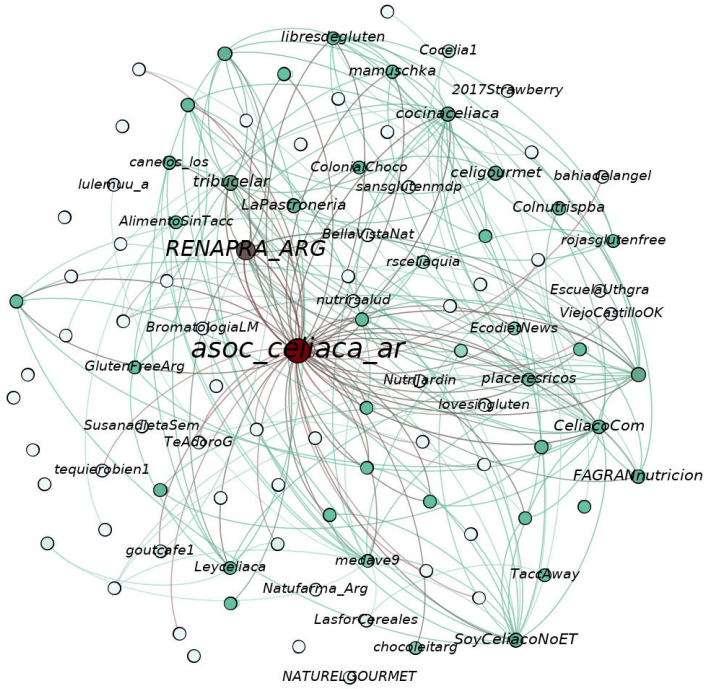

Fig. 3. Most influential users according to PageRank. For anonymity, we only maintain usernames that identify institutions, associations or companies and do not individualize people.

lect users, and we also combine them (a random selection is used as a baseline). In summary, we use four different criteria to select users, as:

1. **Influencers:** Influential users (l_{infl}) are obtained through PageRank (G_{inf}). Figure 3 shows the graph G_{inf} with the size of the nodes proportional to their score. Here, it is possible to see a *hub* user (*"asoc_celiaca_ar"*) as the most influential.
2. **Actives:** We get the list of the most active users (l_{act}) regarding the target topic (i.e. Table 5).
3. **Both(I+A):** We select users from the previous lists (influent and active users) intersecting them ($G_{inf} \cap G_{act}$) at an arbitrary cut threshold p, according to the method described in Section 4.5. Figure 4 shows how the intersection of both lists evolves according to different cut proportions (the same in both lists). For our experiments, in a conservative way, we set $p = 0.2$, obtaining a new subset U_{rec} of only a few users (Figure 4).
4. **Random:** As a baseline, we select users in a random fashion to verify that having an appropriate criterion leads to better results (i.e. better structural properties of nodes and the network itself).

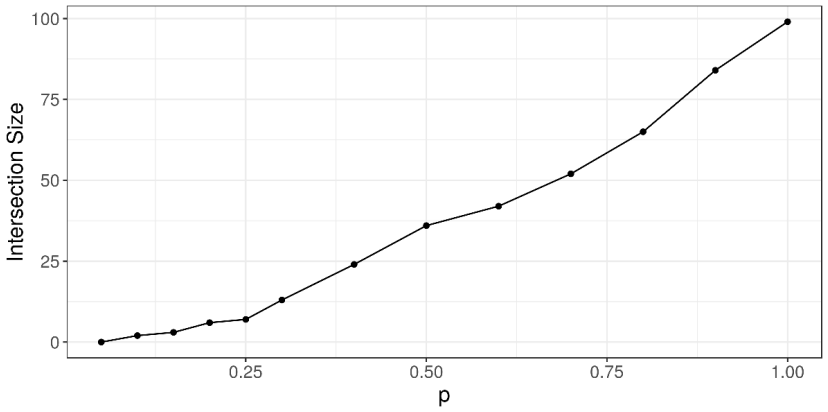

Fig. 4. Intersection size between influential and active users rankings.

The simulation of recommendations is made by taking each user $u_{rec} \in U_{rec}$ and for each user $u_{com\_l\_base} \in U_{com\_l\_base} : u_{com\_l\_base} \neq u_{rec}$, we check if there exists a link $(u_{com\_l\_base}, u_{rec})$. In the negative case, a new link is added with a probability of acceptance $P(a)$.

For the three first criteria, we run 10 trials varying the probability $P(a)$ and averaged the results. For the last one (Random), we run 25 trials to select the different set of users and then, for each one, we run 10 trials varying $P(a)$ (and averaged the results).

As evaluation metrics, we use the Clustering Coefficient, Average Closeness Centrality and Diameter of the resulting network. These metrics describe some structural modifications in the network, which may benefit (or not) the spread of information.

Figure 5 shows the results for Clustering Coefficient. Selecting users from the Influencers list or combining it with Actives ones, (Both(I+A) in the figure) perform quite similar. However, the combined list performs around 5% better than Actives' list and 7.5% better than Random's. Taking into account only low values of the threshold ($P(a) \leq 0.3$), which becomes a more realistic setting, the improvements raise up to 12% and 17%, respectively. As another interesting observation, when $P(a) = 0.35$, the Clustering Coefficient reaches its peak value (0.83) and the series change their slopes. This means that a relatively low probability of accepting a link quickly leads to a denser and better connected network.

Similarly, we compute the Closeness Centrality measure for all nodes in the network after the reinforcement process. Figure 6 shows the results. According to this metric, the best performance is achieved when we select users from both lists (Influencers + Actives). This criterion is 25%, 36% and 47% better (on average) than Influencers, Actives and Random, respectively. When considering only $P(a) \leq 0.3$, the improvement rises up 32%, excluding the Random serie which is clearly the poorest one and it is only included as a baseline. This

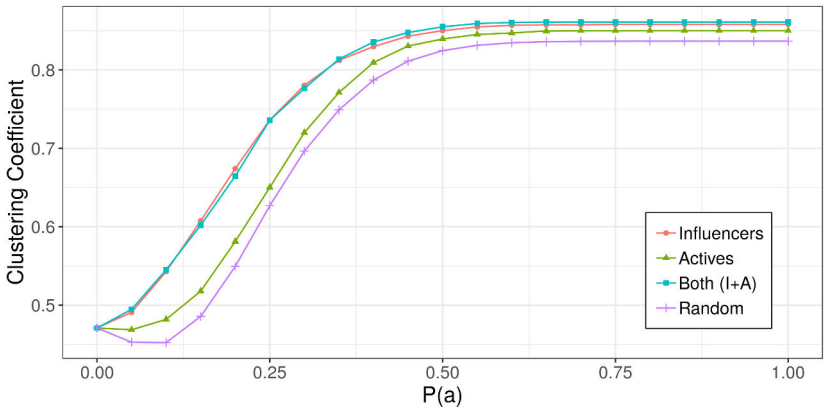

Fig. 5. Improvements on Clustering Coefficient according to $P(a)$.

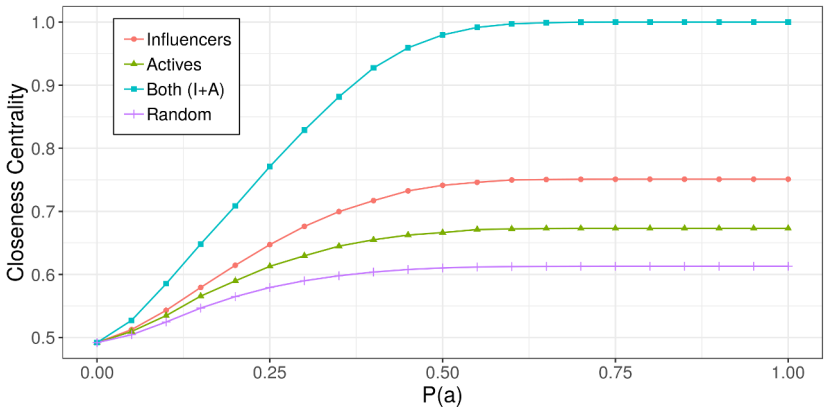

Fig. 6. Improvements on Closeness Centrality according to $P(a)$.

measure enables us to evaluate the speed at which the capacity of these referent users in celiac disease increases to disseminate information. Finally, we measure the diameter of the resulting network in the four cases. The diameter decreases from 5 to 3 in all cases because the network is quite small. However, using the Both (A + I) list to select recommendation allows this process to be faster with low values of $P(a)$ (Figure 7).

6 Conclusions and Future Work

The formation of communities in digital social networks is an interesting phenomenon from multiple points of view. For example, as an underlying structure, communities exhibit particular characteristics such as the density of their connections. Taking into account the users and their interactions, different behaviours

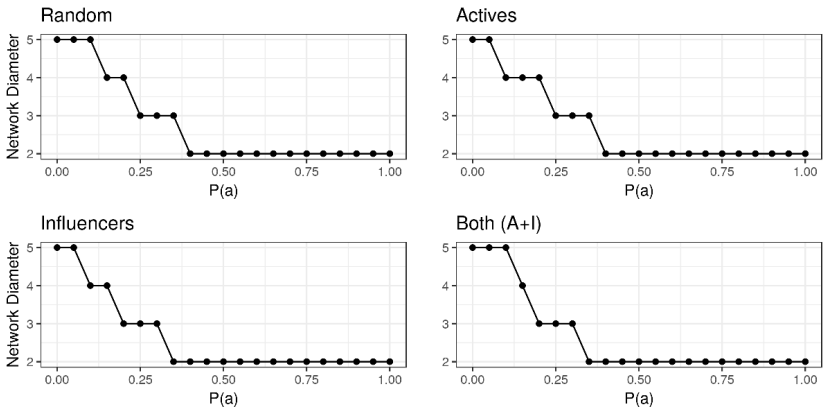

Fig. 7. Diameter of the resulting network according to $P(a)$.

appear, according to the nature of the community and its goals (sharing ideas, tastes, hobbies, etc.).

This paper addresses the problem of detecting and reinforce a community of Twitter users interested in the celiac disease, particularly in Argentina, complementing medical and biological field studies.

Applying combinations of several techniques a target community is found, that is composed by a limited number of users on which highly influential users are identified.

Considering only the structure of the graph we achieve a 65% of accuracy. This value improves when the edges are weighted using the user similarity criterion (up to 77%).

Regarding the use of KMeans combined with the criterion of celebrities, it is shown that it is possible to reach a Precision close to 0.9% with $K \geq 23$.

Finally, the user recommendation strategy based on influencers and active users shows that, by selecting only a small group of users and with a relatively low probability of acceptance of the recommendations, the network quickly becomes denser and better connected, which allows better dissemination of valuable information regarding celiac disease among those interested.

As future work, we plan to expand the study considering the evolution of the community over time, and proposing a strategy for the inclusion in it of users participating in various communities (or those who are partially interested in the topic). This last setup makes the identification of specific users a more challenging issue. In addition, we propose to compare the communities in Argentina with other geographical areas in which there exist current field studies regarding the celiac disease.

References

[1] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999. ISBN: 020139829X.

[2] Punam Bedi and Chhavi Sharma. "Community Detection in Social Networks". In: *Wiley Int. Rev. Data Min. and Knowl. Disc.* 6.3 (May 2016), pp. 115–135. ISSN: 1942-4787.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[4] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008.

[5] Cécile Bothorel et al. "Clustering attributed graphs: models, measures and methods". In: *Network Science* 3.3 (2015), pp. 408–444.

[6] Natacha Cerny et al. "Epidemiological study of Celiac Disease in Chivilcoy, Buenos Aires". In: *IV International Congres in Translational Medicine. School of Pharmacy and Biochemistry of Universidad de Buenos Aires* (2018).

[7] Meeyoung Cha et al. "Measuring user influence in twitter: The million follower fallacy." In: *Icwsm* 10.10-17 (2010), p. 30.

[8] Yves Darmaillac and Sébastien Loustau. "MCMC Louvain for Online Community Detection". In: *CoRR* abs/1612.01489 (2016). eprint: `1612.01489`.

[9] Pasquale DeMeo et al. "Generalized Louvain Method for Community Detection in Large Networks". In: *CoRR* abs/1108.1502 (2011). eprint: `1108.1502`.

[10] Imre Derényi, Gergely Palla, and Tamás Vicsek. "Clique percolation in random networks". In: *Physical review letters* 94.16 (2005), p. 160202.

[11] Himel Dev, Mohammed Eunus Ali, and Tanzima Hashem. "User Interaction Based Community Detection in Online Social Networks". In: *Database Systems for Advanced Applications*. Ed. by Sourav S. Bhowmick et al. Cham: Springer International Publishing, 2014, pp. 296–310. ISBN: 978-3-319-05813-9.

[12] Daniel Edler et al. "Infomap Bioregions: Interactive Mapping of Biogeographical Regions from Species Distributions". In: *Systematic Biology* 66.2 (2017), pp. 197–204.

[13] Davide Feltoni Gurini et al. "Enhancing Social Recommendation with Sentiment Communities". In: *Web Information Systems Engineering – WISE 2015*. Ed. by Jianyong Wang et al. Cham: Springer International Publishing, 2015, pp. 308–315. ISBN: 978-3-319-26187-4.

[14] Santo Fortunato and Claudio Castellano. "Community structure in graphs". In: *Computational Complexity*. Springer, 2012, pp. 490–512.

[15] Olivier Gach and Jin-Kao Hao. "Improving the Louvain Algorithm for Community Detection with Modularity Maximization". In: *Artificial Evo-*

lution. Ed. by Pierrick Legrand et al. Cham: Springer International Publishing, 2014, pp. 145–156. ISBN: 978-3-319-11683-9.

[16] Thomas L. Griffiths and Mark Steyvers. "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101.suppl 1 (2004), pp. 5228–5235. ISSN: 0027-8424. DOI: 10.1073/pnas.0307752101.

[17] Wei-Feng Guo and Shao-Wu Zhang. "A general method of community detection by identifying community centers with affinity propagation". In: *Physica A: Statistical Mechanics and its Applications* 447 (2016), pp. 508–519.

[18] Vairaprakash Gurusamy. "Mining the Attitude of Social Network Users using K-means Clustering". In: *International Journal of Advance Research in Computer Science and Software Engineering* 7 (May 2017), pp. 226–230.

[19] D. A. Huffman. "A Method for the Construction of Minimum-Redundancy Codes". In: *Proceedings of the IRE* 40.9 (Sept. 1952), pp. 1098–1101. ISSN: 0096-8390.

[20] Liping Jing, Michael K Ng, and Joshua Zhexue Huang. "An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data". In: *IEEE Transactions on knowledge and data engineering* 19.8 (2007).

[21] Mohit Naresh Kewalramani. "Community Detection in Twitter". PhD thesis. University of Maryland Baltimore County, 2011.

[22] Emre Kiciman et al. "Analyzing Social Media Relationships in Context with Discussion Graphs". In: *Eleventh Workshop on Mining and Learning with Graphs* (2013).

[23] Kwan Hui Lim and Amitava Datta. "A Topological Approach for Detecting Twitter Communities with Common Interests". In: *Ubiquitous Social Media Analysis*. Ed. by Martin Atzmueller et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 23–43.

[24] Kwan Hui Lim and Amitava Datta. "An interaction-based approach to detecting highly interactive Twitter communities using tweeting links". In: *Web Intelligence* 14.1 (2016), pp. 1–15.

[25] Kwan Hui Lim and Amitava Datta. "Following the Follower: Detecting Communities with Common Interests on Twitter". In: *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*. HT '12. Milwaukee, Wisconsin, USA: ACM, 2012, pp. 317–318. ISBN: 978-1-4503-1335-3.

[26] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. ISBN: 0521865719.

[27] Gunjan Mathur and Hemant Purohit. "Performance Analysis of Color Image Segmentation using K-Means Clustering Algorithm in Different Color Spaces". In: *IOSR Journal of VLSI and Signal Processing* 4 (Dec. 2014), pp. 01–04.

[28] M. E. J. Newman and M. Girvan. "Finding and evaluating community structure in networks". In: *Phys. Rev. E* 69 (2 Feb. 2004), p. 026113.

[29] L. Page et al. "The PageRank citation ranking: Bringing order to the Web". In: *Proceedings of the 7th International World Wide Web Conference*. 1998, pp. 161–172.

[30] S. Papadopoulos et al. "Community Detection in Social Media". In: *Data Mining and Knowledge Discovery* 24.3 (2012), pp. 515–554.

[31] Michel Plantié and Michel Crampes. "Survey on social community detection". In: *Social media retrieval*. Springer, 2013, pp. 65–85.

[32] X. Que et al. "Scalable Community Detection with the Louvain Algorithm". In: *2015 IEEE International Parallel and Distributed Processing Symposium*. May 2015, pp. 28–37.

[33] Y. Ren, R. Kraut, and S. Kiesler. "Applying Common Identity and Bond Theory to Design of Online Communities". In: *Organization studies* 28.3 (2017), pp. 377–408.

[34] M. Rosvall, D. Axelsson, and C. T. Bergstrom. "The map equation". In: *The European Physical Journal Special Topics* 178.1 (Nov. 2009), pp. 13–23. ISSN: 1951-6401.

[35] Yiye Ruan, David Fuhry, and Srinivasan Parthasarathy. "Efficient Community Detection in Large Networks using Content and Links". In: *CoRR* abs/1212.0146 (2012).

[36] María Inés Tamborenea et al. "Impacto sobre la Calidad de Vida en Pacientes con Enfermedad Celíaca (ec) en Tratamiento con Dieta Libre de Gluten (dlg)". In: *Simposio Panamericano de Enfermedad Celíaca. Buenos Aires Argentina* (2018).

[37] Lei Tang and Huan Liu. *Community Detection and Mining in Social Media*. 1st. Morgan and Claypool Publishers, 2010. ISBN: 9781608453542.

[38] Eleni Vathi, Georgios Siolas, and Andreas Stafylopatis. "Mining and categorizing interesting topics in Twitter communities". In: *Journal of Intelligent and Fuzzy Systems* 32.2 (2017), pp. 1265–1275.

[39] Elize Vis and Peer Scheepers. "Social Implications of Celiac Disease or Non-celiac Gluten Sensitivity". In: *International Journal of Celiac Disease* 5.4 (2017), pp. 133–139.

[40] Tao Wang et al. "Detecting and Characterizing Eating-Disorder Communities on Social Media". In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. WSDM '17. Cambridge, United Kingdom: ACM, 2017, pp. 91–100. ISBN: 978-1-4503-4675-7.

[41] Jaewon Yang, Julian McAuley, and Jure Leskovec. "Community detection in networks with node attributes". In: *Data Mining (ICDM), 2013 IEEE 13th international conference on*. IEEE. 2013, pp. 1151–1156.

[42] Xin Yu, Jing Yang, and Zhi-Qiang Xie. "A semantic overlapping community detection algorithm based on field sampling". In: *Expert Systems with Applications* 42.1 (2015), pp. 366–375.

[43] Yang Zhang, Yao Wu, and Qing Yang. "Community Discovery in Twitter Based on User Interests". In: *Journal of Computational Information Systems* (2012).

[44] Zhongying Zhao et al. "Topic oriented community detection through social objects and link analysis in social networks". In: *Knowledge-Based Systems* 26 (2012), pp. 164–173. ISSN: 09507051.